Multi-Layer Personalized Federated Learning for Mitigating Biases in Student Predictive Analytics

Yun-Wei Chu, Seyyedali Hosseinalipour, Elizabeth Tenorio, Laura Cruz, Kerrie Douglas, Andrew S. Lan, Christopher G. Brinton

***Abstract*—Conventional methods for student modeling, which involve predicting grades based on measured activities, struggle to provide accurate results for minority/underrepresented student groups due to data availability biases. In this paper, we propose a Multi-Layer Personalized Federated Learning (MLPFL) methodology that optimizes inference accuracy over different layers of student grouping criteria, such as by course and by demographic subgroups within each course. In our approach, personalized models for individual student subgroups are derived from a global model, which is trained in a distributed fashion via meta-gradient updates that account for subgroup heterogeneity while preserving modeling commonalities that exist across the full dataset. The evaluation of the proposed methodology considers case studies of two popular downstream student modeling tasks, knowledge tracing and outcome prediction, which leverage multiple modalities of student behavior (e.g., visits to lecture videos and participation on forums) in model training. Experiments on three real-world online course datasets show significant improvements achieved by our approach over existing student modeling benchmarks, as evidenced by an increased average prediction quality and decreased variance across different student subgroups. Visual analysis of the resulting students' knowledge state embeddings confirm that our personalization methodology extracts activity patterns clustered into different student subgroups, consistent with the performance enhancements we obtain over the baselines.**

***Index Terms*—federated learning, student modeling, personalization, de-biasing**

I. Introduction

ONLINE learning, underscored by its substantial surge during the COVID-19 pandemic [2], has become a critical component of contemporary educational platforms [3]. The lack of in-person interaction in online education poses challenges for instructors in giving personalized attention to individual students and delivering tailored feedback. This need for tailored feedback has motivated the exploration of AI-driven methods for providing personalized guidance and feedback in online learning based on measured student progress in online learning activities [4].

Student modeling [5] and its associated research area aims to produce analytics that may inform such personalization efforts. A broad spectrum of student models have emerged, including those focused on (i) evaluating student *knowledge*, such as item response theory [6] and models for knowledge tracing [7], and those addressing (ii) student *behavior*, for instance, to recognize psychological states [8], unveil learning tendencies [9], and uncover engagement patterns in discussion forums [10]. Because these student models are constructed using data acquired from actual learning platforms, they are naturally prone to any biases present within the accessible data [11]. As a result, addressing biases in data-driven student models has gained traction, and research has delved into examining inherent algorithmic biases in educational contexts [12]. The methodologies explored to mitigate bias inherited from the data include introducing constraints during model training that ensure equitable outcomes among diverse student groups [13].

A shared focus in these prior de-biasing studies is the examination of bias/fairness within a singular *global* student model that encompasses data from *all* students [14]. This setup is typically effective in AI applications since more data generally leads to improved model fit. However, this setup ignores the fact that *underrepresented groups* may not be well-captured by a population-level model, resulting in unfair predictions that could severely impact some students [15], [16]. In contrast, training separate *local* models for each student subgroup might prove ineffective, as smaller subgroups lack sufficient data for accurate model training. This study endeavors to formulate a *personalized student modeling methodology* to tackle the data availability issue challenges across subgroups.

A. Federated Learning and Student Modeling

The recently introduced federated deep knowledge tracing (FDKT) approach [17] makes an initial attempt to coordinate between global and local student models. It leverages federated learning (FL) [18], a popular technique for enabling collaboration among models trained on heterogeneous local datasets through periodic global model aggregations.

Beyond the conventional FedAvg algorithm [19], research in *Global FL* has introduced diverse aggregation techniques to optimize the obtained global model, in terms of relevant objectives for different learning settings, e.g., [20]. However, statistical heterogeneity across local data distributions – i.e., non-i.i.d. (independent and identically distributed) data – is an

Y. Chu and C. Brinton are with the Elmore Family School of Electrical and Computer Engineering, Purdue University, IN, USA. e-mail: {chu198, cgb}@purdue.edu

S. Hosseinalipour is with the Department of Electrical Engineering, University at Buffalo (SUNY), NY, USA. email: alipour@buffalo.edu

L. Cruz is with the Department of Engineering Education, University of Florida, FL, USA. e-mail: cruzcastrol@ufl.edu

K. Douglas is with the School of Engineering Education, Purdue University, IN, USA. e-mail: douglask@purdue.edu

A. Lan is with the Manning College of Information and Computer Sciences, University of Massachusetts Amherst, MA, USA. email: andrewlan@cs.umass.edu

An abridged version [1] of this paper has been published in the 31st ACM International Conference on Information and Knowledge Management.

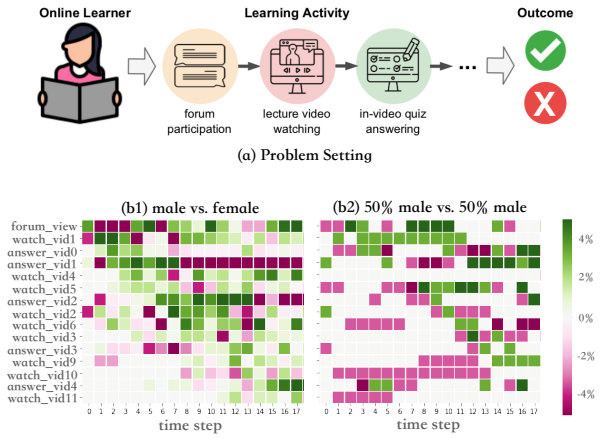


Fig. 1. (a) Our problem setting of tracing online student learning activity to predict learning outcomes. (b) Heatmaps illustrate differences in learning behaviors observed (b1) across distinct student subgroups, such as males versus females, and (b2) within the male subgroup, using our QDS dataset.

inherent characteristic of most FL applications that presents fundamental challenges to Global FL [21]. This has motivated *Personalized FL* which aims to redefine the relationship between the local and global models for a better match to individual local datasets [22]. In particular, personalized FL shifts the optimization focus to individual local objectives, often by incorporating meta-functions into the definition of the global learning objective [23].

In student modeling, this statistical heterogeneity property manifests along several dimensions. FDKT [17] deals with the challenge of heterogeneity stemming from the division of student training data across multiple schools. Applying the FL principle, individual school-specific models are trained, and communication between the global and local models is managed without sharing data across schools. Although this setup mirrors certain real-world educational scenarios, it overlooks the broader context of diversity among student subgroups, including factors like race, gender, and demographic variables. In AI fairness research, it has been noted that minority groups with limited data can be overlooked during model training, leading to biases and ethical concerns [24], [25]. This happens because AI models are frequently trained on datasets that contain more samples from majority population groups, causing them to be better suited for these groups while being less effective for minority groups with limited representation. This issue has been identified in various AI application domains, such as medicine [26], language modeling [27], and image recognition [28]. As a result, biased models tend to be less accurate and unfair when applied to data from underrepresented groups. In the online education field, data from user interactions are similarly skewed according to the demographics of students who have historically taken these courses the most, naturally creating biases in AI-driven student models [29], [30]. This bias has been observed in several applications, such as knowledge tracing [31] and graduation prediction [32] tasks, where AI models produce lower-quality outputs for students from underrepresented backgrounds. For instance, students from lower-income support levels are often disadvantaged in these educational predictions, further exacerbating fairness issues in AI models [31], [32]. Such challenges highlight the importance of addressing bias and ethical concerns in the training data to promote fairness and accuracy for all groups.

The aggregation mechanism in FDKT assigns weights to local models based on their "*data quality*," which is assessed by fitting local psychometric models like classical test theory and item response theory. This leads to an implicit devaluation of underrepresented student subgroups due to insufficient data volume. However, this is unacceptable in our setting as the data from each group holds crucial information about these students' learning processes. We address this in our work by developing a Personalized FL methodology which accounts for heterogeneity across courses and demographics in a multi-tiered student modeling framework.

B. Activity-Based Student Modeling

Our student modeling approach will build upon two significant insights from previous research in educational data mining that can further mitigate biases. These will come into play in Sec. III as we formalize two downstream student modeling applications for case-studies of our methodology.

The first insight is that student learning activity patterns, encompassing actions like video viewing, participation in discussion forums, and clickstream logs (as depicted in Figure 1(a)), contain signals of student achievement in online courses [33], [34]. Given that each student typically generates a significant number of activity records over a course's duration, inclusion of such data could alleviate the sparsity challenges faced by prediction models trained on subgroups with fewer students.

The second observation is that unique learning activity patterns can be identified both within and across student groups (e.g., different gender groups) [35], [36]. This phenomenon was initially found in traditional classrooms, including behaviors like participation and engagement with course materials, and more recently in digital learning environments as well [37]. Hence, we aim to consider how variations in sequences of learning behaviors, such as accessing a forum, viewing specific videos, and answering particular quiz questions, across student subgroups can be integrated into personalized federated learning models. Yet, online learning behaviors are noisier than those occurring in in-person settings (e.g., accidental video access, inadvertent skipping through a video), posing challenges in identifying these patterns through conventional data mining methods [33].

For example, considering one of the datasets in this paper, Figure 1(b) depicts the heatmaps of differences observed when students engage in particular learning activities. We explore two cases: (b1) across learners of a demographic group (all males vs. all females), and (b2) within a subgroup (50% of males vs. the other 50%, randomly chosen). Each heatmap value represents the difference in the proportion of students participating in the activity at that specific point in their learning journey. (b1) reveals varied trajectories compared to (b2), consistent with diverse learning behaviors among different subgroups. However, (b2) also shows significant differences despite being a comparison within the same subgroup. This motivates the meta-learning approach we propose

for subgroup personalization, wherein shared patterns in data across subgroups are integrated into a global model, which can then be fine-tuned using local subgroup information.

