INFORMS Journal on Data Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Fair Collaborative Learning (FairCL): A Method to Improve Fairness amid Personalization

Feng Lin; , Chaoyue Zhao; , Xiaoning Qian; , Kendra Vehik, Shuai Huang;

informs
https://pubsonline.informs.org/journal/ijds

INFORMS JOURNAL ON DATA SCIENCE
Vol. 4, No. 1, January–March 2025, pp. 67–84
ISSN 2694-4022 (print), ISSN 2694-4030 (online)

Fair Collaborative Learning (FairCL): A Method to Improve Fairness amid Personalization

Feng Lin,[a] Chaoyue Zhao,[a] Xiaoning Qian,[b] Kendra Vehik,[c] Shuai Huang[a,*]

[a] Department of Industrial and Systems Engineering, University of Washington, Seattle, Washington 98195; [b] Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas 77843; [c] Health Informatics Institute, University of South Florida, Tampa, Florida 33620

*Corresponding author

Contact: ryanflin@uw.edu, https://orcid.org/0009-0009-6661-7955 (FL); cyzhao@uw.edu, https://orcid.org/0000-0003-0442-2991 (CZ); xqian@ece.tamu.edu, https://orcid.org/0000-0002-4347-2476 (XQ); kendra.vehik@epi.usf.edu, https://orcid.org/0000-0001-6243-6772 (KV); shuaih@uw.edu, https://orcid.org/0000-0003-2493-6818 (SH)

Received: January 4, 2024
Revised: August 13, 2024
Accepted: October 6, 2024
Published Online in Articles in Advance: December 4, 2024

https://doi.org/10.1287/ijds.2024.0029

Abstract. Model personalization has attracted widespread attention in recent years. In an ideal situation, if individuals' data are sufficient, model personalization can be realized by building models separately for different individuals using their own data. But, in reality, individuals often have data sets of varying sizes and qualities. To overcome this disparity, collaborative learning has emerged as a generic strategy for model personalization, but there is no mechanism to ensure fairness in this framework. In this paper, we develop fair collaborative learning (FairCL) that could potentially integrate a variety of fairness concepts. We further focus on two specific fairness metrics, the bounded individual loss and individual fairness, and develop a self-adaptive algorithm for FairCL and conduct both simulated and real-world case studies. Our study reveals that model fairness and accuracy could be improved simultaneously in the context of model personalization.

History: Bianca Maria Colosimo served as the senior editor for this article.
Funding: This work was supported by the Breakthrough T1D Award [Grant 2-SRA-2022-1259-S-B].
Data Ethics & Reproducibility Note: The code capsule is available on Code Ocean at https://codeocean.com/capsule/1331847/tree/v1 and in the e-Companion to this article (available at https://doi.org/10.1287/ijds.2024.0029). The real-world data, including the transportation demand management and surgical site infection data sets, are proprietary and not publicly available. Other results are available at https://github.com/ryanlif/FairCL.

Keywords: collaborative learning (CL) • fairness • model personalization • disparity

1. Introduction

Model personalization has been an important research problem in artificial intelligence (AI) and machine learning (McAuley 2022) with widespread applications ranging from advertising (Bilenko and Richardson 2011, Guo et al. 2021), to healthcare (Andreu-Perez et al. 2015, Abul-Husn and Kenny 2019, Gao et al. 2022) and robotics (Ma et al. 2021, Firdaus et al. 2022). This need for model personalization stems from the rapid advancements in technologies, such as smart devices, intelligent agents, and recommendation systems, which are designed to provide customized interactions and services to individuals coming from a diverse population. An essential task for model personalization is to deal with this dilemma: given a disparate collection of data sets from a population of individuals, shall we build one model as a one-size-fits-all solution by pooling all individuals' data together or build an individual model based solely on each individual's data? In an ideal situation, with sufficient data for each individual, there should be no such dilemma. However, in reality, a great disparity exists among the data sets of different individuals in size and quality.

To address this dilemma, collaborative learning (CL) has been developed in recent years. It provides a generic framework for model personalization and can be readily extended to a wide range of models, such as regression (Lin et al. 2015), classification (Feng et al. 2020), Markov models (Lin et al. 2018), and other utility-based behavioral models (Feng et al. 2020, Zhu et al. 2020). It also has a theoretical connection with the mixed effects models (Feng et al. 2020). As collaborative learning has demonstrated remarkable efficiency in improving accuracy of individual models with the increasing concern on fairness in machine learning models, it is natural to ask whether collaborative learning has any fairness guarantees. To our knowledge, there has been a lack of investigation into this question in the literature. Fairness in machine learning has emerged as a critical issue in AI ethics because of the use of AI in high-stake decision-making applications, such as law enforcement, justice, finance, education, healthcare, etc. Fairness is broadly depicted as the absence of preference for an individual or a group with specific characteristics (Mehrabi et al. 2021b). As we strive for model personalization, it is important to investigate whether personalization of the models could potentially exacerbate existing disparities among individuals and elevate the degree of unfairness. If so, how can we reformulate the collaborative learning method to pursue both model personalization and fairness simultaneously?

In this paper, we show that model personalization does not inevitably come at the expense of fairness. Think of the aforementioned dilemma: the one-size-fits-all solution only seems fair through unawareness of the different individuals. However, it does not guarantee the same prediction accuracy for them. As a matter of fact, this one-size-fits-all approach causes substantial disparities, which can result in bad model performance in practice. On the other hand, fully individual learning (IL) with no fairness considerations might also lead to such a disparity. But collaborative learning was actually motivated to reduce this disparity (Lin et al. 2017). In this sense, compared with the one-size-fits-all and IL approaches, collaborative learning has been already more concerned with fairness. Nonetheless, there is much room for collaborative learning to further enhance the fairness of the individual models because fairness has not been explicitly formulated in collaborative learning. In this paper, we propose a framework named fair collaborative learning (FairCL) that can potentially integrate a wide range of fairness concepts. We further focus on two specific fairness metrics, the bounded individual loss (BIL) and individual fairness (IF), and develop a self-adaptive algorithm for FairCL and conduct both simulated and real-world case studies. One fundamental difference between our problem and general fair machine learning is that we build many different models for personalization, whereas the latter usually assumes one model for all groups. The remainder of the article unfolds as follows: In Section 2, we provide a literature review of fair machine learning and model personalization. In Section 3, we review the formulation of collaborative learning and its basic rationale in detail. In Section 4, we show that the collaborative learning framework is versatile to incorporate a wide range of fairness metrics. In Section 5, we present a self-adaptive fair collaborative learning method and explore its connection with other fair collaborative learning methods. Experimental results on both simulated and real-world data are shown in Section 6, and conclusion is given in Section 7.

2. Related Work

2.1. Fair Machine Learning

Fairness in machine learning has attracted tremendous attention from the research community (Corbett-Davies and Goel 2018, Zhang et al. 2020, Mehrabi et al. 2021b). Even though there is no consensus on the universal definition of fairness (Caton and Haas 2020), researchers have proposed various metrics from different perspectives to quantify fairness in learning systems and develop specific algorithms to reduce biases. Given that this is a vast field, here, we focus on the most relevant work that concerns the design of the learning formulations and develops fairness metrics or constraints.

One of the earliest and most widely used metrics is demographic parity (DP), also known as statistical parity (Zemel et al. 2013, Corbett-Davies et al. 2017), which aims to ensure that the distribution of the model outcome remains the same regardless of sensitive attributes such as gender or race. Another similar notion is disparate impact (Feldman et al. 2015), which uses the probability ratio between protected and unprotected groups to measure fairness. However, these metrics can only support pairwise comparison, and their effectiveness might be undermined when the sensitive attribute is not binary. To handle more general sensitive attributes, Jiang et al. (2022) further propose generalized demographic parity, which extends the parity-based metric to incorporate continuous sensitive attributes, maintaining computational efficiency. Different from parity-based metrics, equal opportunity (Hardt et al. 2016) and equalized odds (Berk et al. 2021) were proposed to take not only predictions, but also actual outcomes into account. Such group metrics more or less help with balancing between groups, but they may ignore protections for specific individuals in a group. In addition to these metrics, there are also other metrics that are not limited to comparison across groups. For example, Dwork et al. (2012) present IF to advocate that similar individuals should receive similar treatment. However, this metric is also criticized for potentially leading to universal rejection and the difficult in measuring similarity (Fleisher 2021). As the fairness metrics might be controversial in their use, it is very important to choose appropriate metrics in different contexts, which still remains challenging.

With the emergence of many fairness metrics, researchers are also developing generic optimization formulations and computational strategies to incorporate these fairness metrics into various models. For example, Alabi et al. (2018) develop an efficient optimization framework for fairness by minimizing the specific bounded group loss (BGL). To incorporate more general fairness metrics, from the perspective of nonconvex optimization, Cotter et al. (2019) propose to express fairness as a kind of rate constraint in a proxy-Lagrangian formulation and provide a theoretically guaranteed procedure to solve it. This framework provides a general way to incorporate fairness as constraints but suffers from dealing with nonlinearity in such constraints. Zafar et al. (2019) narrow the problem down and propose a flexible constraint-based framework enabling the design of fair margin-based classifiers. This framework uses covariance between sensitive and nonsensitive attributes as a tractable proxy for unfairness and aims to limit unfairness with small performance cost. However, it might result in poor performance when the data are unbalanced. Whereas most of these studies focus on classification problems, there are also some works considering regression problems, such as Chzhen et al. (2020) and Zhao and Chen (2019), which is beyond the scope of this paper.

2.2. Model Personalization

Models that can be used for personalization include the classic mixed effects model (Cudeck 1996, Laird and Ware 1982) and more modern ones, such as transfer learning (Jiang et al. 2020, Yang et al. 2020) or multitask learning (Li et al. 2021, Zhang and Yang 2021). Transfer learning is a learning paradigm that improves learning in a target domain by transferring knowledge from a source domain (Zhuang et al. 2020), whereas multitask learning aims to leverage information shared by multiple related tasks to help with learning for all the tasks (Zhang and Yang 2021). Despite attaching different degrees of importance to various tasks of interest, both of them are similar in the strategies adopted for modeling (Yang et al. 2020). Within the general umbrella of transfer learning, CL was developed (Lin et al. 2015, 2017; Feng et al. 2020). It features a concept known as canonical models, which distinguishes itself from other transfer learning or multitask learning models and employs a probabilistic relationship to characterize individual models.

Another line of work that is related to model personalization is distributed learning (Miao et al. 2015, Verbraeken et al. 2020), in which data/computing resources are allocated to different clients. As a distributed learning approach that emphasizes data privacy protection, federated learning (FL) has become increasingly popular in recent years (Yang et al. 2019, Liu et al. 2022). To compensate for its poor performance caused by data heterogeneity, personalization techniques have started to be used in federated learning (Kulkarni et al. 2020, Tan et al. 2022, Zhang et al. 2022). At first sight, this line of work is somewhat similar to CL. However, it is worth noting that there are several differences between them: (i) CL aims to model individuals by utilizing population-level information, whereas personalization in FL strives for better FL on heterogeneous data. (ii) CL allows the sharing of data together with model information among individuals, whereas the privacy requirement in FL prevents data from being shared among clients and only allows sharing the knowledge of model updates. (iii) CL assumes the individuals share a latent canonical structure, whereas FL usually defaults to independence among clients. (iv) CL emphasizes model personalization through centralized training but with no explicit central server, whereas personalized FL allows training personalized models in a decentralized way through information sharing between an explicit central server and clients. (v) CL is useful when individuals have different amounts of data, whereas traditional FL sometimes relies on sufficient data from its distributed clients to reach a minimum model quality (Shamir and Srebro 2014, Kamp et al. 2023). It should be noted that CL and FL are not competitive with each other, but instead, they solve different questions. And CL can be compatible with FL. Another approach worth mentioning is the clustered federated learning (CFL) (Mansour et al. 2020, Sattler et al. 2020). It addresses heterogeneity by explicitly identifying the clusters of the users and performing FL to learn separate models for each cluster instead of one global model. Whereas some works use centralized clustering algorithms such as K-means (Ghosh et al. 2019, Sattler et al. 2020), hierarchical clustering (Briggs et al. 2020, Mehta and Shao 2023), etc., others use decentralized algorithms (Ghosh et al. 2020, He et al. 2023) to recognize the cluster identities. In addition to hard clustering, Morafah et al. (2023), Ruan and Joe-Wong (2022), and Marfoq et al. (2021) propose to use soft clustering strategies. CFL methods provide a promising way to handle user heterogeneity through a clustering structure, but their performance depends on the ability to find good clusters. In contrast to CFL, CL doesn't need to cluster the users, but encodes the sharing characteristics among the users through a latent canonical structure that is not necessarily a clustering structure, which also allows simple and efficient personalization through one unified optimization framework.