C. Outline and Summary of Contributions

In this paper, we develop a federated student modeling methodology that personalizes prediction models for different data grouping criteria and mitigates the bias in data availability for underrepresented groups. Specifically, we make the following major contributions:

- We design Multi-Layer Personalized Federated Learning (MLPFL), where local models associated with different student groupings are adapted from a global model (Section II). Unlike prior methods that invoke data quality heuristics, this adaptation involves meta-gradient updates on localized data from particular courses or student subgroups. We show that this mitigates biases from heterogeneous data availability.
- We consider two popular downstream student modeling tasks for our methodology, knowledge tracing and outcome prediction (Section III-B). We formulate these based on multi-modal activity logs on learner interactions with course content and in discussion forums available in real-world online course datasets (Section III-A).
- Through experiments on three online education datasets (Section IV), we demonstrate that our methodology significantly outperforms existing approaches in terms of improving prediction accuracy and reducing variance across subgroups (Section IV-A-IV-B). We find this advantage appears at both the global and local model levels, resulting in a robust prediction framework that can adapt between different student groupings.
- Our experiments reveal that demographic-based MLPFL provides substantial improvements in student modeling compared to course-level personalization without demographic information. We also visually examine how the student activity embeddings generated by our approach cluster according to subgroups, and find that they exhibit informative clusterings (Section IV-C).

This paper is an extension of our prior conference version [1]. Compared with [1], this extension adds the following major components: (1) We consider a *hierarchical* personalization approach that incorporates subgroup variables across both courses and demographic groups within courses, while [1] only considers demographic groups and treats each course independently. (2) In addition to student outcome prediction considered in [1], we extend our exploration to include the knowledge tracing use case. (3) Our experiments have been augmented with various baselines to evaluate different components of hierarchical personalization, and have also introduced a practical scenario where access to students' demographic information is unavailable.

II. Multi-Layer Personalized Federated Learning

This section introduces the training procedures of Multi-Layer Personalized Federated Learning (MLPFL) based on general modeling assumptions. We detail the model structures for specific downstream tasks in Section III-B. The two layers we consider for our modeling hierarchy are *courses* and *demographics*. Since demographic information may not always be available, we consider two different scenarios, as depicted in Figure 2. The first scenario (Figure 2(a)) conducts adaptable student modeling by course only. The second scenario (Figure 2 (b)) further considers personalization for each demographic subgroup within each course, i.e., when students have provided this information.

A. Scenario I: Course-specific Adaptation

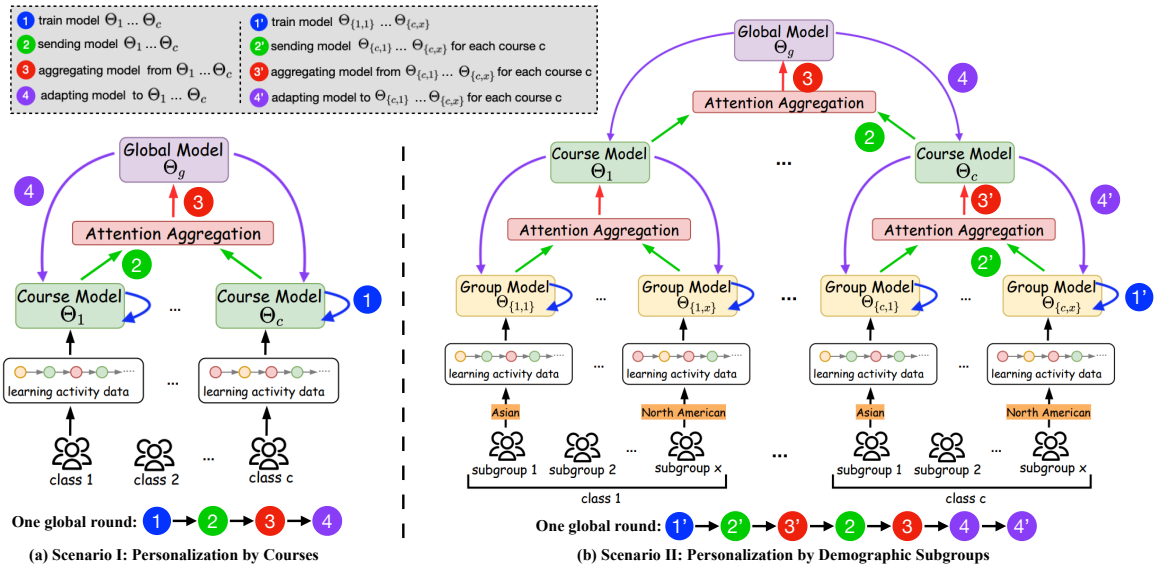
1) Data Partitioning by Courses: Well-defined dataset partitioning is essential for scenarios I and II since it is the basis for personalization. We use Ω_c to represent the students of each course $c \in \{1, 2, ..., C\}$, where C is the number of courses. For each course, the partition is done by splitting students into train-test sets at a ratio of 4:1, denoted as $\Omega_c = \{\Omega_c^{\mathsf{train}}, \Omega_c^{\mathsf{test}}\}$. Therefore, our entire collected dataset is $\Omega = \bigcup_{c=1}^{C} \Omega_c = \{\Omega^{\mathsf{train}}, \Omega^{\mathsf{test}}\}$, where $\Omega^{\mathsf{train}} = \bigcup_{c=1}^{C} \Omega_c^{\mathsf{train}}$ and $\Omega^{\mathsf{test}} = \bigcup_{c=1}^{C} \Omega_c^{\mathsf{test}}$. The methodology proposed in this section uses $\Omega_c^{\mathsf{train}}$ $(c = 1, \cdots, C)$ to build C models for courses, denoted as Θ_c, for the downstream tasks formulated in Section III. The prediction performance of each trained model Θ_c is subsequently evaluated by using test set Ω_c^{test}.

2) Meta-Learning Adaptation: To facilitate adaptation to courses, our methodology obtains separate local models for each course. Locally, we address two tasks: (i) generating a course-customized model Θ_c for global aggregations and (ii) adapting the global model Θ_g according to the characteristics of students within each course. Our objective is to develop a global model that can be *easily adaptable* to each course. For this purpose, we employ a personalized federated learning framework based on meta-learning [23] to find the global model that addresses the optimization problem:

$$\min_{\Theta} \sum_{c \in \mathcal{C}} F_c(\Theta) := f_c\big(\underbrace{\Theta - \nabla f_c(\Theta)}_{(a)}\big), \tag{1}$$

where $F_c(\cdot)$ is the meta-loss function of course c, and $f_c(\cdot)$ is the original loss function, which is the sum of prediction losses (i.e., binary cross entropy loss as we will formulate in Section III-B1 and III-B2) over course c. The key distinction between our loss function (equation (1)) and those used in prior student modeling studies [38], [39] is that ours minimizes the loss of the *adapted* versions of the global model. This adaptation is performed through a single gradient descent step, represented as term (a) in (1), executed on the dataset of course c. In essence, our approach leverages the *commonality* of data across courses to develop an adaptable global model which can be easily tailored to individual courses.

To address (1), we derive local update steps based on meta-gradients. The training process is executed in a series of training *rounds* $k \in 1, \cdots, K$, where each round comprises multiple local training *iterations* $e \in 1, \cdots, E$. Within each iteration e, course $c \in \mathcal{C}$ updates its local model $\Theta_c^{(k,e)}$ using solely the student data associated with that course. After each local model completes its E iterations, these local models are synchronized through a global aggregation process. In



(a) Scenario I: Personalization by Courses

(b) Scenario II: Personalization by Demographic Subgroups

Fig. 2. Overview of the personalized methodology we develop in this paper for customizing prediction models for (a) each course and (b) each demographic student subgroup. Our Multi-Layer Personalized Federated Learning (MLPFL) methodology refines global models in a hierarchical manner to create personalized local student models for distinct student groups, leveraging both commonality and subgroup diversity.

every round k, the local model for each course c begins with initialization as $\Theta_c^{(k,0)} = \Theta_g^{(k)}$, where $\Theta_g^{(k)}$ represents the global model achieved at the completion of round $k-1$. Then, each course c performs its meta-gradient updates according to

$$\Theta_c^{(k,e)} = \Theta_c^{(k,e-1)} - \eta \nabla F_c \left(\Theta_c^{(k,e-1)} \right), \ e = 1, \cdots, E, \quad (2)$$

where η is the step size, and following (1), the meta gradient ∇F_c is computed as:

$$\nabla F_c(\Theta) = \left(\mathbf{I} - \nabla^2 f_c(\Theta) \right) \nabla f_c \left(\Theta - \nabla f_c(\Theta) \right), \ \forall \Theta, \quad (3)$$

where ∇^2 is the Hessian operator. The second-order Hessian term in (3) can be first-order approximated without causing deterioration in performance [23]. Following the local updates, we compute the global aggregation $\Theta_g^{(k+1)}$ through an attention-based technique explained in the following section.

3) Attention-Based Global Model Aggregation: The global modeling phase in our approach involves two tasks: (i) aggregating local models and (ii) synchronizing course models with the resulting parameter vector for subsequent rounds of local model training. Rather than relying on the conventional averaging-based aggregation FedAvg [19], we utilize an attention-based aggregation, FedAtt [20]. FedAtt incorporates an attention mechanism to assign weights to local models based on the extent to which the parameters in their individual layers have diverged from the previous global model.

In each aggregation step, the global model is informed by two elements: (i) the entirety of local models $\Theta_c^{(k,E)}$ $(c = 1, \cdots, C)$, and (ii) the respective attention weight associated with each local model. The attention weight $\alpha_c^{(k)}$ for the local course model Θ_c is calculated across its layers $\mathcal{L}$ as follows:

$$\alpha_c^{(k)} = \sum_{\ell \in \mathcal{L}} \text{softmax} \left(\left\| \Theta_g^{(k)}(\ell) - \Theta_c^{(k,E)}(\ell) \right\| \right), \quad (4)$$

where $\ell \in \mathcal{L}$ represents a specific model layer, and $\Theta_g^{(k)}(\ell)$ and $\Theta_c^{(k,E)}(\ell)$ denote the parameter vectors of the ℓ-th layer

Algorithm 1 Course Level MLPFL

Global Model at k training round: Θ_g^k; Local Model of course c at k training round and e local iteration: $\Theta_c^{(k,e)}$
Global Execution:
Initialize global model $\Theta_{\text{g}}^{(0)}$
for each global round $k = 1, 2, ..., K$ **do**
  for each course $c \in \mathcal{C}$ **in parallel do**
    $\Theta_c^{(k,E)} \leftarrow$ **LocalAdaptation** $(\Theta_{\text{g}}^{(k-1)})$
  for each course $c \in \mathcal{C}$ **do**
    Compute attention weight $\alpha_c^{(k-1)}$ using (4)
  Obtain the global model $\Theta_{\text{g}}^{(k)}$ by (5)

LocalAdaptation $(\Theta_g^{(k-1)})$:
Initialize the local model $\Theta_c^{(k,0)} = \Theta_{\text{g}}^{(k-1)}$
for Each local iteration $e = 1, \cdots, E$ **do**
  Obtain $\Theta_c^{(k,e)}$ using the meta-update rule (2)
Return parameters $\Theta_c^{(k,E)}$

in the global and local models, respectively, during global aggregation. The resulting aggregated global model is:

$$\Theta_g^{(k+1)} = \Theta_g^{(k)} - \epsilon \sum_{c \in \mathcal{C}} \alpha_c^{(k)} \left(\Theta_g^{(k)} - \Theta_c^{(k,E)} \right), \quad (5)$$

where ϵ is a tunable step-size. The full personalized FL process is outlined in Algorithm 1 and Figure 2(a).

B. Scenario II: Personalized by Course and Demographics

1) Data Partitioning by Demographic Variables: Open platforms, such as edX and Coursera, usually solicit students' demographic details during registration, which they can fill out voluntarily. In this section, we add an additional layer of personalization based on this demographic information. Next, we illustrate the partitioning criterion using the training set as an example; the test set follows the same partitioning rule. We use $\Omega_{\{c,I\}}^{\text{train}}$ to represent the students belonging to the training

set who answered specific demographic information I within course c. For each variable I, we assume there are a finite set of categories (groups) for students to select from, e.g., as defined by a dropdown list. Based on our available datasets from edX (see Section III-A), the demographic information $I \in \{\texttt{G}, \texttt{C}, \texttt{Y}\}$ includes three frequently employed variables: Gender (G), Country (C), and Year of Birth (Y). For country, we group students into five continents since some countries receive no responses at all. $\Omega^{\text{train}}_{\{c,\texttt{G}\}}$, $\Omega^{\text{train}}_{\{c,\texttt{C}\}}$, and $\Omega^{\text{train}}_{\{c,\texttt{Y}\}}$ are not necessarily the same since students may choose to not respond to certain demographic questions.

For the set $\Omega^{\text{train}}_{\{c,I\}}$, we further define $\Omega^{\text{train}}_{\{c,I,x\}}$ as the set of students belonging to subgroup $x \in \mathcal{X}$, where $\mathcal{X}$ is the set of groups in variable I. For example, $\mathcal{X} = \{\texttt{M}, \texttt{F}\}$ represents male and female subgroups for gender information ($I = \texttt{G}$) when it is provided. For country information ($I = \texttt{C}$), $\mathcal{X} = \{\texttt{AS}, \texttt{AF}, \texttt{EU}, \texttt{NA}, \texttt{SA}\}$ represents Asian (AS), African (AF), European (EU), North American (NA), and South American (SA) student subgroups. Furthermore, $\mathcal{X} = \{\texttt{<80}, \texttt{80-90}, \texttt{>90}\}$ represent year of birth prior to 1980, between 1980 and 1990, and after 1990 for birth year information ($I = Y$). To conclude, $\Omega^{\text{train}}_{\{c,I\}} = \bigcup_{x\in\mathcal{X}} \Omega^{\text{train}}_{\{c,I,x\}}$

Moreover, we introduce $\Omega^{\text{train}}_{\{c,\neg I\}}$ to denote the set of students that chose not to provide demographic information I in the training set of course c. To be specific, $\Omega^{\text{train}}_{c} = \{\Omega^{\text{train}}_{\{c,I\}}, \Omega^{\text{train}}_{\{c,\neg I\}}\}$. We will use $\Omega^{\text{train}}_{c}$ to build several local models, denoted as $\Theta_{\{c,I,x\}}$ or $\Theta_{\{c,\neg I\}}$. $\Theta_{\{c,\neg I\}}$ denotes the model built by $\Omega^{\text{train}}_{\{c,\neg I\}}$, and $\Theta_{\{c,I,x\}}$ represents the model constructed by the specific x subgroup of demographic information I within course c. When performing Multi-Layer Personalized Federated Learning in this section, we consider two settings: (i) training models by using $\Theta_{\{c,I,x\}}$ and (ii) training an additional model $\Theta_{\{c,\neg I\}}$ alongside $\Theta_{\{c,I,x\}}$. The prediction performance of each model then is tested on $\Omega^{\text{test}}_{\{c,I,x\}}$ or $\Omega^{\text{test}}_{\{c,\neg I\}}$ after the training procedure.

2) Meta-Learning Personalization: We build separate models for demographic subgroups to facilitate more fine-granular prediction personalization within each course. Specifically, as shown in Figure 2(b), we build separate subgroup-personalized models and increase the hierarchy during MLPFL training. Similar to the local model personalization for courses as explained in Section II-A2, we undertake two tasks at the subgroup level: (i) generating a subgroup-personalized model for global aggregations and (ii) adapting the global model using local subgroup data. Our aim is to train a global model that is *easily adaptable* to each local student subgroup. Therefore the optimization problem will change from (1) to:

$$\min_{\Theta} \sum_{c\in\mathcal{C}} \sum_{x\in\mathcal{X}} F_{\{c,I,x\}}(\Theta) := f_{\{c,I,x\}}\big(\underbrace{\Theta - \nabla f_{\{c,I,x\}}(\Theta)}_{(b)}\big), \tag{6}$$

where $F_{\{c,I,x\}}(\cdot)$ is the meta-loss function of student subgroup x (corresponding to information I) in course c, and $f_{\{c,I,x\}}(\cdot)$ is the original loss function, which is the sum of prediction loss over the dataset of group x in course c. This adaptation is performed through a single gradient descent step, indicated by term (b) in (6), conducted on the local dataset of subgroup x. In essence, our approach leverages the *commonality* of data across subgroups to develop an adaptable global model, readily customizable for each specific subgroup.

To solve (6), we derive meta-gradient based local update steps. Similar to Scenario I, the training also proceeds through a sequence of training rounds $k \in \{1, \cdots, K\}$, with each round consisting of multiple local training iterations $e \in \{1, \cdots, E\}$. During each iteration e, the local subgroup-model $\Theta^{(k,e)}_{\{c,I,x\}}$ is updated by meta-gradient:

$$\Theta^{(k,e)}_{\{c,I,x\}} = \Theta^{(k,e-1)}_{\{c,I,x\}} - \eta \nabla F_{\{c,I,x\}}\left(\Theta^{(k,e-1)}_{\{c,I,x\}}\right),\ e = 1, \cdots, E, \tag{7}$$

where η is the step size, and $\nabla F_{\{c,I,x\}}$ is the meta gradient.

3) Global Model Aggregation and Adaptation: We employ attention-based aggregation for global model construction. As we have additional hierarchy for training, the global modeling stage in this scenario encompasses four tasks: (i) aggregating local subgroup-models into course-models, (ii) aggregating course-models into a global model, (iii) synchronizing course-models with the resulting global parameters for course-level adaptation, and (iv) synchronizing subgroup-models with course-level parameters for subgroup-level adaptation.

At each aggregation step, local subgroup-models are first aggregated into course-models as:

$$\Theta^{(k+1)}_{c} = \Theta^{(k)}_{c} - \epsilon \sum_{x\in\mathcal{X}} \alpha^{(k)}_{\{c,I,x\}} \left(\Theta^{(k)}_{c} - \Theta^{(k,E)}_{\{c,I,x\}}\right) \forall I, \tag{8}$$

where $\alpha^{(k)}_{\{c,I,x\}}$ is the layer-wisely computed attention weight for local subgroup-models:

$$\alpha^{(k)}_{\{c,I,x\}} = \sum_{\ell\in\mathcal{L}} \text{softmax}\left(\left\|\Theta^{(k)}_{c}(\ell) - \Theta^{(k,E)}_{\{c,I,x\}}(\ell)\right\|\right), \tag{9}$$

with $\mathcal{L}$ denoting the set of layers. After aggregating the subgroup-models into course-models, the global model Θ_g receives the parameters of each course-model Θ_c and performs attention-based aggregation as follows:

$$\Theta^{(k+1)}_{\text{g}} = \Theta^{(k)}_{\text{g}} - \epsilon \sum_{c\in\mathcal{C}} \alpha^{(k)}_{c} \left(\Theta^{(k)}_{\text{g}} - \Theta^{(k+1)}_{c}\right), \tag{10}$$

where $\alpha^{(k)}_{c}$ is the attention weight for course-models:

$$\alpha^{(k)}_{c} = \sum_{\ell\in\mathcal{L}} \text{softmax}\left(\left\|\Theta^{(k)}_{\text{g}}(\ell) - \Theta^{(k+1)}_{c}(\ell)\right\|\right), \tag{11}$$

and $\ell \in \mathcal{L}$ denotes a particular model layer.