Finally, different from model personalization, a remotely related line of work is the design of personalization strategies. For example, personalized recommendation (Benhamdi et al. 2017, Amara and Subramanian 2020, Knisely et al. 2021) is becoming increasingly pervasive to handle user heterogeneity (McAuley 2022). Such personalization is usually conducted through collaborative filtering or content-based filtering. Collaborative filtering makes recommendations to an individual user based on the preference of other similar users, whereas content-based filtering recommends similar items to the user without taking other users' information into consideration. These personalized strategies are tied to many specific application domains, such as healthcare (Alfian et al. 2018, Emmert-Streib and Dehmer 2018), e-commerce (Yoganarasimhan 2020, Ban and Keskin 2021), education (Liu et al. 2019, Xu et al. 2021), etc. And they have prospered rapidly in recent years with the development of deep learning techniques (Naumov et al. 2019, Tobore et al. 2019).

2.3. Fairness-Aware Federated Learning

There have been works in FL focusing on ensuring fairness in different stages of FL, building on various fairness metrics proposed in literature (Shi et al. 2023a). Some studies aim to enforce fairness constraints on client selection (Shi et al. 2023b). For example, Huang et al. (2020) propose to incorporate a fairness constraint in Lyapunov optimization to guarantee the participation rate of clients, whereas Yang et al. (2021) and Huang et al. (2022) used bandit-based algorithms to encourage a fairer selection. Whereas these methods ignore a real-time client contribution, it is further addressed in Song et al. (2021) by combining a reputation table with fairness constraints in selection. Other work promotes the selection of underrepresented clients through personalization of training procedures, such as Caldas et al. (2018), and Li et al. (2020b). Nevertheless, as the biases in training may not just come from client selection, more studies are concerned with fairness in the model optimization process (Shi et al. 2023b). Li et al. (2021) developed a multitask personalized FL framework called Ditto to promote fairness in FL through regularization on the discrepancy between personalized models and the global model in local training. Similarly, Yue et al. (2023) also introduce regularization terms in their GIFAIR-FL framework to ensure fairness. Instead of penalizing the models, they regularize between the local loss functions and show that their formulation can be equivalently written as loss reweighting. Reweighting is also applied in other recent work. Extending the idea of agnostic federated learning (Mohri et al. 2019), which optimized their loss over an unknown mixture of the client data distributions, Du et al. (2021) propose the AgnosticFair framework. It reweights an agnostic loss function with kernel function parameterization and constructs an agnostic fairness constraint from demographic parity. However, its reliance on prior knowledge limits its applications (Shi et al. 2023a). Li et al. (2020a) propose a q-fair federated learning framework in which different clients are weighted differently in the aggregated loss through a power parameter q. But its effectiveness is affected by the estimation of a Lipschitz constant and the setting of q. When q is large, it may also suffer numerical issues in computation. In addition to model optimization, there are also studies aiming to improve fairness in the stage of incentive distribution, for instance, Lyu et al. (2020), Gao et al. (2021), and Yu et al. (2020) take fairness into consideration through the design of incentive mechanisms.

3. Collaborative Learning

In this section, we present the basic framework of CL as developed in (Lin et al. 2015, 2018; Feng et al. 2020). This framework can be applied to both regression (Lin et al. 2017) and classification (Feng et al. 2020). Whereas the majority of fair machine learning works center on classification, in this paper, we also focus on the classification problem to develop our models and algorithms.

3.1. The Basic Framework of CL

Suppose we have N individuals, each with its own set of observations. Let n_i denote the number of observations collected from the ith individual. Each observation can be represented as a pair $(\boldsymbol{x}_{ij}, y_{ij})(j = 1,\ldots,n_i)$, where $\boldsymbol{x}_{ij} \in \mathbb{R}^d$ is a feature vector and $y_{ij} \in \{0,1\}$ is a binary label. Our task is to build a personalized model for each individual; that is, for the ith individual, we find a mapping $h_i : \mathbb{R}^d \to \{0,1\}$ to predict the label y_{ij} given the feature vector $\boldsymbol{x}_{ij}$.

In an ideal scenario, each individual would have a sufficiently large data set to train an accurate personalized model for this task through fully IL. However, this is rarely the case in practice. Whereas it seems more realistic to simply pool all individuals' data together to build one population-level model, this strategy could not characterize population heterogeneity and thereby causes a great disparity on prediction accuracy among individuals. CL offers a middle ground between these two extremes. Rather than building one population-level model, CL builds a set of canonical models that can provide an effective representation of individual models: each individual model is characterized as a probabilistic combination of these canonical models, and this relationship is parameterized by a membership vector (Feng et al. 2020). CL learns the parameters of the canonical models and the membership vectors of the individuals.

More specifically, let K denote the number of canonical models, in which the kth canonical model is represented as $g_k(\boldsymbol{x})$:

$$g_k(\boldsymbol{x}) = \log\frac{P(y=1)}{P(y=0)} = \boldsymbol{x}^\top \boldsymbol{q}_k.$$

Here, $\boldsymbol{q}_k$ denotes the unknown parameters of the kth canonical model. For the ith individual, their individual model can be written as

$$\log\frac{P(y_{ij}=1\,|\,\boldsymbol{x}_{ij})}{P(y_{ij}=0\,|\,\boldsymbol{x}_{ij})} = \sum_{k=1}^{K} c_{ik} g_k(\boldsymbol{x}_{ij}) = {\boldsymbol{x}_{ij}}^\top \sum_{k=1}^{K} c_{ik}\boldsymbol{q}_k,$$

where $c_{ik} \in [0,1]$ is the membership parameter that characterizes how likely it is that the model of the ith individual comes from the kth canonical model. The sum of $c_{ik}(k=1,\ldots,K)$ equals one, ensuring that the individual's model is a probabilistic combination of the canonical models. Denote $\mathbf{Q} = [\,\boldsymbol{q}_1 \cdots \boldsymbol{q}_K\,]$ and $\boldsymbol{c}_i = [\,c_{i1} \cdots c_{iK}\,]$. The individual model can be rewritten as a two-parameter logistic regression model:

$$p_{ij} = P(y_{ij}=1\,|\,\boldsymbol{x}_{ij}) = \frac{\exp({\boldsymbol{x}_{ij}}^\top \mathbf{Q}\boldsymbol{c}_i)}{1+\exp({\boldsymbol{x}_{ij}}^\top \mathbf{Q}\boldsymbol{c}_i)}.$$

The likelihood function of the CL model can be written as

$$\begin{aligned} L(\mathbf{Q},\mathbf{C}) &= \prod_{i=1}^{N}\prod_{j=1}^{n_i} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}} \\ &= \prod_{i=1}^{N}\prod_{j=1}^{n_i}\left[\frac{\exp(\boldsymbol{x}_{ij}^\top \mathbf{Q}\boldsymbol{c}_i)}{1+\exp(\boldsymbol{x}_{ij}^\top \mathbf{Q}\boldsymbol{c}_i)}\right]^{y_{ij}}\left[\frac{1}{1+\exp(\boldsymbol{x}_{ij}^\top \mathbf{Q}\boldsymbol{c}_i)}\right]^{1-y_{ij}}. \end{aligned}$$

This gives the log-likelihood function

$$l(\mathbf{Q},\mathbf{C}) = \log L(\mathbf{Q},\mathbf{C}) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}[y_{ij}\log \exp(\boldsymbol{x}_{ij}^{\top}\mathbf{Q}\boldsymbol{c}_i) - \log(1+\exp(\boldsymbol{x}_{ij}^{\top}\mathbf{Q}\boldsymbol{c}_i))].$$

The maximization of the log-likelihood is equivalent to the minimization of the negative log-likelihood loss $-l(\mathbf{Q},\mathbf{C})$. We further normalize each individual's loss by its sample size, which leads to the following optimization problem:

$$\begin{aligned}\min_{\mathbf{C},\mathbf{Q}} \quad & \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}[\log(1+\exp(\boldsymbol{x}_{ij}^{\top}\mathbf{Q}\boldsymbol{c}_i)) - y_{ij}(\boldsymbol{x}_{ij}^{\top}\mathbf{Q}\boldsymbol{c}_i)] \\ \text{s.t.} \quad & \boldsymbol{c}_i^{\top}\mathbf{1}=1,\ \boldsymbol{c}_i \geq 0, \quad i=1,\ldots,N.\end{aligned} \tag{1}$$

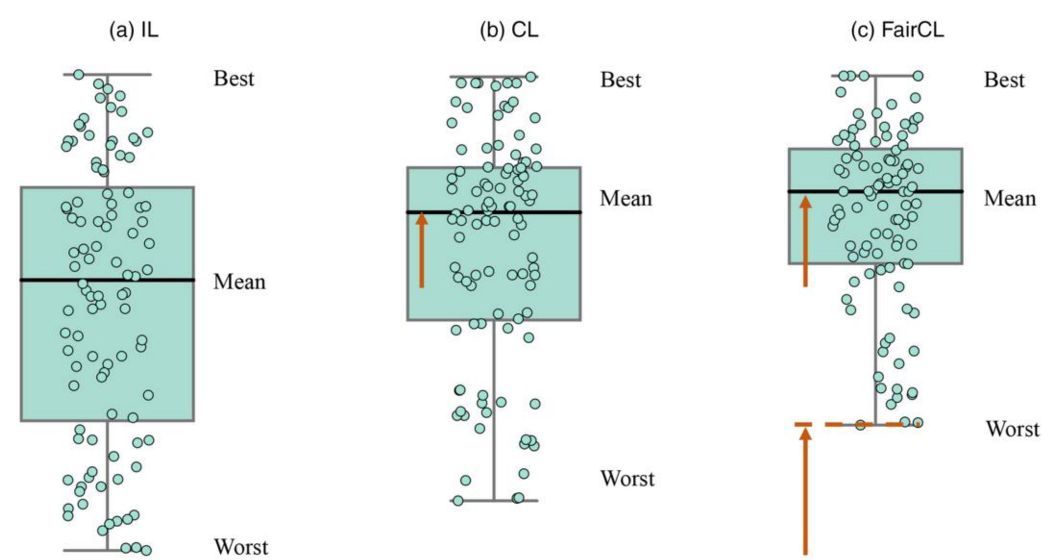
3.2. The Limitation of the CL Framework to Achieve Fairness

Compared with the fully IL scheme, CL promotes fairness by enabling high-level data sharing among the individuals and considering the canonical models as extra data that is fused with the individual's data, thus reducing the performance disparity among the individual models. However, CL has no explicit control over fairness. It does not incorporate any fairness constraints into the learning formulation as is done in many existing fair machine learning models. Meanwhile, if we take a close look at the objective function of CL in (1), the goal of CL is to maximize the mean performance of the models on all the individuals (as illustrated in Figure 1(b)). This indicates that, to further improve fairness on CL, we can enhance the CL formulation by minimizing variance at the same time, still improving on the mean performance (as illustrated in Figure 1(c)). The remaining question is whether these two objectives are in conflict with each other. After all, it is commonly known in fair machine learning literature that there is a trade-off between model accuracy and fairness, and one cannot increase both simultaneously. This trade-off can be quite intuitively conceived in the existing works of fair machine learning because they build one model to serve all groups. Therefore, if one aims to increase fairness, the single model shared by all groups has to be changed, and the overall accuracy can only go down instead of going up as fairness also goes up. However, in our problem, we build personalized models rather than one single model, so this trade-off is not necessarily true. Intuitively speaking, as the current CL framework adopts a complex and nonlinear optimization problem that suffers from suboptimal solutions and overfitting, our FairCL models may help find a better optimal solution and achieve better generalization and thereby improve both the accuracy of the models and reduce the disparity among them. Indeed, our experimental results in Section 6 show that the fairness and accuracy of the personalized models can actually be improved at the same time.