The full MLPFL algorithm is summarized in Algorithm 2 and Figure 2(b). Note that, in Scenario II, there are two steps of synchronization after $\Theta^{(k)}_{\text{g}}$ has been aggregated, compared with Scenario I where there is just one step. The first synchronization is adapting the global model Θ_g as a temporary course-model Θ'_c, as shown in line 6 of Algorithm 2. This course adaptation is done through a single-step update by using a batch of students from course c. We use stratified sampling to select a few students within each subgroup to form this batch without bias. After the course adaptation, the second step is to synchronize and adapt the subgroup models, shown in line 8. In each round k, the subgroup model for demographic x in course c is initialized as $\Theta^{(k,0)}_{c,x} = \Theta^{(k)'}_{c}$, where $\Theta^{(k)'}_{c}$ is the temporary course-model. Referring to Figure 2(b), the $\Theta^{(k,E)}_{c,x}$

Algorithm 2 Demographic Level MLPFL

1: Global Model at k training round: Θ_g^k; Global Course-level Model of course c at k training round: Θ_c^k; Local Demographic-level Model of demographic subgroup x in course c at k training round and e local iteration: $\Theta_{\{c,x\}}^{(k,e)}$
2: **Global Execution**:
3: Initialize global model $\Theta_{\mathrm{g}}^{(0)}$
4: **for** each global round $k = 1, 2, ..., K$ **do**
5: **for** each course $c \in \mathcal{C}$ **in parallel do**
6: $\Theta_c^{'(k)} \leftarrow$ **Course-level Adaptation** $(\Theta_{\mathrm{g}}^{(k-1)})$
7: **for** each subgroup $x \in \mathcal{X}$ **in parallel do**
8: $\Theta_{\{c,x\}}^{(k,E)} \leftarrow$ **LocalAdaptation** $(\Theta_c^{'(k)})$
9: **for** each subgroup $x \in \mathcal{X}$ **do**
10: Compute attention weight $\alpha_x^{(k-1)}$ using (9)
11: Obtain the course-level model $\Theta_{\mathrm{c}}^{(k)}$ by (8)
12: **for** each course $c \in \mathcal{C}$ **do**
13: Compute attention weight $\alpha_c^{(k-1)}$ by (11)
14: Obtain global model $\Theta_{\mathrm{g}}^{(k)}$ based on (10)

1: **Course-level Adaptation** $(\Theta_g^{(k-1)})$:
2: Initialize the course-level model $\Theta_c^{(k)} = \Theta_g^{(k-1)}$
3: Obtain $\Theta_c^{'(k)}$ using one step of meta-update (2)

1: **LocalAdaptation** $(\Theta_c^{'(k)})$:
2: Initialize the local model $\Theta_{\{c,x\}}^{(k,0)} = \Theta_c^{'(k)}$
3: **for** each local iteration $e = 1, \cdots, E$ **do**
4: Obtain $\Theta_c^{(k,e)}$ using the meta-update (7)
5: Return parameters $\Theta_c^{(k,E)}$ for each course c

correspond to the models at the bottom level of the hierarchy, which have the most fine-granular adaptation. Then, in line 11, the updated course-level models $\Theta_c^{(k+1)}$ are obtained by aggregating the bottom-level models, corresponding to the middle level of the hierarchy. Finally, in line 14, the updated global model $\Theta_{\mathrm{g}}^{(k+1)}$ is obtained by aggregating the middle-level models, corresponding to the top level of the hierarchy.

III. Datasets and Prediction Tasks

A. Datasets

Our datasets are sourced from online courses hosted on the edX platform (www.edx.org) at Purdue University. We study three graduate-level courses: Fiber Optic Communications (FOC), Quantum Detectors and Sensors (QDS), and Essentials of MOSFETs (MOSFETs). Within each course, a series of lecture videos is available, some of which include end-of-video quizzes to evaluate student comprehension. Additionally, every course features a discussion forum where students can engage in interactions with one another. The platform also logs the final grade of each student, indicating whether they passed or failed according to the grading policy. Table I provides summary statistics for the three courses, highlighting a diverse range of activity levels for our evaluation. Details about the dataset are available in Appendix A.

TABLE I
Summary details of three datasets acquired from edX.

	FOC	QDS	MOSFETs
# of students	1,265	2,304	886
# of lecture videos	43	31	26
# end-of-video quizzes	25	23	11
# of discussion threads	20	35	17
Avg. reply per thread	1.95	0.48	1.44
Avg. activities per student	50.42	41.44	35.41

1) Video-watching behavior and quiz responses: Each time a student u accesses a lecture video v_c of course c, their activity is recorded with the following information: student ID, course ID, video ID, and UNIX timestamp. In each course, the in-video quizzes comprise either a single multiple-choice question or a True/False question, presented at the end of the video. When student u submits an answer to the in-video question for video v_c, the platform records the student's response $r_{u,v_c} \in \{0, 1\}$. $r_{u,v_c} = 0$ indicates an incorrect response, and $r_{u,v_c} = 1$ otherwise.

2) Discussion forum interactions: When student u visits the discussion forum page, the learning platform logs forum participation activities as $f_u \in \{\mathsf{forum\_post}, \mathsf{forum\_reply}, \mathsf{forum\_view}\}$. forum_post represents a student initiating a new thread or making a post to an existing one, forum_reply indicates a student's response to another student's post, and forum_view marks a student's visit to a thread without posting or replying.

B. Student Prediction Modeling Tasks

To evaluate MLPFL, we implement two downstream tasks, knowledge tracing [38] and student outcome prediction [1].

1) Knowledge Tracing: We formulate knowledge tracing task by using students' video-and-quiz interactions in our dataset. Let $x_{u,t} = (i_{u,t}, r_{u,v_c,t})$ be a tuple representing the item i attempted by student u at time t, where t serves as the activity index for each student. $r_{u,v_c,t} \in \{0, 1\}$ represents the result of the student's response to the quiz on video v_c at time t. $i_{u,t} = \mathbb{1}(c) \oplus \mathbb{1}(v_c)$ is a combined representation of course ID c and video v_c by using one-hot function $\mathbb{1}(\cdot)$, where $\oplus$ denotes vector concatenation.

Given a series of interactions $X_u = \{x_{u,1}, x_{u,2}, ..., x_{u,t}\}$, the goal of knowledge tracing task is to predict the value of $r_{u,v_c,t+1}$ representing if student u will answer the new item correctly based on their current knowledge state $\mathbf{h}_{u,t}$. The current knowledge state of a student will be highly correlated with their previous knowledge state. Therefore, we follow [38]'s suggestion that models student learning process by Long Short Term Memory network (LSTM) as:

$$\mathbf{h}_{u,t} = \mathrm{LSTM}\left(i_{u,t},\ \mathbf{h}_{u,t-1}\right). \tag{12}$$

The probability of correctly answering item $i_{u,t}$ is:

$$\mathbf{p}_{u,t+1} = \mathrm{softmax}\left(\mathbf{W} \cdot \mathbf{h}_{u,t} + \mathbf{b}\right), \tag{13}$$

where $\mathbf{W} \in \mathbb{R}^{k \times 2}$ is the weight for linearly transforming $\mathbf{h}_{u,t}$, $k = 48$ is the hidden dimension, and $\mathbf{b} \in \mathbb{R}^2$ is the bias vector. For student u answering item $i_{u,t}$ at time t, the prediction loss $l_{u,t}$ can be modeled with the binary cross entropy loss:

$$l_{u,t} = \mathbf{r}_{u,v_c,t} \log(\mathbf{p}_{u,t} + (\mathbf{1} - \mathbf{r}_{u,v_c,t}) \log(\mathbf{1} - \mathbf{p}_{u,t})), \tag{14}$$

where $\mathbf{1}$ is an all-one vector and $\mathbf{r}_{u,v_c,t}$ is the one-hot encoding vector of the binary response $r_{u,v_c,t}$. Finally, the total loss $\mathcal{L}_{KT}$ for knowledge tracing (KT) can be represented as

$$\mathcal{L}_{KT} = \sum_{u \in \Omega} \sum_{t}^{L_u} l_{u,t}, \tag{15}$$

where L_u is the length of the time series sequence X_u.

2) Outcome Prediction: In addition to tracing students' knowledge state, we consider both video interaction and forum-activity to form a complete knowledge representation for outcome prediction [1]. For each student u, the complete activity record at time step t is defined as $\mathbf{a}_{u,t} = \mathbb{1}(c) \oplus \mathbb{1}(v_c) \oplus \mathbb{1}(r_{u,v_c,t}) \oplus \mathbb{1}(f_{u,t})$, where $f_{u,t}$ represents the forum-participation activity f_u made by student u at time t. We adopt this concatenation of video interaction and forum-participation activity to maintain uniformity in the activity dimension. Note that video and forum activities occur separately; whenever a student watches a video, the forum-participation part of this encoding, i.e., $\mathbb{1}(f_{u,t})$, is set as $\mathbf{0}$. Likewise, whenever a student makes a forum-participation activity, the video part of $\mathbf{a}_{u,t}$, i.e., $[\mathbb{1}(v_c) \oplus \mathbb{1}(r_{u,v_c,t})] = \mathbf{0}$.

Given a series of activities $A_u = \{\mathbf{a}_{u,1}, \mathbf{a}_{u,2}, ..., \mathbf{a}_{u,t_u}\}$, the goal of outcome prediction is to infer a binary classification label $s_u \in \{0, 1\}$ indicating whether the student successfully completed the course or not. Note that t_u, the total number of activities for student u used for modeling, can vary based on the time-frame of interest for the prediction tasks; it can represent any segment of the course duration, such as the entire course, or segments more appropriate for early detection prediction, e.g., the first couple weeks of the course. This flexibility allows for tailored analyses at different stages of the course, accommodating various intervals of student engagement. The specific designations of t_u used in our experiments will be detailed in the experiments section. We follow [1]'s suggestion and leverage attention-based Gated Recurrent Units (GRU) to capture dependencies over long time periods [34]. The model takes encoded activities as an input, generates learned representations for each student, and predicts learning outcome. The hidden state of the GRU model is defined as follows:

$$\mathbf{h}_{u,t} = \text{GRU}\left(\mathbf{a}_{u,t},\ \mathbf{h}_{u,t-1}\right). \tag{16}$$

Encoding a long temporal sequence $\mathbf{a}_{u,t}$ into a single final state $\mathbf{h}_{u,t}$ can lead to substantial loss of information. To address this concern, [40] introduced a self-attention mechanism that assigns weights to $\mathbf{h}_{u,t}$ across time. Rather than forcing the network to condense all information into the final state, an attention module accepts all $\mathbf{h}_{u,t}$ as input and produces the learned representation $\widetilde{\mathbf{h}}_u$ as output. Applying this concept to our context, we define an attention module as:

$$\widetilde{\mathbf{h}}_u = \sum_{t} \alpha_t \mathbf{h}_{u,t}, \tag{17}$$

where the weight $\alpha_t = \frac{\exp(e_t)}{\sum_t \exp(e_t)}$, $e_t = p_t^\top \tanh\left(\mathbf{W}_\alpha \mathbf{h}_{u,t}\right)$, and $\mathbf{W}_\alpha$ is a learned parameter. Then, a linear layer converts the representation into the predicted pass/fail probability:

$$\mathbf{s}'_u = \text{softmax}\left(\mathbf{W} \cdot \widetilde{\mathbf{h}}_u + \mathbf{b}\right), \tag{18}$$

where $\mathbf{W} \in \mathbb{R}^{k \times 2}$ is the weight matrix for linearly transforming $\widetilde{\mathbf{h}}_u$, $k = 48$ is hidden dimension, and $\mathbf{b} \in \mathbb{R}^2$ is the bias vector. The loss $\mathcal{L}_{OP}$ for assessing the quality of outcome prediction (OP) is defined as binary cross-entropy loss:

$$\mathcal{L}_{OP} = -\sum_{u \in \Omega} \mathbf{s}_u \log(\mathbf{s}'_u) + (\mathbf{1} - \mathbf{s}_u) \log(\mathbf{1} - \mathbf{s}'_u), \tag{19}$$

where $\mathbf{s}'_u$ is the model's prediction, $\mathbf{1}$ is an all-one vector, and $\mathbf{s}_u$ is the one-hot encoded vector of the binary label s_u.

IV. Experimental Evaluation

We now carry out experiments on our three online course datasets from Section III to evaluate our personalization methodology MLPFL. We employ the standard AUC metric for evaluation. See Appendix B for implementation details.

A. Models and Baselines

We compare MLPFL against baselines in several different configurations. These configurations differ in terms of several attributes: (i) scenario (i.e., scenario I (`sc1`) and II (`sc2`) we introduced in Section II-A and II-B), (ii) architecture (i.e., local (`L`), global (`G`), and personalized (`P`) model), (iii) FL aggregation method (i.e., average-based aggregation (`AV`), attention-based aggregation (`AT`)), and (iv) hierarchy information (for the scenario with demographic personalization). For conciseness, we use a general structure:

$$\begin{aligned}\texttt{Method} = [&\texttt{<Scenario>-<Architecture>-}\\ &\texttt{<Aggregation>-<Hierarchy>}]\end{aligned} \tag{20}$$

to represent each baseline according to the attributes listed in Table II. In Scenario II, `<Hierarchy>` specifies the level within our hierarchical adaptation structure where the model is taken from–bottom (B), middle (M), or top (T). As discussed in Sec. II-B, "B" models are tailored to demographic subgroups within courses, providing the most fine-granular adaptation. "M" describes course-specific models that are adapted based on the learning patterns of students within a particular course. "T" indicates the global model, which is aggregated across all courses and demographics to offer the most coarse-granular yet comprehensive overview. This hierarchical approach enables nuanced modeling that ranges from highly personalized to broadly generalized analyses. For example, `[sc2-P-AT-B]` corresponds to the personalized model that adapts the attention-aggregated global model based on the data of the bottom level (demographic subgroups) in scenario II.

Also, under the `<Architecture>` attribute, note that both the "G" and "P" architectures have global federated models, despite the "global (G)" name. The difference is that the global federated models in the "G" category are obtained through aggregations in standard, non-personalized federated learning, whereas those in the "P" category are defined according to our meta-learning adaptation procedures formalized in Section II. Thus, the "G" category contains important baselines for evaluating our meta-learning-based student modeling. We will clarify the exact procedure followed in each algorithm setup from Table II.

TABLE II
ILLUSTRATION OF ALGORITHM SETUPS: sc1 AND sc2 DENOTE THE FIRST AND SECOND SCENARIOS, AND L, G, AND P STAND FOR LOCAL, GLOBAL, AND PERSONALIZED MODELS. <AGGREGATION> INCLUDES AV FOR AVERAGE-BASED AND AT FOR ATTENTION-BASED AGGREGATIONS. IN SCENARIO II, THREE HIERARCHY LEVELS—DEMOGRAPHIC (B), COURSE (M), AND GLOBAL (T)—ARE USED.

Configuration	Attribute
<Scenario>	sc1, sc2
<Architecture>	L, G, P
<Aggregation>	AV, AT
<Hierarchy> (optional)	B, M, T

1) Algorithms for scenario I (<Scenario> = sc1):
Local Modeling (<Architecture> = L) We construct a local baseline [sc1-L] by building several local models based on student data in each course. More specifically, we train three local models, one for each course $c \in \mathcal{C}$, using datasets Ω_c^{train} and evaluate each model on Ω_c^{test}.

Global Modeling (<Architecture> = G) We implement three global models for scenario I. One of them, [sc1-G], is a centralized model without FL, while the others, [sc1-G-AV] & [sc1-G-AT], are Global FL models.

• [sc1-G]: A centralized global model is trained on all the students' training data Ω^{train} collected from all the courses and evaluated on Ω^{test}.

• [sc1-G-AV]: A federated global model is implemented from FedAvg [19]. We train several local models $\Theta_c^{(k,e)}$ on Ω_c^{train}, without meta-learning. Upon completing E iterations, FedAvg aggregates local models, considering the student count in each course as the weighting factor for the aggregation:

$$\Theta_{\text{g}}^{(k+1)} = \sum_{c \in \mathcal{C}} \frac{N_c}{N} \Theta_c^{(k,E)}, \tag{21}$$

where k is training round, N_c is the number of students of course c, and $N = \sum_{c \in \mathcal{C}} N_c$. After K global aggregations, we take $\Theta_{\text{g}}^{(K)}$ as [sc1-G-AV] to evaluate FedAvg on Ω_c^{test}.

• [sc1-G-AT]: An attention-based federated global algorithm is implemented from FedAttn [20]. After E local iterations, FedAttn aggregates local models based on the attention mechanism introduced in (4). After K rounds, the global model $\Theta_{\text{g}}^{(K)}$ defined in (5) is evaluated on Ω_c^{test}.

Adaptive Modeling (<Architecture> = P) We consider two versions of our local model adaptation in FL for scenario I: [sc1-P-AV] and [sc1-P-AT]. Both of these models adapt the global model based on meta-updates from the data in each course.

• [sc1-P-AV]: This follows the algorithm we introduced in Section II-A, while replacing our attention-based aggregation method in (4) and (5) with the averaging-based aggregation of PerFed [23].

• [sc1-P-AT]: This is the full version of our course-based adaptive method from Section II-A.

2) Algorithms for scenario II (<Scenario> = sc2):
Local Modeling (<Architecture> = L) In scenario II, the bottom hierarchy is grouped by demographic variables. Therefore, we build a local baseline [sc2-L] with a separate model for each of these demographic subgroups. Each model in [sc2-L] is trained on student data $\Omega_{\{c,I,x\}}^{\text{train}}$ for a specific subgroup $\{c, I, x\}$ (i.e., subgroup x of demographic variable I in course c) and evaluated on $\Omega_{\{c,I,x\}}^{\text{test}}$.

Global Modeling (<Architecture> = G) We also consider a centralized global [sc2-G] and several federated global models [sc2-G-AV-M] & [sc2-G-AT-M] & [sc2-G-AV-T] & [sc2-G-AT-T]. For the federated global models, we first train several local models based on the demographic subgroups within each course. Then, we aggregate them to form different global models at each level (i.e., course(middle)-level global models ([sc2-G-AV-M] & [sc2-G-AT-M]) and top level global models ([sc2-G-AV-T] & [sc2-G-AT-T])).

• [sc2-G]: Architecturally, this is the same as [sc1-G] because they both use all students Ω^{train} to build one model. However, we evaluate [sc2-G] on each demographic subgroup $\Omega_{\{c,I,x\}}^{\text{test}}$ instead of each course.

• [sc2-G-AV-M]: In this case, the demographic models are first locally trained. Then, this baseline obtains the aggregated course(middle)-level model Θ_c^K at round K using the weighted average method of FedAvg [19]:

$$\Theta_{\text{c}}^{K+1} = \sum_{x \in \mathcal{X}} \frac{N_x}{N_c} \Theta_x^{(K,E)}, \tag{22}$$

where $\Theta_x^{(K,E)}$ is a demographic model trained locally, N_x is the number of students within subgroup x, N_c is the number of students within course c, and $N_c = \sum_{x \in \mathcal{X}} N_x$.

• [sc2-G-AT-M]: This is the course(middle)-level model Θ_g^{K+1} from (8), that aggregates local demographic models via the attention-based aggregation method.

• [sc2-G-AV-T]: After obtaining the course(middle)-level global models from (22), we aggregate them again to the top level global model. Specifically, [sc2-G-AV-T] is obtained by aggregating Θ_c^K as:

$$\Theta_g^{K+1} = \sum_{c \in \mathcal{C}} \frac{N_c}{N} \Theta_c^{K+1}. \tag{23}$$

• [sc2-G-AT-T]: This applies the attention aggregation mechanism to [sc2-G-AT-M] and obtains the top-level global model based on (10).

Personalized Modeling (<Architecture> = P) We consider five personalized algorithms in this scenario. One of the baselines, FedIRT, uses educational theory for the aggregation, while four – [sc2-P-AV-M], [sc2-P-AT-M], [sc2-P-AT-B], and [sc2-P-AT-B]– are different configurations of our MLPFL algorithm. [sc2-P-AV-M] and [sc2-P-AT-M] are the course(middle)-level personalizations adapted from the top global model, while [sc2-P-AV-B] and [sc2-P-AT-B] are the corresponding demographic(bottom)-level personalizations adapted from the course(middle)-level models.