4. Collaborative Learning with Fairness Constraints (FairCL)

In this section, we explore how the CL can be integrated with existing fairness constraints to mitigate disparity among individuals and discuss the advantages and disadvantages of different FairCL formulations.

Figure 1. (Color online) An Illustration of the Boxplots of the Accuracy of Individual Models Learned by IL (a), CL (b), and FairCL (c)



Note. CL improves on the mean level, whereas our FairCL improves on the worst cases and also improves the mean level.

4.1. A General Framework for FairCL

The field of fairness machine learning focuses on developing general fairness constraints, which are usually designed to bound a certain fairness metric on each group (i.e., the groups are defined based on some sensitive attributes, such as gender and race). Then, fair learning models can be straightforwardly developed by integrating these fairness constraints into the learning formulation. In our context, it is natural to extend these fairness definitions on individuals (i.e., each individual is a "group"). A general formulation of our FairCL can be written as

$$\begin{aligned} \min_{\mathbf{C},\mathbf{Q}} \quad & \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} [\log(1+\exp(\boldsymbol{x}_{ij}^\top \mathbf{Q} \boldsymbol{c}_i)) - y_{ij}(\boldsymbol{x}_{ij}^\top \mathbf{Q} \boldsymbol{c}_i)] \\ \text{s.t.} \quad & \boldsymbol{c}_i^\top \mathbf{1} = 1,\ \boldsymbol{c}_i \geq 0, \quad i = 1, \ldots, N \\ & f_i(\mathbf{C},\mathbf{Q}) \leq 0, \quad i = 1, \ldots, N, \end{aligned} \tag{2}$$

where $f_i(\mathbf{C},\mathbf{Q}) \leq 0 (i = 1, \ldots, n)$ are the fairness constraints (i.e., some specific examples are introduced in Section 4.2). It can be reformulated by introducing the Lagrangian multipliers $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ as

$$\begin{aligned} \min_{\mathbf{C},\mathbf{Q}} \quad & \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} [\log(1+\exp(\boldsymbol{x}_{ij}^\top \mathbf{Q} \boldsymbol{c}_i)) - y_{ij}(\boldsymbol{x}_{ij}^\top \mathbf{Q} \boldsymbol{c}_i)] \\ & + \sum_{i=1}^{N} \lambda_i f_i(\mathbf{C},\mathbf{Q}) \\ \text{s.t.} \quad & \boldsymbol{c}_i^\top \mathbf{1} = 1,\ \boldsymbol{c}_i \geq 0, \quad i = 1, \ldots, N. \end{aligned} \tag{3}$$

Online Algorithm S1 presents the general computational framework of FairCL. In the subsequent section, we introduce various specific FairCL formulations that aim to achieve fairness by different definitions. For brevity, we denote the empirical loss of individual i by $\ell_i(\mathbf{X}_i, \boldsymbol{y}_i; \mathbf{Q}, \boldsymbol{c}_i) \triangleq \frac{1}{n_i} \sum_{j=1}^{n_i} [\log(1+\exp(\boldsymbol{x}_{ij}^\top \mathbf{Q} \boldsymbol{c}_i)) - y_{ij}(\boldsymbol{x}_{ij}^\top \mathbf{Q} \boldsymbol{c}_i)]$.

4.2. Some Specific FairCL Models and Algorithms

In theory, our FairCL framework shown in Equation (3) is general and can incorporate a variety of fairness concepts. But further modeling and computational improvement can be made when adapting this framework to a specific fairness constraint. In what follows, we present some examples of how this can be done. We focus on widely adopted fairness metrics including the following:

- Equality of opportunity (EO) (Hardt et al. 2016). This requires the model to assign a positive outcome to the positive population with equal opportunity regardless of their groups; that is, $P(\hat{Y} = 1 | A = 0, Y = 1) = P(\hat{Y} = 1 | A = 1, Y = 1)$, where A is the sensitive attribute to be protected, Y is the label, and $\hat{Y}$ is the prediction of the outcome given by the learned model.
- DP (Dwork et al. 2012). This requires that the model's predictions should not depend on the sensitive attributes; that is, $P(\hat{Y} | A = 0) = P(\hat{Y} | A = 1)$.
- Fairness through unawareness (FTU) (Chen et al. 2019). This requires that the sensitive attributes should not be observed or explicitly used in the model prediction.
- BGL (Alabi et al. 2018, Agarwal et al. 2019). This requires that the average group loss should be bounded; that is, $\mathbb{E}[\ell(\hat{Y}, Y) | A] \leq \zeta$, where ζ is an upper bound and $\ell(\cdot,\cdot)$ is the loss function.
- IF (Dwork et al. 2012). This requires that similar individuals should be assigned with similar predictions; that is, $D(\hat{Y}(X_1, A_1), \hat{Y}(X_2, A_2)) \leq d(X_1, X_2)$, where X_i and A_i are the insensitive and sensitive attributes of an individual, respectively. Here, $D(\cdot)$ and $d(\cdot)$ are distance metrics for predictions $\hat{Y}$ and attributes X, respectively.
- Counterfactual fairness (Kusner et al. 2017). This requires that, in both actual and counterfactual worlds in which an individual has different sensitive attribute values, predictions should be the same; that is, $P(\hat{Y}_{A\leftarrow a}(U) | X, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) | X, A = a)$, where A is the sensitive attribute taking the value of a or a', U denotes the latent variables, and X denotes the insensitive attributes.

4.2.1. EO and DP. Whereas these two fairness concepts are foundational to the field of fair machine learning, their practical relevance with personalized learning is quite limited. As Pessach and Shmueli (2020) point out, the DP metric might mistakenly deem a model unfair when the outcome distributions of various groups are different. This can be particularly problematic in the context of personalized learning, in which the population heterogeneity can lead to significant variations in the proportion of actual positive outcomes across individuals. In such cases, enforcing similar outcome distributions for all individuals may not necessarily improve fairness but instead result in misleading models. Only if the data distributions of individuals are very much alike (which is quite uncommon in practice), DP and EO would be practical and realistic. Therefore, in this paper, we do not further develop algorithms or conduct experiments for these two fairness concepts, but only present the FairCL formulation of DP here for theoretical interest:

$$\begin{aligned} \min_{\mathbf{C},\mathbf{Q}} \quad & \sum_{i=1}^{N} \ell_i(\mathbf{X}_i, \boldsymbol{y}_i; \mathbf{Q}, \boldsymbol{c}_i) \\ \text{s.t.} \quad & \boldsymbol{c}_i^\top \mathbf{1} = 1,\ \boldsymbol{c}_i \geq 0, \quad i = 1, \ldots, N \\ & |P(h_i(\mathbf{x}; \mathbf{Q}, \mathbf{C} \geq z) | i) - P(h_i(\mathbf{x}; \mathbf{Q}, \mathbf{C}) \geq z)| \leq \xi_i, \\ & \forall z, \quad i = 1, \ldots, N. \end{aligned}$$

Here, ξ_i is a slack variable that is a hyperparameter to create tolerance and make the formulation feasible because, in practice, we will not likely achieve absolute DP.

4.2.2. FTU. It is worthy of mentioning that the FTU actually corresponds to the one-size-fits-all approach for personalized learning.

4.2.3. Similarity-Based IF. The notion of IF (Dwork et al. 2012) provides a trade-off between equal treatment of individuals and individual heterogeneity. It does not necessitate absolute equality, but rather mandates that individuals with similar characteristics should be treated similarly. To achieve IF, usually a similarity matrix is constructed to measure the similarity between the individuals, denoted as $\mathbf{S} = [s_{ij}]_{N\times N}$. We can obtain the Laplacian matrix $\mathbf{L}_s = \mathbf{D} - \mathbf{S}$. Here, $\mathbf{D} = \mathrm{diag}(d_{11}, \ldots, d_{NN})$ is a $n \times n$ diagonal matrix, where $d_{ii} = \sum_{j=1}^{N} s_{ij}$. Using $\mathbf{L}_s$, it is easy to incorporate the similarity knowledge among individuals to obtain a graph-regularized (Belkin et al. 2005) formulation of CL to enforce IF:

$$\begin{aligned} \min_{\mathbf{C},\mathbf{Q}} \quad & \sum_{i=1}^{N} \ell_i(\mathbf{X}_i, \mathbf{y}_i; \mathbf{Q}, \mathbf{c}_i) + \lambda Tr(\mathbf{C}^T \mathbf{L}_s \mathbf{C}) \\ \text{s.t.} \quad & \mathbf{c}_i^\top \mathbf{1} = 1,\ \mathbf{c}_i \geq 0, \quad i = 1, \ldots, N. \end{aligned}$$

Actually, this formulation of FairCL-IF has been "unconsciously" developed in the prior work on collaborative learning (Feng et al. 2020), motivated not by fairness concerns, but rather for accuracy improvement. However, IF alone is inadequate to guarantee protection on vulnerable individuals because it only calls for similar treatment on similar individuals (Fleisher 2021). Another difficulty is the construction of the similarity matrix, which usually demands a strong prior knowledge.

4.2.4. BIL. BIL, as a natural generalization of BGL (Agarwal et al. 2019, Chi et al. 2021, Hu et al. 2022), bounds the loss function of each individual to reduce disparity. Formally, BIL is defined as

$$\mathbb{E}[\ell_i(\mathbf{X}_i, \mathbf{y}_i; \mathbf{Q}, \mathbf{c}_i) | i] \leq \zeta_i, \quad i = 1, \ldots, N.$$

The formulation of FairCL-BIL can be cast as

$$\begin{aligned} \min_{\mathbf{C},\mathbf{Q}} \quad & \sum_{i=1}^{N} \ell_i(\mathbf{X}_i, \mathbf{y}_i; \mathbf{Q}, \mathbf{c}_i) \\ \text{s.t.} \quad & \mathbf{c}_i^\top \mathbf{1} = 1,\ \mathbf{c}_i \geq 0, \quad i = 1, \ldots, N \\ & \ell_i(\mathbf{X}_i, \mathbf{y}_i; \mathbf{Q}, \mathbf{c}_i) \leq \zeta_i, \\ & i = 1, \ldots, N. \end{aligned} \tag{4}$$

To make it easier to solve, we can reformulate the problem using Lagrangian techniques. The Lagrangian function of the above formulation is

$$\ell(\mathbf{Q}, \mathbf{C}; \boldsymbol{\lambda}) = \sum_{i=1}^{N} \ell_i(\mathbf{X}_i, \mathbf{y}_i; \mathbf{Q}, \mathbf{c}_i) + \sum_{i=1}^{N} \lambda_i [\ell_i(\mathbf{X}_i, \mathbf{y}_i; \mathbf{Q}, \mathbf{c}_i) - \zeta_i].$$

To prevent $\boldsymbol{\lambda}$ from being arbitrarily large as demonstrated by Agarwal et al. (2019), the FairCL-BIL problem can be reformulated by constraining the L_1 norm of $\boldsymbol{\lambda}$ to a limit of B:

$$\begin{aligned} \min_{\mathbf{Q},\mathbf{C}} \max_{\boldsymbol{\lambda}} \quad & \ell(\mathbf{Q}, \mathbf{C}; \boldsymbol{\lambda}) \\ \text{s.t.} \quad & \mathbf{c}_i^\top \mathbf{1} = 1,\ \mathbf{c}_i \geq 0, \quad i = 1, \ldots, N \\ & \|\boldsymbol{\lambda}\|_1 \leq B,\ \boldsymbol{\lambda} \geq 0. \end{aligned} \tag{5}$$

We propose to solve FairCL-BIL following the scheme in Agarwal et al. (2019) that takes an exponentiated gradient approach for optimization (Online Algorithm S2), which is an alternating iterative strategy to iterate through $\mathbf{C}$, $\mathbf{Q}$, and $\boldsymbol{\lambda}$.

5. A Self-Adaptive Formulation: Rearrangement by Loss Reweighting

The FairCL formulations developed in Section 4 show the flexibility of incorporating fairness constraints into the CL framework. Among all these possibilities, FairCL-BIL provides the most practical formulation to incorporate fairness into CL. More specifically, it provides a way to achieve fairness by controlling over the individual loss functions. When the worst individual loss is confined to a certain level, the overall loss is pushed to stay low, and the disparity between the individual losses can be shrunk by adjusting the tolerance level. Unlike DP, FairCL-BIL does not impose strict restriction, and it does not require additional information, such as the similarity matrix in IF, which can be challenging to obtain. But still, FairCL-BIL also has its limitations. The most significant one is that the formulation can be infeasible in some cases. Also, how to tune the large number of hyperparameters (such as $\boldsymbol{\zeta}$) is also a problem. To address these issues, in this section, we propose a novel approach called rearrangement by loss reweighting (RRW) that has no feasibility issue and also enjoys easy hyperparameter tuning.

5.1. The Basic Formulation of FairCL-RRW

Instead of forcing fairness by incorporating fairness constraints into the learning formulation, here, we introduce weights $(w_1, \ldots, w_N)$ over individual loss functions and obtain a weighted optimization problem, which leads to the formulation of FairCL-RRW:

$$\begin{aligned} \min_{\mathbf{C},\mathbf{Q}} \quad & \sum_{i=1}^{N} w_i \ell_i(\mathbf{X}_i, \mathbf{y}_i; \mathbf{Q}, \mathbf{c}_i) \\ \text{s.t.} \quad & \mathbf{c}_i^\top \mathbf{1} = 1,\ \mathbf{c}_i \geq 0, \quad i = 1, \ldots, N. \end{aligned} \tag{6}$$

Note that, here, the weights $\boldsymbol{w}$ are not fixed, but rather iteratively updated (i.e., as shown in Algorithm 2) in order to dynamically rebalance the weights of the individual losses. This allows the individual models who have larger losses in the last iteration to be further minimized in the next iteration. The essence of RRW does not lie in the weighted form of the objective function, but in the automatic updates of the weights during the optimization process.

5.2. Automatic Reweighting by a Self-Adaptive Strategy

To make sure the iterative update of w can rebalance the individual losses and drive the optimization process more toward the improvement on the worse cases as shown in Figure 1(c), the following Lemma 1 provides a good hint about how this rebalancing process can be systematically and automatically managed by a simple self-adaptive strategy of reweighting.

Lemma 1 (Rearrangement Inequality). *Given the ascending sequences of weights $w_1 \le \cdots \le w_N$ and values $f_1 \le \cdots \le f_N$, we have*

$$\sum_{i=1}^{N} w_i f_{N-i+1} \le \sum_{i=1}^{N} w_i f_{\sigma(i)} \le \sum_{i=1}^{N} w_i f_i,$$

where $\sigma(1), \ldots, \sigma(N)$ is a random permutation of $1 \sim N$.

As our optimization problem is a minimization problem, this lemma suggests that, if we set up the weights in the same ordering as the individual losses, the minimization problem in Equation (6) forces a reordering of the individual losses in a reverse order of the weights. In other words, it forces the larger losses to be smaller in the next iteration.

Motivated by this idea, our proposed strategy is that, in each iteration t, the assigned weight w_i for each individual is an increasing function of the contribution of the individual loss to the total loss in the last iteration; that is, $w_i^t = g\left(\ell_i^{(t-1)}/\sum_{j=1}^{N} \ell_j^{(t-1)}\right) > 0$. And one of the simplest reweighting functions $g(\cdot)$ is the power function given in Algorithm 2; that is, $g(x) = x^q$. The following theorems show the rationale of using Algorithm 2 for reweighting.

Theorem 1. *FairCL-RRW reweighted by Algorithm 2 is equivalent to solving a min-max problem with the objective function* $\min_{\mathbf{C},\mathbf{Q}} \max_{w} \sum_{i=1}^{N} w_i \ell_i(\mathbf{X}_i, \boldsymbol{y}_i; \mathbf{Q}, \boldsymbol{c}_i)$.

The proof can be found in the online appendix. Theorem 1 reveals the nature of FairCL-RRW as demonstrated in Figure 1(c) that aims to improve on the worse cases. That is, larger weight is assigned to the individual with worse performance so that its impact on the overall performance is amplified in the next iteration. Note that a key element in the automatic reweighting method in FairCL-RRW is the power parameter $q (> 0)$, which regulates the balance between individual heterogeneity and commonality. The following theorem further uncovers some nature of this parameter.

Theorem 2. *When $q = 1$, the update of $\mathbf{Q}$ and $\mathbf{C}$ in Algorithm 1 aims to shrink the variance of the individual losses at the cost of the loss change incurred in the update.*

It is worthy of mentioning that, when $q \ne 1$, Algorithm 1 still shrinks the variation of the individual losses. It is just that, when $q = 1$, the shrunk variation coincides with the exact form of variance (i.e., as a quadratic form). Actually, when q is large, the algorithm pays more attention to the interindividual differences and focuses on improving the low-achieving models to achieve better equity. In contrast, when q is small, it tends to create equal weights for all individuals and treat their loss functions indifferently in optimization. In practice, we tune q through cross-validation with a grid search.

5.3. Enhance the Stability of FairCL-RRW by Temporal Smoothing (TS) Regularization

In the proof of Theorem 2 (see Online Appendix A.2), a byproduct is that we find out the objective function that Algorithm 1 aims to minimize when updating $\mathbf{Q}$ and $\mathbf{C}$ can be equivalently decomposed into three parts:

$$\mathrm{Var}^{(t+1)} + \frac{1}{N}\left(\sum_{i=1}^{N} \ell_i^{(t+1)}\right)^2 - \|\boldsymbol{\ell}^{(t)} - \boldsymbol{\ell}^{(t+1)}\|_2^2,$$

where the first term is the variance of the individual losses and the second term is proportional to the square of the basic CL loss $\sum_{i=1}^{N} \ell_i^{(t+1)}$, which makes sense. But the third term is to maximize the difference between the losses in the current iteration and the losses in the last iteration $\|\boldsymbol{\ell}^{(t)} - \boldsymbol{\ell}^{(t+1)}\|_2$. Obviously, it is undesired that $\|\boldsymbol{\ell}^{(t)} - \boldsymbol{\ell}^{(t+1)}\|_2$ gets larger because it might lead to the instability of the optimal solution if the solution drastically changes from iteration to iteration. To eliminate this potential instability, a natural idea is to consider an additional term to offset it. It is not easy to directly penalize this distance term in consideration of the complexity of $\boldsymbol{l}$. Instead, we propose a solution that is based on the following theorem.

Theorem 3. *The loss $\boldsymbol{\ell}^{(t)}$ is temporally Lipschitz continuous on $\mathbf{Q}$ and $\mathbf{C}$.*

Theorem 3 gives the upper bound of $\|\boldsymbol{\ell}^{(t)} - \boldsymbol{\ell}^{(t+1)}\|$. Thus, instead of incorporating $\|\boldsymbol{\ell}^{(t)} - \boldsymbol{\ell}^{(t+1)}\|$ directly into Formulation (6), we propose a TS regularization on the change of $\mathbf{Q}$ and $\mathbf{C}$ over time. With this temporal smoothing regularization, we can revise Formulation (6) with the following formulation:

$$\begin{aligned} \min_{\mathbf{C},\mathbf{Q}} \quad & \sum_{i=1}^{N} w_i \ell_i(\mathbf{X}_i, \boldsymbol{y}_i; \mathbf{Q}, \boldsymbol{c}_i) + \lambda_1 \|\mathbf{Q}^{(t)} - \mathbf{Q}^{(t+1)}\|_F^2 \\ & + \lambda_2 \|\mathbf{C}^{(t)} - \mathbf{C}^{(t+1)}\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{c}_i^\top \mathbf{1} = 1,\ \boldsymbol{c}_i \ge 0, \quad i = 1, \ldots, N. \end{aligned} \tag{7}$$

The regularization parameters λ_1 and λ_2 can be tuned by cross-validation. In Section 6, we show that this TS regularization strategy can lead to a great stability of the FairCL-RRW formulation.

Algorithm 1 (FairCL-RRW)
Require: Training data from $\{\{\mathbf{X}_i, \boldsymbol{y}_i\}_{i=1,\ldots,N}\}$.
Ensure: Canonical model $\mathbf{Q}$, membership $\mathbf{C}$
1: Initialize $\mathbf{Q}^{(0)}, \mathbf{C}^{(0)}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: **for** $i = 1, 2, \ldots, N$ **do**
4: Compute the individual loss $\ell_i^{(t-1)}$ for individual i given $\mathbf{Q}^{(t-1)}$, $\mathbf{C}^{(t-1)}$.
5: **end for**
6: Update $\boldsymbol{w}^{(t)} = \text{Reweight}(\ell_1^{(t-1)}, \ldots, \ell_N^{(t-1)})$.
7: Update $\mathbf{Q}^{(t)}, \mathbf{C}^{(t)}$ by minimizing (7) given $\boldsymbol{w}^{(t)}$.
8: **end for**
9: Return $\mathbf{Q}^{(t)}, \mathbf{C}^{(t)}$.

Algorithm 2 (Reweight)
Require: Individual losses $\ell_1, \ell_2 \ldots, \ell_N$, hyperparameter $q > 0$.
Ensure: Individual weights $\boldsymbol{w}$
1: **for** $i = 1, 2, \ldots, N$ **do**
2: $w_i = \left(\frac{\ell_i}{\sum_{j=1}^N \ell_j}\right)^q$
3: **end for**
4: $\boldsymbol{w} = [w_1, w_2, \ldots, w_N]$
5: Return $\boldsymbol{w}$.

5.4. Connection with the FairCL-BIL

The following theorem builds a connection between FairCL-BIL (Equation (4)) and FairCL-RRW (Equation (6)).

Theorem 4. *There exist $w_1, \ldots, w_n$ such that the optimal solution to (4) is also optimal to (6).*

This theorem shows that, in theory, the best performance of FairCL-BIL can be achieved by FairCL-RRW if the weights are appropriately set. Meanwhile, when FairCL-BIL is infeasible, FairCL-RRW can still be feasible. Therefore, we can expect that, with fine-tuned weights, FairCL-RRW could, in theory, always outperform FairCL-BIL, which is verified by our numerical results in Section 6.

5.5. Computational Efficiency

To learn personalized models of N individuals, given the number of features d and the number of canonical models K ($K < d, N$, usually set small values), whereas IL requires $N \cdot d$ parameters, FairCL methods can reduce the number of parameters to $K \cdot d + N \cdot K$. It can significantly save a space with $O(K(N + d) + N)$ compared with IL when K is far smaller than d and N. However, in each iteration, FairCL-RRW needs to deal with more parameters than the separate training in IL. Assume the sample size for each individual is m. The update of $\mathbf{Q}$ requires a cost of $O(NKmd)$, whereas the update of $\mathbf{C}$ needs $O(NKmd + NKm)$. Given that the cost of reweighting time is linear to N individuals, the total time complexity can be roughly written as $O(TKNmd)$, where T is the number of iterations. Apparently, running FairCL-RRW may incur more computational cost than IL. But given that K is usually small, the cost of each iteration should be comparable to IL. And, in practice, the algorithm can usually converge to a good solution in a few iterations. It should be mentioned that, because of the matrix operations, when the data are large, there might be some numerical issues. Currently, our methods can efficiently deal with data from hundreds of individuals, but when the data set or resulting optimization problem is too large, we might need more efficient and scalable computation strategies, which is one of our future directions.

6. Experiments

In this section, we conduct numerical evaluations of the proposed FairCL formulations using both simulated and real-world data. The real-world data includes data sets collected from three different applications. We assess the performance of the algorithms in terms of both prediction accuracy and fairness, that is, the disparity between the prediction accuracy of the individual models evaluated by variance, and the accuracy of low-achieving individual models. All algorithms are implemented in Python3 on a MacBook Pro with Apple M1 chips.

6.1. Simulation Studies

We simulate data from $N = 100$ individuals. For individual i, we randomly generate a sample size n_i by adding 5 to n_i', which is uniformly drawn from $\{1, 2, \ldots, 50\}$. Here, we set $n_i > 5$ to ensure that each individual has a minimum number of samples to enable model training. Following the setup in Feng et al. (2020), we generate the canonical models encoded in $\mathbf{Q}$ and the individual membership mappings encoded in $\mathbf{C}$ separately. In our data generation, we fix the true number of canonical models to $K = 3$. And we choose the dimension of the data d from {5, 10, 100}. Given K and d, $\mathbf{Q} = [q_{ij}]_{(d+1)\times K}$ is randomly generated from the standard normal distribution, that is, $q_{ij} \sim \mathcal{N}(0, 1)$. To generate $\mathbf{C}$, we consider two generation strategies.