• FedIRT: This is based on the federated deep knowledge tracing (FDKT) [17] method discussed in Section I. FDKT employs classical test theory and item response theory (IRT) [41] to compute a score that reflects the "data quality" of local subgroups, which then influences the weighting of each subgroup during model aggregation. For our setting, we execute the model update and aggregation process by

computing the IRT confidence α_x^{IRT} for each student subgroup x in scenario II, using the students' responses to the end-of-video quizzes. Within each course, local models are trained for each demographic variable. Following [17], local model updates use interpolation, where the initial model for training round k is defined as:

$$\Theta_x^{(k,0)} = \lambda_x^{(k)}\Theta_x^{(k-1,E)} + (1-\lambda_x^{(k)})\Theta_g^{(k)}, \tag{24}$$

where

$$\lambda_x^{(k)} = \left(\Theta_x^{(k-1,E)} \cdot \Theta_{\text{g}}^{(k)}\right) / \left(\|\Theta_x^{(k-1,E)}\| \times \|\Theta_{\text{g}}^{(k)}\|\right), \tag{25}$$

and the final local model $\Theta_x^{(k,E)}$ is obtained after E local epochs of conventional gradient descent. The aggregated global model is then computed as

$$\Theta_{\text{g}}^{(k+1)} = \sum_{x\in\mathcal{X}} \alpha_x^{\text{IRT}}\Theta_x^{(k,E)}. \tag{26}$$

- `[sc2-P-AV-M]`: This is built by taking one meta-update from `[sc2-G-AV-T]`, i.e., following (7). This creates one personalized model for each course, using stratified sampling of students in each subgroup for each course.
- `[sc2-P-AT-M]`: This baseline is adapted from `[sc2-G-AT-T]` by taking one meta-update.
- `[sc2-P-AV-B]`: This is constructed through one further meta-update step from `[sc2-P-AV-M]`. It uses the datasets from each demographic subgroup to create a separate model for each subgroup in each course.
- `[sc2-P-AT-B]`: Taking one further meta-update step from `[sc2-P-AT-M]`, `[sc2-P-AT-B]` corresponds to the full MLPFL method we introduced in Section II-B2- II-B3.

B. Experimental Results

In Tables III and IV, we compare the predictive quality of knowledge tracing and outcome prediction for each method across courses in scenario I. For scenario II, Tables VII, VIII, and IX compare the performance of knowledge tracing on each student group, while Tables XIV, XV, and XIII in Appendix C are the results for outcome prediction. Additionally, Tables V and VI provide insights into early prediction performance using different time-frames of students' activities.

In this section, all results are shown on the set of students who provided specific demographic information I; our models are trained on $\Omega_{c,I}^{\text{train}}$ and evaluated on each subgroup of $\Omega_{c,I}^{\text{test}}$. More specifically, for the experiments in Tables VII&XIV, VIII&XV, and IX&XIII, the dataset is partitioned into groups according to gender, continent, and age, respectively. We present the results for students who did not disclose their personal information (i.e., the models trained on $\{\Omega_{c,I}^{\text{train}}, \Omega_{\neg I}^{\text{train}}\}$ and tested on each subgroup of $\{\Omega_{c,I}^{\text{test}}, \Omega_{\neg I}^{\text{test}}\}$) to Appendix D; these results exhibit similar qualitative patterns.

1) Discussion for scenario I: From Tables III and IV, we see that in many cases, there is not a large difference in performance between the locally trained `[sc1-L]` for each course and the centrally trained `[sc1-G]` across courses. This is a key motivation for course adaptation: despite `[sc1-G]` containing significantly larger training data, it is not preserving course-specific information. On the other hand, `[sc1-G]` generally outperforms both versions of global FL (`[sc1-G-AV]` and `[sc1-G-AT]`), indicating that centralized training across the entire dataset is more effective than decentralized models in scenarios without adaptation.

TABLE III
AUC RESULTS FOR KNOWLEDGE TRACING OF DIFFERENT MODELS

Model	FOC	QDS	MOSFETs
`[sc1-L]`	.533 (.013)	.537 (.009)	.529 (.011)
`[sc1-G]`	.528 (.007)	.529 (.010)	.531 (.011)
`[sc1-G-AV]`	.521 (.010)	.517 (.009)	.529 (.011)
`[sc1-G-AT]`	.526 (.008)	.525 (.012)	.519 (.007)
`[sc1-P-AV]`	.614 (.016)	.618 (.009)	.603 (.014)
`[sc1-P-AT]`(MLPFL)	**.621 (.019)**	**.625 (.015)**	**.619 (.014)**

TABLE IV
AUC RESULTS FOR OUTCOME PREDICTION OF DIFFERENT MODELS.

Model	FOC	QDS	MOSFETs
`[sc1-L]`	.553 (.021)	.558 (.017)	.549 (.019)
`[sc1-G]`	.547 (.017)	.559 (.010)	.553 (.013)
`[sc1-G-AV]`	.550 (.013)	.546 (.017)	.551 (.009)
`[sc1-G-AT]`	.548 (.018)	.556 (.011)	.548 (.015)
`[sc1-P-AV]`	.678 (.020)	.689 (.011)	.658 (.017)
`[sc1-P-AT]`(MLPFL)	**.684 (.014)**	**.701 (.021)**	**.679 (.015)**

Across Tables III and IV, our proposed adaptive models (`[sc1-P-AV]` and `[sc1-P-AT]`) obtain anywhere from 15-25% improvement in AUC over the non-adaptive models, depending on the course, aggregation, and downstream task. The improvements obtained by the adaptive models show the significance of adapting the global models based on local students within each course. Compared to `[sc1-P-AV]`, MLPFL (`[sc1-P-AT]`) performs better on most courses, confirming our intuition on the advantages of layer-wise model aggregations with an attention mechanism. *Importantly, these findings suggest that even without demographic information, MLPFL is able to obtain substantial improvement in student modeling through course-level personalizations.*

Tables III and IV demonstrate that our method outperforms other baselines in two downstream tasks. We further analyze the breakdown of performance by demographic, particularly when demographic information is not considered in the modeling. Figure 3 shows the outcome prediction performance by gender across three courses using our method. Although our method generally improves performance for each course, significant variations are observed between demographic subgroups. This observation motivates the need for Scenario II, which incorporates demographic information into the modeling to address these disparities more effectively.

In addition to the outcome prediction that utilizes all activities from each student throughout the entire course (Table IV), we have expanded our analysis to include early prediction

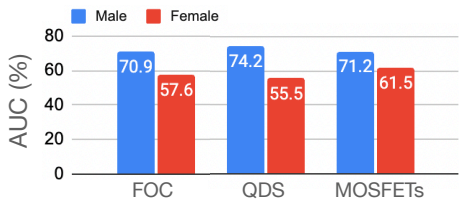


Fig. 3. Breakdown of outcome prediction performance by gender for Scenario I using our method, MLPFL (`[sc1-P-AT]`).

TABLE V
AUC RESULTS FOR EARLY PREDICTION USING THE FIRST 3 WEEKS OF DATA, ACROSS DIFFERENT MODELS.

Model	FOC	QDS	MOSFETs
[sc1-L]	.462	.459	.488
[sc1-G]	.473	.487	.499
[sc1-G-AV]	.470	.469	.489
[sc1-G-AT]	.479	.478	.500
[sc1-P-AV]	.498	.503	.518
[sc1-P-AT](MLPFL)	**.519**	**.534**	**.537**

TABLE VI
AUC RESULTS FOR EARLY PREDICTION USING HALF OF EACH STUDENT'S DATA, ACROSS DIFFERENT MODELS.

Model	FOC	QDS	MOSFETs
[sc1-L]	.487	.473	.491
[sc1-G]	.499	.508	.489
[sc1-G-AV]	.503	.486	.512
[sc1-G-AT]	.498	.511	.505
[sc1-P-AV]	.539	.553	.549
[sc1-P-AT](MLPFL)	**.568**	**.592**	**.583**

tasks. Specifically, we focused on two time-frames: (a) using only the first half of each student's recorded interactions in the course (Table VI) and (b) using only the first three weeks' worth of data, out of approximately 17 weeks total for each course (Table V). Compared with using the full duration of data, the results show a slight decrease in AUC scores for each method. However, our MLPFL continues to outperform other baselines in these early prediction scenarios, demonstrating its effectiveness even with limited data.

2) Discussion for scenario II: In Tables VII, VIII, and IX, for knowledge tracing, we observe that the local model [sc2-L] provides on average 54% AUC on each subgroup according to different demographic groupings, and in Tables XIV, XV, and XIII, we find on average 60% AUC for outcome prediction. While providing moderate improvements over [sc1-L], these results confirm our hypothesis that small subgroups do not have enough data to train high accuracy models individually. On the other hand, unlike in scenario I, the centralized global model [sc2-G] does not necessarily outperform the global federated models ([sc2-G-AV-M]&[sc2-G-AT-M] and [sc2-G-AV-T]&[sc2-G-AT-T]). This is particularly true for underrepresented subgroups. For example, in Table XIV, for the MOSFETs dataset, [sc2-G] obtains the highest performance for the male subgroup among the non-personalized models, while two of the global FL models noticeably outperform it on the female subgroup. As another example, in Table XV, for the QDS dataset, [sc2-G-AT-M] outperforms [sc2-G] noticeably for all subgroups except North America. By incorporating such subgroup heterogeneity, the personalized models are able to obtain improvements across datasets and subgroups, as we discuss further next.

For both the knowledge tracing task (Tables VII, VIII, and IX) and the outcome prediction task (Tables XIV, XV, and XIII), *our MLPFL obtains substantial improvements over the non-personalized models on each demographic subgroup*. Specifically, for the FOC and QDS datasets, we obtain 15-35% improvement in AUC over the highest performing non-personalized model across subgroups. The improvements obtained by the personalized models are noticeable but more limited in the MOSFETs dataset, possibly due to its overall smaller size (see Table I). For both tasks, MLPFL consistently outperforms FedIRT except in 4 out of approximately 50 subgroup cases across courses, underscoring the advantages of our meta learning-based personalization strategy. This enhancement confirms our premise that assigning weights to student subgroups according to their "data quality" as done in FedIRT might result in the loss of crucial information regarding subgroup diversity, a concern addressed by our redefinition in (6). Moreover, FedIRT does not consider course-specific modeling, whereas our personalized architecture adapts the student models across different courses and demographics.