- Totally random strategy: With this strategy, for each individual i, we draw $\boldsymbol{c}_i$ from a normal distribution, that is, $\boldsymbol{c}_i \sim N(\mathbf{0}, \mathbf{I}_{K\times K})$. Then, they are normalized so that $\boldsymbol{c}_i^\top \mathbf{1} = 1$ and concatenated into a matrix $\mathbf{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N]$.
- Hierarchical strategy: For each individual, first, k_i is randomly drawn from $\{1, \ldots, K\}$ by the uniform distribution, which represents that the k_ith model is chosen. Then, the individual membership $\boldsymbol{c}_i$ is drawn following $\boldsymbol{c}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{E}_{k_i})$, where $\mathbf{E}_k = \text{diag}(1, \ldots, 1, \sigma^2, 1, \ldots, 1)$ is an

almost identity matrix with the kth entry replaced by σ^2. Just like the totally random strategy, we also normalize the vectors before concatenation.

With $\mathbf{Q}$ and $\mathbf{C}$ being generated, the individual models can be easily obtained by $\boldsymbol{\beta}_i = \mathbf{Q}\boldsymbol{c}_i$. We then generate their data sets $\mathcal{D}_i = \{(\boldsymbol{x}_{ij}, y_{ij})\}_{j=1,\ldots,n_i}$. Specifically, we generate $\boldsymbol{x}_{ij} \in \mathbb{R}^d$ from a multivariate normal distribution, that is, $\boldsymbol{x}_{ij} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$. Then, for individual i, the labels can be generated following $y_{ij} = \mathrm{sgn}(\boldsymbol{\beta}_i^\top[\boldsymbol{x}_{ij}, 1] + \epsilon_{ij})$ for $j = 1, \ldots, n_i$.

6.1.1. An Overall Evaluation. We compare the models that include IL, CL, FairCL-IF, FairCL-BIL, FairCL-RRW, and FairCL-RRW-IF in Figure 2 on a toy data set generated under hierarchical strategy with d = 10. Recall that IL is the learning strategy that learns each individual's model using the individual's data alone. As we can observe, CL significantly boosts the overall performances of the models more than IL and also reduces the disparity among them. Its median accuracy is far higher than that of IL. Meanwhile, all the FairCL methods can further reduce the disparity among the performance of the individual models, achieving higher population-level average. We can also observe that, if we have the knowledge on individual similarity, we may further improve the FairCL algorithms with their IF-regularized versions. For example, FairCL-RRW+IF seems better than FairCL-RRW, and FairCL-IF is clearly better than CL.

6.1.2. The Impact of $\hat{K}$. In what follows, we present more experimental results to show how the various hyperparameters impact the results. We first show how CL perform differently from IL for different $\hat{K}$s (recall that the true K = 3) under the hierarchical data-generation strategy with d = 10 fixed. The histograms of the distributions of the model accuracy are shown in Figure 3. When $\hat{K} = 3$ (which is the true number of canonical models), one can easily see that CL outperforms IL as the entire distribution of the accuracy of the individual models of CL is more compact and both the mean and peak of the distribution are larger than the ones of IL. CL also has fewer low-performance individual models and more high-performance individual models. As $\hat{K}$ increases to five (which is a little larger than the true value), CL still outperforms IL. However, when $\hat{K} = 10$, CL seems to downgrade. The peak of CL shifts to the left of the peak of IL, and more low-performance individual models occur. This result shows that a proper selection of the parameter $\hat{K}$ is important for the success of CL, which is consistent with the observations in prior works (Feng et al. 2020). In practice, this parameter can also be effectively identified using the Akaike information criterion/Bayes information criterion and cross-validation (Feng et al. 2020).

The comparisons between FairCL-BIL and FairCL-RRW with CL are shown in Online Figures S1 and S2 (see Online Appendix C.1). As we mentioned before, CL itself doesn't focus on the fairness issue though it does lead to more fairness compared with IL. Compared with FairCL-BIL and FairCL-RRW, the performances of the individual models of CL are more disparate. We can also observe that the advantage of FairCL-BIL will not sustain if there is serious misspecification of the parameter $\hat{K}$; that is, when K = 10, the gap between FairCL-BIL and CL seems to shrink and is not more significant. Different from FairCL-BIL, FairCL-RRW outperforms CL for all $\hat{K}$.

6.1.3. The Impact of d. The above results show the performance of the models with d = 10. When d is increased to 100, the results are shown in Figure 4 (with $\hat{K} = 3$). It can be seen that CL still outperforms IL. And the superiority of FairCL-BIL and FairCL-RRW over CL is even more significant than it is when d = 10. So is CL more than IL. As the sample size of the individual data are relatively small given such a large d, it is foreseeable that IL

Figure 2. (Color online) Comparison of Different Methods (d = 10) (From Left to Right: IL, CL, FairCL-IF, FairCL-BIL, FairCL-RRW, FairCL-RRW+IF)

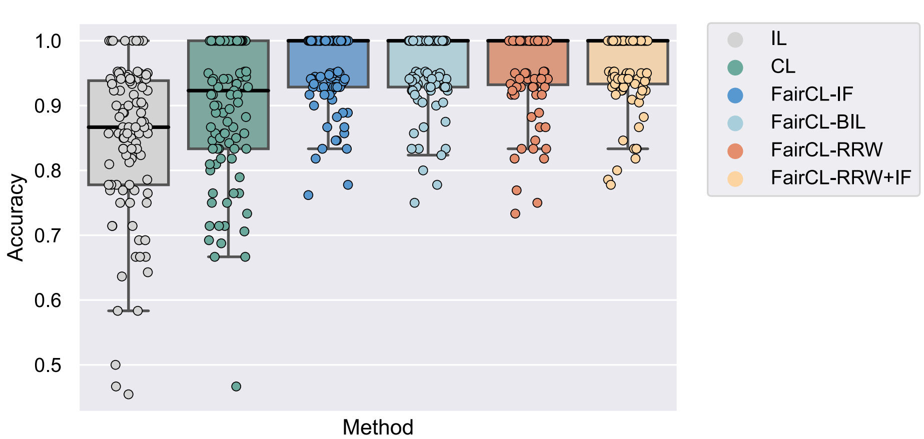
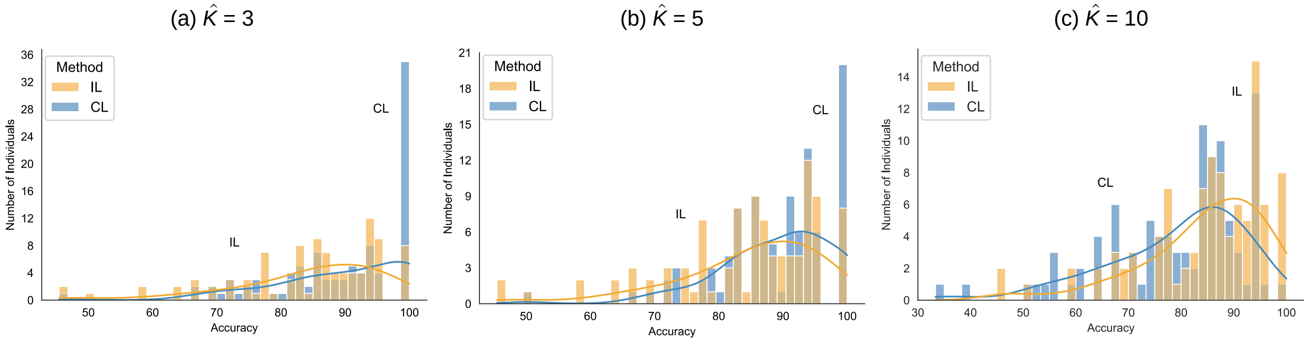


Figure 3. (Color online) Comparison of IL and CL on Simulated Data Under a Hierarchical Strategy (d = 10)



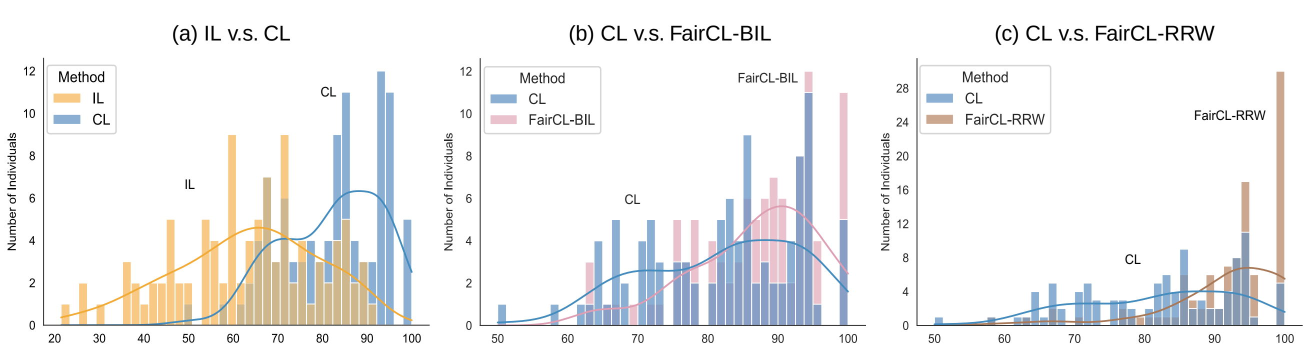
couldn't perform well, but CL-based approaches really help the individuals model to learn collectively and reinforce each other.

6.1.4. The Impact of IF Regularization. To evaluate the effect of further incorporating individual similarity information into the CL-based models, we conduct more experiments, and some results are shown in Online Appendix C.2 (with $d = 10$). It can be observed in Online Figure S3(a) that the IF regularization can significantly improve CL's performance with the corresponding $\hat{K}$ even when we set $\hat{K} = 10$ that quite deviates from the true value $K = 3$. Now, we fix $\hat{K}$ to this true value and explore the performance when d is changed. Obviously, regardless of d, FairCL-IF outperforms CL to a great extent as is shown in Online Figure S4(a) and Figure 5(a). For FairCL-RRW, IF regularization can also shift the distributions more toward the right side as shown in Online Figure S4(b). But the improvement may not be that significant as FairCL-RRW has already shown great performance. IF regularization is not necessarily always useful in improving fairness. As Fleisher (2021) argues, the IF principle that requires similar treatment on similar individuals can't always guarantee fairness in the population. Sometimes, it may even degrade the performance of some FairCL methods. An example of such degradation can be seen in Figure 5(b).

6.1.5. In-Depth Analysis of the Detailed Experimental Results. To reach a more insightful understanding of FairCL algorithms, we further examine some detailed statistics of the models' accuracy in Tables 1–3. In these tables, we not only present the mean test accuracy, but also the best 10% and the worst 10% test accuracy. FairCL-RRW can always beat IL, CL, and FairCL-BIL in the worst 10% mean accuracy and the overall variance for whatever $\hat{K}$ (see the underlined results). We also include experiment results from additional methods: group contribution matching (GCM) (Li et al. 2023), postprocess for α-DP (Xian et al. 2023), and subgroup mixup (SGMixup) (Zhang et al. 2017, Navarro et al. 2024). These three methods represent three types of fair-aware machine learning approaches, that is, matching, postprocessing and data augmentation. From the tables, it can be observed that our FairCL-RRW still outperformed these methods.

When IF regularization is added, FairCL-IF improves upon CL and can usually achieve high mean performance. Similarly, FairCL-RRW+IF might also improve upon FairCL-RRW for some cases. But we can also observe that IF regularization doesn't necessarily shrink the variance as we compare the variances of FairCL-RRW+IF with FairCL-RRW, especially when model complexity increases. Looking at the statistics of different segments of the population, we see that the way FairCL-RRW works is to enhance the performance of

Figure 4. (Color online) Comparison of Test Accuracy Distributions on Simulated Data Under a Hierarchical Strategy Without Prior Knowledge ($\hat{K} = 3, d = 100$)



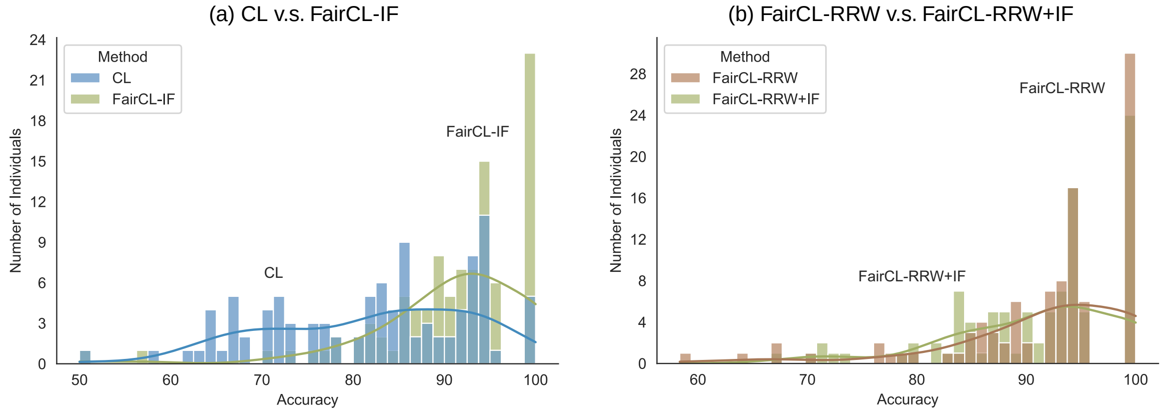


Figure 5. (Color online) Comparison of Test Accuracy Distributions on Simulated Data Under a Hierarchical Strategy with Prior Knowledge on IF ($\hat{K} = 3, d = 100$)

low-achieving individual models (i.e., evidenced by the higher mean accuracy of the worst 10%), which, in turn, further increases the mean accuracy of all the models and reduces the variance. Note that these conclusions are drawn from experiments using the hierarchical strategy. Similar conclusions can be obtained if we simulate data using the totally random strategy (see Online Appendix D).

6.1.6. The Impact of TS Regularization on Computational Stability.

Figure 6 illustrates that the TS-regularized FairCL-RRW behaves more stably compared with the nonregularized FairCL-RRW algorithm. It can be observed that the TS-regularized FairCL-RRW can converge quicker and maintains its stable performance over iterations with a good prediction performance and low variance among the individual models on the testing data set. Note that this experiment is done on a simulated data set generated under the totally random strategy with $K = 3, d = 10$, but overall, this is a commonly observed phenomenon on other data sets in our experiments.

6.2. Real-World Studies

We further evaluate the FairCL models using data sets collected from three different sectors, including transportation, manufacturing, and healthcare. Similarly, we compare the FairCL methods with not only IL and CL, but also three fairness-aware learning methods, that is, GCM, postprocess and SGMixup-augmented CL.

6.2.1. Transportation Demand Management.

We started the real-world experiments from a real-world data set collected from a personalized transportation demand

Table 1. Statistics of the Models' Accuracy on Simulated Data Under a Hierarchical Strategy ($d = 5$)

$\hat{K}$	Methods	Mean	Best 10%	Worst 10%	Variance
	IL	88.56	100.00	64.28	113.57
	GCM	74.75	93.79	43.47	222.21
	Postprocess	78.11	97.89	44.96	234.70
3	CL	93.91	100.00	73.48	84.28
	SGMixup+CL	93.51	100.00	74.46	72.24
	FairCL-BIL	94.74	100.00	76.42	67.68
	FairCL-RRW	<u>95.99</u>	100.00	<u>80.18</u>	<u>48.30</u>
	FairCL-IF	95.61	100.00	79.30	50.10
	FairCL-RRW+IF	**96.10**	100.00	**81.26**	**42.21**
5	CL	92.18	100.00	72.80	88.93
	SGMixup+CL	92.53	100.00	73.02	79.48
	FairCL-BIL	93.24	100.00	74.60	62.66
	FairCL-RRW	<u>**95.92**</u>	100.00	<u>**80.20**</u>	<u>**47.54**</u>
	FairCL-IF	95.08	100.00	77.43	57.86
	FairCL-RRW+IF	95.62	100.00	78.31	50.58
10	CL	89.86	100.00	67.99	92.25
	SGMixup+CL	88.76	100.00	68.48	105.71
	FairCL-BIL	90.74	100.00	70.44	88.84
	FairCL-RRW	<u>**95.03**</u>	100.00	<u>**78.25**</u>	<u>**56.98**</u>
	FairCL-IF	94.47	100.00	76.24	66.72
	FairCL-RRW+IF	94.85	100.00	77.76	59.86

Notes. Bold results are the best results among all methods for corresponding parameters. Underlined results are the best results without prior knowledge on IF.

Table 2. Statistics of the Models' Accuracy on Simulated Data Under a Hierarchical Strategy (d = 10)

$\hat{K}$	Methods	Mean	Best 10%	Worst 10%	Variance
	IL	83.94	100.00	57.38	149.42
	GCM	71.55	93.95	44.69	192.24
	Postprocess	77.82	96.46	56.96	132.17
3	CL	92.29	100.00	72.95	83.30
	SGMixup+CL	92.86	100.00	73.75	74.65
	FairCL-BIL	94.34	100.00	77.25	62.45
	FairCL-RRW	95.44	100.00	83.39	34.64
	FairCL-IF	95.12	100.00	84.17	32.75
	FairCL-RRW+IF	**96.38**	100.00	**85.62**	**25.27**
5	CL	91.94	100.00	69.78	89.65
	SGMixup+CL	92.03	100.00	70.67	81.64
	FairCL-BIL	93.66	100.00	76.01	68.44
	FairCL-RRW	**96.28**	100.00	**81.49**	**42.72**
	FairCL-IF	95.50	100.00	78.30	60.04
	FairCL-RRW+IF	95.61	100.00	79.81	52.30
10	CL	89.57	100.00	67.05	94.43
	SGMixup+CL	85.78	100.00	63.91	124.06
	FairCL-BIL	90.27	100.00	70.76	84.68
	FairCL-RRW	**94.54**	100.00	**78.75**	**49.40**
	FairCL-IF	93.56	100.00	71.73	78.73
	FairCL-RRW+IF	93.96	100.00	77.14	55.61

Notes. Bold results are the best results among all methods for corresponding parameters. Underlined results are the best results without prior knowledge on IF.

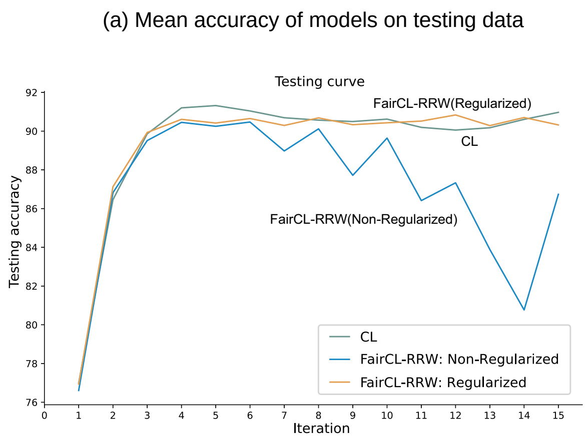
management (TDM) system (Zhu et al. 2019). This TDM data set studies travel behaviors in a reward system and consists of 828 participants, whereas each participant has 13 observations. The number of attributes in each travel alternative is four, which includes early schedule delay, late schedule delay, travel time saving, and reward points offered. And the participants make decisions on whether to accept the travel alternative given the rewards, which leads to binary responses. We randomly split each individual's data into training and testing sets by a ratio of 7:6. Following Feng et al. (2020), we use threefold cross-validation to determine the number of canonical models K using the training data set.

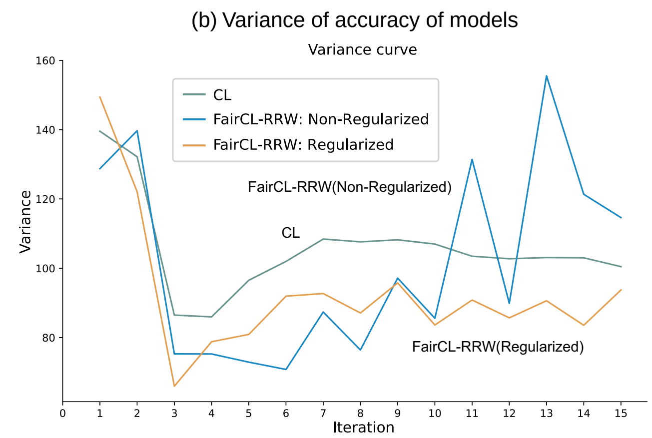
Table 3. Statistics of the Models' Accuracy on Simulated Data Under a Hierarchical Strategy (d = 100)

$\hat{K}$	Methods	Mean	Best 10%	Worst 10%	Variance
	IL	63.17	84.00	38.44	161.63
	GCM	70.64	90.45	43.80	180.98
	Postprocess	74.33	95.08	51.26	161.32
3	CL	88.26	100.00	65.26	113.35
	SGMixup+CL	86.41	100.00	65.73	108.10
	FairCL-BIL	89.01	100.00	70.48	87.98
	FairCL-RRW	**92.87**	100.00	**77.08**	**50.32**
	FairCL-IF	89.74	100.00	73.57	66.38
	FairCL-RRW +IF	90.35	100.00	73.89	59.29
5	CL	82.74	100.00	58.16	143.19
	SGMixup+CL	81.53	99.00	60.44	120.82
	FairCL-BIL	80.32	97.50	53.41	166.51
	FairCL-RRW	**89.11**	100.00	**70.17**	**89.41**
	FairCL-IF	85.80	100.00	58.51	157.89
	FairCL-RRW +IF	87.94	100.00	68.25	98.60
10	CL	75.76	96.04	47.22	196.72
	SGMixup+CL	74.66	91.55	43.83	190.85
	FairCL-BIL	74.87	94.49	50.27	160.43
	FairCL-RRW	**85.78**	100.00	**65.69**	**104.01**
	FairCL-IF	78.94	97.34	50.96	178.33
	FairCL-RRW +IF	84.39	100.00	58.47	151.50

Notes. Bold results are the best results among all methods for corresponding parameters. Underlined results are the best results without prior knowledge on IF.

Figure 6. (Color online) The Impact of TS Regularization on FairCL-RRW





Results of the different FairCL models are shown in Table 4. As the data dimension is small and the sample size is not large, although IL doesn't perform very well on average and the mean accuracy of the worst 10% is quite low, it still achieves a mean accuracy of more than 70%. CL performs better than not only IL, but also GCM and postprocess, and its performance can be improved by SGMixup, FairCL-BIL, and FairCL-RRW. Particularly, FairCL-RRW reaches the smallest variance of accuracy, achieving a better mean performance. It is also easy to see that the mean accuracy of the worst 10% is much enhanced by FairCL-RRW compared with other methods. We further try to incorporate IL regularization, but because we don't have readily available prior knowledge of the individuals' similarity, here, following Feng et al. (2020), the individual similarity is computed based on the radial basis function kernel; that is, the similarity between individual i and j is defined as $s_{ij} = \exp(-\|x_i - x_j\|_2^2/\sigma^2)$, where x represents the individual attributes. It can be observed that FairCL-IF can achieve smaller variance than FairCL-BIL and higher mean accuracy of the overall population as well as the worst 10% but is worse than FairCL-RRW. This is expected because the similarity information is not accurate enough.

Table 4. Statistics of the Models' Accuracy on the TDM Data Set

Methods	Mean	Best 10%	Worst 10%	Variance
IL	72.36	100.00	30.38	498.37
GCM	64.77	100.00	16.11	605.85
Postprocess	67.19	100.00	24.72	487.69
CL	74.32	100.00	35.33	454.33
SGMixup+CL	75.05	100.00	36.11	474.25
FairCL-BIL	75.17	100.00	38.33	413.06
FairCL-RRW	<u>76.92</u>	100.00	<u>42.22</u>	<u>412.91</u>
FairCL-IF	75.32	100.00	41.47	417.12
FairCL-RRW+IF	**77.76**	100.00	**42.69**	**395.92**

Notes. Bold results are the best results among all methods for corresponding parameters. Underlined results are the best results without prior knowledge on IF.