These results also show how our methodology benefits from the personalization hierarchy. Specifically, in conducting the meta-updates that move from [sc2-G-AV-T] to [sc2-P-AV-M] and [sc2-G-AT-T] to [sc2-P-AT-M], we find that the course-level personalized models outperform the global models by at least 15% AUC for both downstream tasks. By considering the lowest-layer subgroupings and performing additional meta-updates that move from [sc2-P-AV-M] to [sc2-P-AV-B] and [sc2-P-AT-M] to [sc2-P-AT-B], demographic-level personalized models further improve the AUC by 7%-9% on the FOC and QDS datasets over course-level personalized models. Bulding upon our MLPFL results for scenario I, this shows how our personalization methodology benefits from increasing amounts of information available about students, whereas the non-personalized FL models do not show this trend when moving from top to middle layer modeling.

To directly assess the bias mitigation in Scenario II, in Table X, we show the standard deviation in AUC (expressed as a percentage of the mean) across each demographic group obtained by different algorithms. A model with less bias for a demographic variable will have lower standard deviation across the subgroup, as it indicates closer performance achieved across the different categories. Compared to the globally trained [sc1-G] model, we see that our algorithms in Scenario I actually slightly increase the standard deviation across demographic groups, since these characteristics are not explicitly accounted for. In contrast, our algorithms in Scenario II lead to substantial reductions in standard deviation for all demographic groups in each of the courses and tasks compared to the global model (65-70%). Thus, when furnished with demographic information, our approach can both mitigate prediction biases for minority groups while leading to improvements in performance for all groups.

Finally, Figure 4 presents both the mean and variance in prediction quality for each demographic subgroup obtained by FedIRT and our bottom-layer MLPFL methodology, averaged across courses. We can see that MLPFL consistently provides an improvement in mean AUC over FedIRT, with the attention mechanism providing an extra boost. Importantly, we also see that [sc2-P-AT-B] reduces the AUC variance across subgroups in five of the six cases. The exception case is the one with the lowest variance across all methods. *The increase in mean and reduction in variance of prediction quality provided*

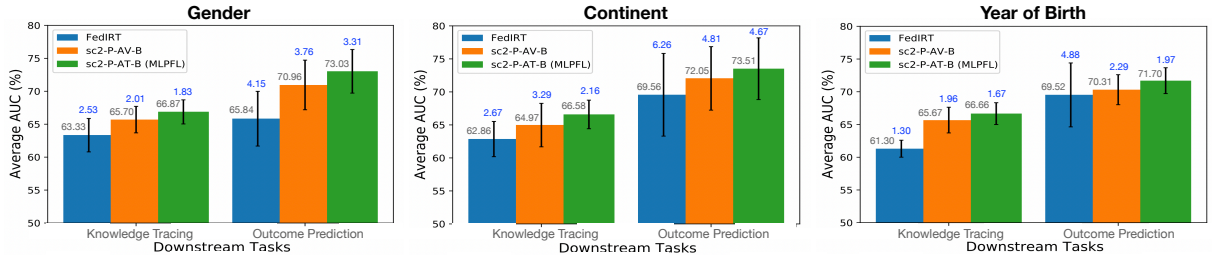


Fig. 4. Mean and standard deviation (in blue text) AUC scores of personalized methods (`FedIRT`, `[sc2-P-AV-B]`, `[sc2-P-AT-B]`) across all subgroups in three courses. The increased mean AUC scores and the reduced variances show that MLPFL improves prediction quality while mitigating biases.

TABLE VII
THE PERFORMANCE ON KNOWLEDGE TRACING OBTAINED BY DIFFERENT MODELS ON EACH GENDER SUBGROUP.

Dataset	FOC		QDS		MOSFETs	
Model	male	female	male	female	male	female
`[sc2-L]`	.532 (.009)	.519 (.004)	.541 (.010)	.523 (.006)	.538 (.011)	.519 (.003)
`[sc2-G]`	.539 (.005)	.557 (.012)	.538 (.008)	.543 (.006)	.567 (.009)	.551 (.010)
`[sc2-G-AV-M]`	.521 (.003)	.539 (.008)	.547 (.010)	.551 (.013)	.543 (.011)	.538 (.008)
`[sc2-G-AT-M]`	.529 (.002)	.534 (.012)	.552 (.013)	.546 (.015)	.557 (.017)	.568 (.020)
`[sc2-G-AV-T]`	.517 (.002)	.526 (.008)	.510 (.004)	.527 (.009)	.530 (.006)	.526 (.011)
`[sc2-G-AT-T]`	.523 (.008)	.515 (.006)	.523 (.013)	.538 (.018)	.549 (.023)	.530 (.017)
`FedIRT`	.667 (.018)	.652 (.025)	.621 (.023)	.667 (.022)	.609 (.015)	.598 (.011)
`[sc2-P-AV-M]`	.665 (.029)	.630 (.024)	.649 (.031)	.651 (.028)	.624 (.022)	.602 (.018)
`[sc2-P-AT-M]`(MLPFL)	.680 (.033)	.633 (.028)	.655 (.024)	.668 (.031)	.637 (.027)	.620 (.019)
`[sc2-P-AV-B]`	.679 (.033)	**.661 (.028)**	.658 (.026)	.671 (.031)	.630 (.017)	.625 (.019)
`[sc2-P-AT-B]`(MLPFL)	**.692 (.031)**	.654 (.028)	**.671 (.028)**	**.684 (.024)**	**.644 (.021)**	**.631 (.016)**

TABLE VIII
PREDICTION PERFORMANCE ON KNOWLEDGE TRACING OBTAINED WITH DIFFERENT MODELS ON EACH STUDENT SUBGROUP GROUPED BY CONTINENT. AS, AF, EU, NA, AND SA REPRESENT ASIAN, AFRICAN, EUROPEAN, NORTH AMERICAN, AND SOUTH AMERICAN SUBGROUPS.

Dataset	FOC				
Model	AS	AF	EU	NA	SA
`[sc2-L]`	.521 (.007)	.519 (.005)	.536 (.008)	.517 (.004)	.532 (.009)
`[sc2-G]`	.537 (.009)	.546 (.013)	.535 (.007)	.521 (.005)	.530 (.010)
`[sc2-G-AV-M]`	.555 (.009)	.562 (.013)	.559 (.011)	.538 (.007)	.556 (.013)
`[sc2-G-AT-M]`	.578 (.015)	.563 (.016)	.557 (.013)	.569 (.011)	.568 (.009)
`[sc2-G-AV-T]`	.546 (.007)	.532 (.011)	.543 (.013)	.550 (.010)	.567 (.008)
`[sc2-G-AT-T]`	.557 (.016)	.549 (.018)	.561 (.020)	.548 (.013)	.557 (.015)
`FedIRT`	.601 (.007)	.637 (.019)	.605 (.010)	.649 (.016)	.633 (.015)
`[sc2-P-AV-M]`	.614 (.020)	.603 (.024)	.627 (.023)	.600 (.018)	.611 (.016)
`[sc2-P-AT-M]`(MLPFL)	.628 (.021)	.617 (.026)	.639 (.024)	.624 (.020)	.643 (.024)
`[sc2-P-AV-B]`	.640 (.023)	**.653 (.027)**	.635 (.013)	**.648 (.017)**	.647 (.025)
`[sc2-P-AT-B]`(MLPFL)	**.652 (.023)**	.648 (.019)	**.667 (.015)**	.643 (.023)	**.652 (.021)**
Dataset	QDS				
Model	AS	AF	EU	NA	SA
`[sc2-L]`	.510 (.004)	.533 (.006)	.548 (.008)	.564 (.010)	.523 (.007)
`[sc2-G]`	.556 (.006)	.533 (.011)	.548 (.012)	.540 (.015)	.536 (.010)
`[sc2-G-AV-M]`	.547 (.016)	.583 (.009)	.546 (.013)	.557 (.018)	.589 (.021)
`[sc2-G-AT-M]`	.583 (.013)	.554 (.011)	.562 (.015)	.567 (.019)	.565 (.016)
`[sc2-G-AV-T]`	.532 (.010)	.568 (.018)	.521 (.007)	.556 (.013)	.574 (.018)
`[sc2-G-AT-T]`	.564 (.015)	.576 (.014)	.543 (.016)	.575 (.013)	.580 (.009)
`FedIRT`	.659 (.019)	.632 (.022)	.655 (.017)	.619 (.015)	.628 (.017)
`[sc2-P-AV-M]`	.627 (.025)	.604 (.019)	.632 (.016)	.628 (.020)	.617 (.023)
`[sc2-P-AT-M]`(MLPFL)	.651 (.015)	.638 (.018)	.667 (.022)	.662 (.024)	.654 (.016)
`[sc2-P-AV-B]`	.670 (.028)	.654 (.025)	.668 (.014)	.658 (.016)	**.677 (.017)**
`[sc2-P-AT-B]`(MLPFL)	**.698 (.018)**	**.671 (.015)**	**.683 (.020)**	**.692 (.027)**	.674 (.023)
Dataset	MOSFETs				
Model	AS	AF	EU	NA	SA
`[sc2-L]`	.531 (.010)	.518 (.004)	.510 (.006)	.537 (.008)	.555 (.015)
`[sc2-G]`	.528 (.013)	.537 (.018)	.550 (.015)	.542 (.007)	.536 (.009)
`[sc2-G-AV-M]`	.548 (.015)	.532 (.007)	.569 (.006)	.547 (.008)	.563 (.015)
`[sc2-G-AT-M]`	.569 (.017)	.583 (.016)	.552 (.010)	.543 (.009)	.538 (.010)
`[sc2-G-AV-T]`	.532 (.010)	.518 (.015)	.547 (.009)	.535 (.004)	.533 (.013)
`[sc2-G-AT-T]`	.553 (.015)	.547 (.008)	.532 (.016)	.514 (.005)	.528 (.008)
`FedIRT`	.592 (.008)	.620 (.010)	**.617 (.015)**	.584 (.009)	.603 (.015)
`[sc2-P-AV-M]`	.588 (.016)	.563 (.015)	.551 (.020)	.549 (.018)	.543 (.012)
`[sc2-P-AT-M]`(MLPFL)	.607 (.025)	.584 (.017)	.586 (.015)	.553 (.010)	.567 (.015)
`[sc2-P-AV-B]`	.624 (.018)	**.632 (.016)**	.615 (.020)	.602 (.014)	.607 (.016)
`[sc2-P-AT-B]`(MLPFL)	**.632 (.018)**	.618 (.012)	.600 (.023)	**.613 (.015)**	**.629 (.020)**

by MLPFL confirms that it lessens the impact of data subgroup availability biases in student modeling.