6.2.2. Turbofan Engine Degradation. Another data set we use in our numerical studies is the run-to-failure degradation trajectories of turbofan engines. The degradation simulation was carried out by the NASA commercial modular aero-propulsion system simulation (C-MAPSS) dynamical model (Saxena et al. 2008). This data set includes continuous observations of 90 different turbofan engines collected during multiple operational cycles before failure in which each observation has 18 discriminative sensor measurements. And two failure modes, including high-pressure compressor degradation and fan degradation, were considered in this data set. We evaluate the ability of our methods to predict the failure of the engines indicated by the number of operation cycles. That is, if an engine is operated more than 150 cycles, it is labeled as being at higher risk of failure. This gives rise to a binary classification task (Loyer et al. 2014, Aribi et al. 2023).

We compare FairCL-BIL and FairCL-RRW with IL, GCM, postprocess, CL, and SGMixup-augmented CL. Because of the lack of prior knowledge, IF-regularized versions of our methods are not taken into consideration in the experiments. The results are reported in Table 5. Given the varying data sizes and data quality for different engines, IL can show extremely poor prediction performance on some individuals. This results in a low test accuracy on the worst 10%, indicating significant unfairness. GCM and postprocess significantly outperform IL on both accuracy and fairness, but they perform worse than CL. Though SGMixup can further shrink the variance, which indicates better fairness, our FairCL-BIL and FairCL-RRW can improve the performance of CL to a larger extent. Particularly, FairCL-RRW demonstrates the highest mean accuracy, achieving the best performance fairness (highest worst 10% accuracy, lowest accuracy variance).

6.2.3. Surgical Site Infections. We further test the methods using a data set from healthcare that concerns

Table 5. Statistics of the Models' Performance on the CMAPSS Data Set

Methods	Mean	Best 10%	Worst 10%	Variance
IL	77.61	100.00	15.78	784.57
GCM	82.55	100.00	28.48	563.01
Postprocess	80.36	100.00	35.90	525.45
CL	83.56	100.00	36.23	525.67
SGMixup+CL	83.79	100.00	37.48	496.41
FairCL-BIL	84.38	100.00	37.91	511.72
FairCL-RRW	**86.34**	100.00	**42.39**	**417.07**

Note. Bold results are the best results among all methods for corresponding parameters.

surgical site infections (SSIs). SSI is one of the most common types of hospital-acquired infections, which are usually associated with increasing morbidity, prolonged hospitalization, and higher mortality (Costabella et al. 2023). The heterogeneous conditions of patients and the high-stake situation call for personalized prediction models that can treat different patients fairly and give accurate diagnosis of SSI. Given this context, we conducted experiments on an SSI data set that involves 259 patients whose wounds are continuously monitored in a range of 7–20 days. Each observation consists of 29 wound measurements, such as wound color, wound temperature, and pain intensity, etc., together with a binary indicator showing the infection status marked by experts. Each individual's data are randomly split into training and testing sets with a ratio of 4:3. And similar to the experiments in the C-MAPSS data set, we didn't consider IF regularized FairCL methods in comparison because of a lack of prior knowledge of the patients.

Table 6 shows our experimental results. Whereas the performance of IL is poor, others perform much better with mean accuracy greater than 90%. Among these methods, FairCL-RRW achieves a mean accuracy much higher than all the other methods. When it comes to fairness, GCM, postprocess, and CL all outperform IL. And we can easily observe the significant improvement on the worst 10% accuracy by SGMixup, FairCL-BIL, and FairCL-RRW compared with CL. Compared with SGMixup, FairCL-BIL can improve the worst 10%

Table 6. Statistics of the Models' Performance on the SSI Data Set

Methods	Mean	Best 10%	Worst 10%	Variance
IL	74.27	100.00	22.37	645.17
GCM	91.34	100.00	57.18	231.32
Postprocess	92.59	100.00	55.38	196.40
CL	92.94	100.00	59.29	190.65
SGMixup+CL	93.27	100.00	62.05	169.05
FairCL-BIL	92.95	100.00	63.17	162.83
FairCL-RRW	**98.14**	100.00	**81.44**	**59.06**

Note. Bold results are the best results among all methods for corresponding parameters.

accuracy and reduce the variance to a larger extent. And, apparently, FairCL-RRW is superior to other methods as it improves the model accuracy of those low-achieving models and leads to a small variance. It can achieve the best performance on both accuracy and fairness.

7. Conclusion

CL has been developed in the literature to learn personalized models for a heterogeneous population of individuals. In this work, we further develop the FairCL framework by incorporating fairness constraints into the CL framework and developing the corresponding computational algorithms. We further develop a self-adaptive FairCL formulation, named FairCL-RRW, that highlights a dynamic reweighting approach to improve fairness of the individual models. We conduct theoretical analysis to reveal the connection of this novel reweighting method with other FairCL formulations and the underlying rationale as to why it can reduce the variance of the performances of the individual models. Extensive numerical studies using both simulated and real-world data sets also show the advantages of the FairCL methods (particularly, FairCL-RRW) that can improve not only fairness, but also the accuracy of the individual models. As the proposed FairCL shows its potential in generalizing to different applications (transportation, healthcare, and manufacturing, etc.), remember that, ultimately, the success of model personalization requires trustworthiness between individuals and the system. To build trustworthy personalized models, beyond fairness, there are also important issues such as privacy, robustness to adversarial tasks, etc., that haven't been fully explored in this paper. Thus, we further extend our FairCL framework in the future in the following directions: (i) Adversarial robustness: although FairCL can somewhat show resistance to data scarcity or noise, how different adversarial attacks (Koh and Liang 2017, Solans et al. 2020, Mehrabi et al. 2021a) affect its performance and how to defend against specific attacks still need further investigation. (ii) Privacy: as there are growing concerns on data security and privacy in machine learning, it is worth exploring how to enable privacy-preserving information sharing in FairCL so as to improve its generalization to problems with privacy concerns. One possible way to do so is combining it with federated learning, which awaits more investigation in the future. (iii) Scalability: more efficient updating algorithms in the optimization of FairCL can provide more scalability when the number of individuals becomes tens of thousands or even more than that and when the data size also correspondingly becomes larger.

Acknowledgments

The authors thank the editors and the anonymous referees for insightful comments and constructive suggestions. This work was supported by the Breakthrough T1D award 2-SRA-2022-

1259-S-B. The authors also would like to thank Dr. Cynthia Chen and her team for sharing the TDM data.

References

Abul-Husn NS, Kenny EE (2019) Personalized medicine and the power of electronic health records. *Cell* 177(1):58–69.

Agarwal A, Dudík M, Wu ZS (2019) Fair regression: Quantitative definitions and reduction-based algorithms. *Internat. Conf. Machine Learn.* (PMLR, New York), 120–129.

Alabi D, Immorlica N, Kalai A (2018) Unleashing linear optimizers for group-fair learning and optimization. *31st Annual Conf. Learn. Theory* (PMLR, New York), 2043–2066.

Alfian G, Syafrudin M, Ijaz MF, Syaekhoni MA, Fitriyani NL, Rhee J (2018) A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing. *Sensors (Basel)* 18(7):2183.

Amara S, Subramanian RR (2020) Collaborating personalized recommender system and content-based recommender system using textcorpus. *Internat. Conf. Advanced Comput. Comm. Systems* (IEEE, Piscataway, NJ), 105–109.

Andreu-Perez J, Leff DR, Ip HM, Yang GZ (2015) From wearable sensors to smart implants—Toward pervasive and personalized healthcare. *IEEE Trans. Biomedical Engrg.* 62(12):2750–2762.

Aribi Y, Boutora S, Zamoum PRB, Menasria H, Abdellaoui A, Kouzou PA (2023) Turbofan engine RUL prediction using ICA and machine learning algorithms. *20th Internat. Multi-Conf. Systems Signals Devices* (IEEE, Piscataway, NJ), 477–484.

Ban GY, Keskin NB (2021) Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Sci.* 67(9):5549–5568.

Belkin M, Niyogi P, Sindhwani V (2005) On manifold regularization. *Internat. Workshop Artificial Intelligence Statist.* (PMLR, New York), 17–24.

Benhamdi S, Babouri A, Chiky R (2017) Personalized recommender system for e-learning environment. *Ed. Inform. Tech.* 22(4):1455–1477.

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2021) Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* 50(1):3–44.

Bilenko M, Richardson M (2011) Predictive client-side profiles for personalized advertising. *Proc. 17th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 413–421.

Briggs C, Fan Z, Andras P (2020) Federated learning with hierarchical clustering of local updates to improve training on non-IID data. *Internat. Joint Conf. Neural Networks* (IEEE, Piscataway, NJ), 1–9.

Caldas S, Konečný J, McMahan HB, Talwalkar A (2018) Expanding the reach of federated learning by reducing client resource requirements. Preprint, submitted December 18, https://arxiv.org/abs/1812.07210.

Caton S, Haas C (2020) Fairness in machine learning: A survey. Preprint, submitted October 4, https://arxiv.org/abs/2010.04053.

Chen J, Kallus N, Mao X, Svacha G, Udell M (2019) Fairness under unawareness: Assessing disparity when protected class is unobserved. *Proc. Conf. Fairness Accountability Transparency* (ACM, New York), 339–348.

Chi J, Tian Y, Gordon GJ, Zhao H (2021) Understanding and mitigating accuracy disparity in regression. *Internat. Conf. Machine Learn.* (PMLR, New York), 1866–1876.

Chzhen E, Denis C, Hebiri M, Oneto L, Pontil M (2020) Fair regression with Wasserstein barycenters. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates, Inc., Red Hook, NY), 7321–7331.

Corbett-Davies S, Goel S (2018) The measure and mismeasure of fairness: A critical review of fair machine learning. Preprint, submitted July 31, https://arxiv.org/abs/1808.00023.

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. *Proc. 23rd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 797–806.

Costabella F, Patel KB, Adepoju AV, Singh P, Mahmoud HAH, Zafar A, Patel T, et al. (2023) Healthcare cost and outcomes associated with surgical site infection and patient outcomes in low-and middle-income countries. *Cureus* 15(7):e42493.

Cotter A, Jiang H, Gupta MR, Wang S, Narayan T, You S, Sridharan K (2019) Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *J. Machine Learn. Res.* 20(172):1–59.

Cudeck R (1996) Mixed-effects models in the study of individual differences with repeated measures data. *Multivariate Behav Res.* 31(3):371–403.

Du W, Xu D, Wu X, Tong H (2021) Fairness-aware agnostic federated learning. *Proc. 2021 SIAM Internat. Conf. Data Mining* (SIAM, Philadelphia), 181–189.

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. *Proc. 3rd Innovations Theoret. Comput. Sci. Conf.* (ACM, New York), 214–226.

Emmert-Streib F, Dehmer M (2018) A machine learning perspective on personalized medicine: An automized, comprehensive knowledge base with ontology for pattern recognition. *Machine Learn. Knowledge Extraction* 1(1):149–156.

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. *Proc. 21st ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 259–268.

Feng J, Zhu X, Wang F, Huang S, Chen C (2020) A learning framework for personalized random utility maximization (RUM) modeling of user behavior. *IEEE Transactions Automation Sci. Engrg.* 19(1):510–521.

Firdaus M, Shandilya A, Ekbal A, Bhattacharyya P (2022) Being polite: Modeling politeness variation in a personalized dialog agent. *IEEE Trans. Comput. Soc. Systems* 10(4):1455–1464.

Fleisher W (2021) What's fair about individual fairness? *Proc. 2021 AAAI/ACM Conf. AI Ethics Soc.* (Association for Computing Machinery, New York).

Gao D, Liu Y, Zeng D (2022) Non-asymptotic properties of individualized treatment rules from sequentially rule-adaptive trials. *J. Machine Learn. Res.* 23(250):1–42.

Gao L, Li L, Chen Y, Zheng W, Xu C, Xu M (2021) FIFL: A fair incentive mechanism for federated learning. *Proc. 50th Internat. Conf. Parallel Processing* (ACM, New York), 1–10.

Ghosh A, Chung J, Yin D, Ramchandran K (2020) An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates, Inc., Red Hook, NY), 19586–19597.

Ghosh A, Hong J, Yin D, Ramchandran K (2019) Robust federated learning in a heterogeneous environment. Preprint, submitted June 16, https://arxiv.org/abs/1906.06629.

Guo L, Jin J, Zhang H, Zheng Z, Yang Z, Xing Z, Pan F, et al. (2021) We know what you want: An advertising strategy recommender system for online advertising. *Proc. 27th ACM SIGKDD Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 2919–2927.

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 29 (Curran Associates, Inc., Red Hook, NY).

He Z, Wang L, Cai Z (2023) Clustered federated learning with adaptive local differential privacy on heterogeneous IoT data. *IEEE Internet Things J.* 11(1):137–146.

Hu S, Wu ZS, Smith V (2022) Provably fair federated learning via bounded group loss. Preprint, submitted March 18, https://arxiv.org/abs/2203.10190.

Huang T, Lin W, Shen L, Li K, Zomaya AY (2022) Stochastic client selection for federated learning with volatile clients. *IEEE Internet Things J.* 9(20):20055–20070.

Huang T, Lin W, Wu W, He L, Li K, Zomaya AY (2020) An efficiency-boosting client selection scheme for federated learning with fairness guarantee. *IEEE Trans. Parallel Distributed Systems* 32(7):1552–1564.

Jiang M, Wang Z, Guo S, Gao X, Tan KC (2020) Individual-based transfer learning for dynamic multiobjective optimization. *IEEE Trans. Cybernetics* 51(10):4968–4981.

Jiang Z, Han X, Fan C, Yang F, Mostafavi A, Hu X (2022) Generalized demographic parity for group fairness. *10th Internat. Conf. Learn. Representations* (OpenReview.net).

Kamp M, Fischer J, Vreeken J (2023) Federated learning from small datasets. *Internat. Conf. Learn. Representations* (OpenReview.net).

Knisely BM, Vaughn-Cooke M, Wagner LA, Fink JC (2021) Device personalization for heterogeneous populations: Leveraging physician expertise and national population data to identify medical device patient user groups. *User Modeling User-Adapted Interaction* 31(5):979–1025.

Koh PW, Liang P (2017) Understanding black-box predictions via influence functions. *Internat. Conf. Machine Learn.* (PMLR, New York), 1885–1894.

Kulkarni V, Kulkarni M, Pant A (2020) Survey of personalization techniques for federated learning. *Fourth World Conf. Smart Trends Systems Security Sustainability (London)*, 794–797.

Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., Red Hook, NY).

Laird NM, Ware JH (1982) Random-effects models for longitudinal data. *Biometrics* 38(4):963–974.

Li T, Hu S, Beirami A, Smith V (2021) Ditto: Fair and robust federated learning through personalization. *Internat. Conf. Machine Learn.* (PMLR, New York), 6357–6368.

Li T, Sanjabi M, Beirami A, Smith V (2020a) Fair resource allocation in federated learning. *10th Internat. Conf. Learn. Representations* (OpenReview.net).

Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V (2020b) Federated optimization in heterogeneous networks. *Proc. Machine Learn. Systems*, vol. 2, 429–450.

Li T, Li Z, Li A, Du M, Liu A, Guo Q, Meng G, Liu Y (2023) Fairness via group contribution matching. *Proc. 32nd Internat. Joint Conf. Artificial Intelligence* (International Joint Conferences on Artificial Intelligence), 436–445.

Lin Y, Liu S, Huang S (2018) Selective sensing of a heterogeneous population of units with dynamic health conditions. *IISE Trans.* 50(12):1076–1088.

Lin Y, Liu K, Byon E, Qian X, Huang S (2015) Domain-knowledge driven cognitive degradation modeling for Alzheimer's disease. *Proc. 2015 SIAM Internat. Conf. Data Mining* (SIAM, Philadelphia), 721–729.

Lin Y, Liu K, Byon E, Qian X, Liu S, Huang S (2017) A collaborative learning framework for estimating many individualized regression models in a heterogeneous population. *IEEE Trans. Reliability* 67(1):328–341.

Liu J, Huang J, Zhou Y, Li X, Ji S, Xiong H, Dou D (2022) From distributed machine learning to federated learning: A survey. *Knowledge Inform. Systems* 64(4):885–917.

Liu Q, Tong S, Liu C, Zhao H, Chen E, Ma H, Wang S (2019) Exploiting cognitive structure for adaptive learning. *Proc. 25th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 627–635.

Loyer JL, Henriques E, Wiseall S (2014) Comparison of binary classifiers for data-driven prognosis of jet engines health. *PHM Soc. Eur. Conf.*, vol. 2 (PHM Society).

Lyu L, Yu J, Nandakumar K, Li Y, Ma X, Jin J, Yu H, Ng KS (2020) Toward fair and privacy-preserving federated deep models. *IEEE Trans. Parallel Distributed Systems* 31(11): 2524–2541.

Ma Z, Dou Z, Zhu Y, Zhong H, Wen JR (2021) One chatbot per person: Creating personalized chatbots based on implicit user profiles. *Proc. 44th Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 555–564.

Mansour Y, Mohri M, Ro J, Suresh AT (2020) Three approaches for personalization with applications to federated learning. Preprint, submitted February 25, https://arxiv.org/abs/2002.10619.

Marfoq O, Neglia G, Bellet A, Kameni L, Vidal R (2021) Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, vol. 34 (Curran Associates Inc., Red Hook, NY), 15434–15447.

McAuley J (2022) *Personalized Machine Learning* (Cambridge University Press, Cambridge, UK).

Mehrabi N, Naveed M, Morstatter F, Galstyan A (2021a) Exacerbating algorithmic bias through fairness attacks. *Proc. Conf. AAAI Artificial Intelligence*, vol. 35 (AAAI Press, Palo Alto, CA), 8930–8938.

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021b) A survey on bias and fairness in machine learning. *ACM Comput. Surveys* 54(6):1–35.

Mehta M, Shao C (2023) A greedy agglomerative framework for clustered federated learning. *IEEE Trans. Indust. Informatics* 19(12):11856–11867.

Miao X, Chu CT, Tang L, Zhou Y, Young J, Bhasin A (2015) Distributed personalization. *Proc. 21th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 1989–1998.

Mohri M, Sivek G, Suresh AT (2019) Agnostic federated learning. *Internat. Conf. Machine Learn.* (PMLR, New York), 4615–4625.

Morafah M, Vahidian S, Wang W, Lin B (2023) FLIS: Clustered federated learning via inference similarity for non-IID data distribution. *IEEE Open J. Comput. Soc.* 4:109–120.

Naumov M, Mudigere D, Shi HJM, Huang J, Sundaraman N, Park J, Wang X, et al. (2019) Deep learning recommendation model for personalization and recommendation systems. Preprint, submitted May 31, https://arxiv.org/abs/1906.00091.

Navarro M, Little C, Allen GI, Segarra S (2024) Data augmentation via subgroup mixup for improving fairness. *2024 IEEE Internat. Conf. Acoustics Speech Signal Processing* (IEEE, Piscataway, NJ), 7350–7354.

Pessach D, Shmueli E (2020) Algorithmic fairness. Preprint, submitted January 21, https://arxiv.org/abs/2001.09784.

Ruan Y, Joe-Wong C (2022) Fedsoft: Soft clustered federated learning with proximal local updating. *Proc. Conf. AAAI Artificial Intelligence*, vol. 36 (AAAI Press, Palo Alto, CA), 8124–8131.

Sattler F, Müller KR, Samek W (2020) Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Trans. Neural Networks Learning Systems* 32(8):3710–3722.

Saxena A, Goebel K, Simon D, Eklund N (2008) Damage propagation modeling for aircraft engine run-to-failure simulation. *Internat. Conf. Prognostics Health Managemen*t (IEEE, Piscataway, NJ), 1–9.

Shamir O, Srebro N (2014) Distributed stochastic optimization and learning. *52nd Annual Allerton Conf. Comm. Control Comput. (Allerton)* (IEEE, Piscataway, NJ), 850–857.

Shi Y, Yu H, Leung C (2023a) Toward fairness-aware federated learning. *IEEE Trans. Neural Networks Learning Systems* 35(9): 11922–11938.

Shi Y, Liu Z, Shi Z, Yu H (2023b) Fairness-aware client selection for federated learning. *IEEE Internat. Conf. Multimedia Expo* (IEEE, Piscataway, NJ), 324–329.

Solans D, Biggio B, Castillo C (2020) Poisoning attacks on algorithmic fairness. *Joint Eur. Conf. Machine Learn. Knowledge Discovery Databases* (Springer, Cham, Switzerland), 162–177.

Song Z, Sun H, Yang HH, Wang X, Zhang Y, Quek TQ (2021) Reputation-based federated learning for secure wireless networks. *IEEE Internet Things J.* 9(2):1212–1226.

Tan AZ, Yu H, Cui L, Yang Q (2022) Toward personalized federated learning. *IEEE Trans. Neural Networks Learning Systems* 34(12): 9587–9603.

Tobore I, Li J, Yuhang L, Al-Handarish Y, Kandwal A, Nie Z, Wang L, et al. (2019) Deep learning intervention for healthcare challenges: Some biomedical domain considerations. *JMIR Mhealth Uhealth* 7(8):e11966.

Verbraeken J, Wolting M, Katzy J, Kloppenburg J, Verbelen T, Rellermeyer JS (2020) A survey on distributed machine learning. *ACM Comput. Surveys* 53(2):1–33.

Xian R, Yin L, Zhao H (2023) Fair and optimal classification via post-processing. *Internat. Conf. Machine Learn.* (PMLR, New York), 37977–38012.

Xu G, Jia G, Shi L, Zhang Z (2021) Personalized course recommendation system fusing with knowledge graph and collaborative filtering. *Comput. Intelligence Neuroscience* 2021(1):9590502.

Yang M, Wang X, Zhu H, Wang H, Qian H (2021) Federated learning with class imbalance reduction. *29th Eur. Signal Processing Conf.* (IEEE, Piscataway, NJ), 2174–2178.

Yang Q, Liu Y, Chen T, Tong Y (2019) Federated machine learning: Concept and applications. *ACM Trans. Intelligent Systems Tech.* 10(2):1–19.

Yang Q, Zhang Y, Dai W, Pan SJ (2020) *Transfer Learning* (Cambridge University Press, Cambridge, UK).

Yoganarasimhan H (2020) Search personalization using machine learning. *Management Sci.* 66(3):1045–1070.

Yu H, Liu Z, Liu Y, Chen T, Cong M, Weng X, Niyato D, Yang Q (2020) A fairness-aware incentive scheme for federated learning. *Proc. AAAI/ACM Conf. AI Ethics Soc.* (Association for Computing Machinery, New York), 393–399.

Yue X, Nouiehed M, Al Kontar R (2023) GIFAIR-FL: A framework for group and individual fairness in federated learning. *INFORMS J. Data Sci.* 2(1):10–23.

Zafar MB, Valera I, Gomez-Rodriguez M, Gummadi KP (2019) Fairness constraints: A flexible approach for fair classification. *J. Machine Learn. Res.* 20(1):2737–2778.

Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C (2013) Learning fair representations. *Internat. Conf. Machine Learn.* (PMLR, New York), 325–333.

Zhang Y, Yang Q (2021) A survey on multi-task learning. *IEEE Trans. Knowledge Data Engrg.* 34(12):5586–5609.

Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2017) mixup: Beyond empirical risk minimization. Preprint, submitted October 25, https://arxiv.org/abs/1710.09412.

Zhang X, Li Y, Li W, Guo K, Shao Y (2022) Personalized federated learning via variational Bayesian inference. *Internat. Conf. Machine Learn.* (PMLR, New York), 26293–26310.

Zhang T, Zhu T, Li J, Han M, Zhou W, Philip SY (2020) Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination. *IEEE Trans. Knowledge Data Engrg.* 34(4): 1763–1774.

Zhao C, Chen F (2019) Rank-based multi-task learning for fair regression. *2019 IEEE Internat. Conf. Data Mining (ICDM)* (IEEE, Piscataway, NJ), 916–925.

Zhu X, Feng J, Huang S, Chen C (2020) An online updating method for time-varying preference learning. *Transportation Res. Part C Emerging Tech.* 121:102849.

Zhu X, Wang F, Chen C, Reed D (2019) Personalized control for promoting sustainable travel behaviors. *Transportation Res. Procedia* 38:730–750.

Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2020) A comprehensive survey on transfer learning. *Proc. IEEE* 109(1):43–76.