C. Embedding Visualization

We utilize t-SNE to visualize the acquired student activity representations in a 2D space, aiming to qualitatively evaluate our approach. t-SNE (t-distributed stochastic neighbor embedding) is a dimensionality reduction technique used to visualize high-dimensional data in two or three dimensions. With t-SNE, datapoints that are statistically close together (i.e., similar) in the original space will be close to one another in the mapped space with high probability, while those far apart (i.e., dissimilar) in the original space will have a low probability of being close in the mapped space. Thus, t-SNE helps reveal similarity patterns in a high-dimensional dataset, such as clusters. For our purposes, a method's embeddings can be considered "better" if its t-SNE visualizations have more discernible clusters, with a high tendency of the points in each cluster to be from a specific course (in Scenario I) or course-demographic pair (in Scenario II): this indicates that the embeddings contain activity patterns that are important to modeling the differences between users in different courses and demographics for personalized analytics.

In Figure 5, we plot the student embeddings for knowledge tracing according to different courses in scenario I, while Figure 6 shows the embeddings based on the gender demographic grouping in scenario II. We defer the visualizations for other demographic subgroups in scenario II to Appendix E; the results are qualitatively similar. All models are trained using Ω^{train}, and student embeddings are derived from the hidden state $\widetilde{\mathbf{h}}_u$ (Section III-B2), which is obtained after applying the attention module to predict their learning outcomes. The colors of the dots represent the corresponding student groups.

In Figure 5(a), the centralized global model does not generate embeddings that differentiate students in different courses for scenario I. In Figure 5(b), unsurprisingly, local modeling produces embeddings which cluster by course with linear separation between the clusters. Compared to these, the embeddings learned by FL (Figure 5(c)&(d)) also demonstrate clustering patterns according to different courses. In Figure 5(d), the distribution learned by MLPFL is more separated by course due to the meta-learning adaptation procedure applied to the model in Figure 5(c).

For scenario II, in Figure 6, it is hard to differentiate

TABLE IX
THE KNOWLEDGE TRACING PERFORMANCE OBTAINED WITH DIFFERENT MODELS ON EACH STUDENT SUBGROUP GROUPED BY AGE. < 80, 80-90, >90 REPRESENT YEAR OF BIRTH PRIOR TO 1980, BETWEEN 1980 AND 1990, AND AFTER 1990, RESPECTIVELY.

Dataset	FOC			QDS			MOSFETs		
Model	< 80	$80-90$	> 90	< 80	$80-90$	> 90	< 80	$80-90$	> 90
[sc2-L]	.539 (.011)	.527 (.008)	.535 (.006)	.528 (.007)	.536 (.005)	.523 (.008)	.536 (.010)	.548 (.011)	.531 (.010)
[sc2-G]	.553 (.007)	.539 (.005)	.548 (.009)	.536 (.007)	.547 (.010)	.551 (.011)	.568 (.011)	.543 (.008)	.556 (.014)
[sc2-G-AV-M]	.554 (.008)	.546 (.007)	.530 (.008)	.541 (.010)	.559 (.010)	.537 (.003)	.560 (.008)	.546 (.007)	.553 (.011)
[sc2-G-AT-M]	.562 (.015)	.557 (.017)	.549 (.008)	.558 (.009)	.567 (.013)	.564 (.015)	.572 (.018)	.568 (.015)	.570 (.019)
[sc2-G-AV-T]	.532 (.008)	.546 (.011)	.538 (.009)	.557 (.009)	.552 (.011)	.546 (.013)	.553 (.015)	.532 (.012)	.548 (.013)
[sc2-G-AT-T]	.546 (.015)	.563 (.018)	.551 (.012)	.562 (.009)	.549 (.013)	.538 (.008)	.546 (.008)	.558 (.015)	.563 (.019)
FedIRT	.610 (.015)	.639 (.017)	.614 (.015)	.631 (.013)	.600 (.011)	.617 (.015)	.608 (.017)	.592 (.009)	.599 (.007)
[sc2-P-AV-M]	.638 (.023)	.652 (.018)	.661 (.018)	.630 (.021)	.643 (.024)	.650 (.019)	.638 (.017)	.620 (.019)	.625 (.021)
[sc2-P-AT-M](MLPFL)	.654 (.017)	.661 (.015)	.673 (.022)	.647 (.019)	.632 (.018)	.656 (.023)	.640 (.015)	.631 (.021)	.617 (.024)
[sc2-P-AV-B]	.672 (.029)	.669 (.025)	.680 (.028)	**.660 (.025)**	.672 (.019)	.648 (.020)	.631 (.017)	.650 (.021)	.628 (.020)
[sc2-P-AT-B](MLPFL)	**.689 (.025)**	**.672 (.022)**	**.699 (.018)**	.654 (.015)	**.683 (.020)**	**.661 (.027)**	**.648 (.022)**	**.657 (.017)**	**.631 (.019)**

TABLE X
PERFORMANCE VARIATION (STANDARD DEVIATION) ACROSS THE GENDER (G), CONTINENT (C), AND YEAR OF BIRTH (Y) DEMOGRAPHIC VARIABLES. THE STANDARD DEVIATION IS REPORTED AS A PERCENTAGE OF THE MEAN.

Task	Knowledge Tracing									Outcome Prediction								
Dataset	FOC			QDS			MOSFETs			FOC			QDS			MOSFETs		
Model	G	C	Y	G	C	Y	G	C	Y	G	C	Y	G	C	Y	G	C	Y
[sc1-G]	5.29	5.64	4.52	8.45	4.23	5.34	4.56	3.34	3.85	7.39	4.56	3.45	9.28	3.95	4.98	5.67	4.44	3.22
[sc1-P-AT](MLPFL)	7.62	4.97	5.67	9.25	6.66	5.99	7.22	3.49	4.58	9.40	3.96	3.95	13.2	4.52	3.88	6.85	5.42	4.98
[sc2-P-AT-M](MLPFL)	3.32	1.07	**0.97**	0.92	1.15	**1.21**	1.20	2.04	**1.16**	1.77	**1.77**	**0.76**	**1.20**	2.16	1.41	3.11	1.82	1.30
[sc2-P-AT-B](MLPFL)	**2.68**	**0.89**	1.55	**0.92**	**1.15**	1.51	**0.92**	**1.29**	1.32	**0.21**	1.84	1.20	1.48	**1.54**	**1.08**	**2.40**	**1.23**	**1.15**

different courses and subgroups from the centralized global model (Figure 6(a)). Starting with the federated global model (Figure 6(b)), a more clustered distribution emerges as personalization steps are taken to the middle and bottom of the hierarchy in Figure 6(c) and 6(d). Compared to the course-level personalized model in Figure 6(c), the MLPFL model personalized by demographic information in Figure 6(d) shows even more well-separated clusters due to additional meta-updates. The same clustering pattern when personalizing according to different continent and age information can be seen from Figure 7 in Section D of the Appendix.

Moreover, these visualizations indicate a reasonable level of correlation between activity embeddings and the respective courses and demographic subgroups. For example, considering gender (Figure 6(c)), males from course 3 tend to cluster on the left side, while females from course 2 are often positioned towards the top, and females from course 1 are located on the upper right part of the figure. Similar patterns are also evident in the remaining subplots of Figure 7(a3)&(b3) in the Appendix. *These results confirm that distinct learning behaviors exist among various student subgroups, which our meta-learning-based personalization method addresses by adapting global models to accommodate these differences.*

V. DISCUSSION ON REAL-WORLD IMPLEMENTATION AND APPLICATIONS

We now discuss how our MLPFL methodology can be integrated into online education platforms, such as the edX platform which hosted the courses we evaluated on in Section IV, to mitigate biases in predictive learning analytics. The most direct application is to integrate MLPFL into the learning management system (LMS) backends, and provide the prediction outputs to students and instructors on their corresponding dashboards. For the outcome prediction and knowledge tracing tasks evaluated in our work, this would include inferences of which students are expected to answer specific in-video quiz questions incorrectly (from knowledge tracing), and which students are at risk of not passing the course (from outcome prediction). Based on this information, instructors could plan targeted interventions to help individual students (or a group of students) in need, e.g., as discussed in [42]. We envision these predictions to be updated regularly as students generate more activities in their courses, and to include a measure of confidence in the prediction, derived from the performance indicators observed during the model training procedure (e.g., the AUC values displayed in our analysis).

To evaluate the effectiveness of MLPFL-enhanced dashboards, it will be important to conduct proper A/B testing. We envision establishing three groups for comparison: (A) courses without predictive analytics in dashboards, (B) courses with baseline predictive learning analytics solutions (e.g., FedIRT, [sc2-L], [sc2-G] from Section IV) integrated into dashboards, and (B) courses with MLPFL integrated in dashboards for mitigating biases. After running such an experiment, one can compare the resulting distributions of content engagement, exam grades, pass/fail results, and other relevant metrics between these groups, with the aim of showing that group (C) has the most successful students, and importantly, the lowest variance across demographic subgroups. One can also quantify the impact of interventions staged by the instructor in each group by comparing student performance before and after such interventions took place. Additionally, qualitative feedback from students and instructors can offer valuable insights into how the dashboard analytics impacted the course experience.

Another application of MLPFL that we envision is to mitigate biases in AI-enabled adaptive learning systems [43], [44] that have been integrated into online education platforms. Many of these systems rely on identifying students at risk of early dropout and/or poor grades to determine refinements of their suggested learning paths [44], [45], including but not limited to providing additional resources on particular topics and personalized quiz difficulties. The aim of MLPFL in these systems will be to improve and reduce the variance in quality of these identifications across demographic subgroups. The evaluation of MLPFL here should similarly follow proper

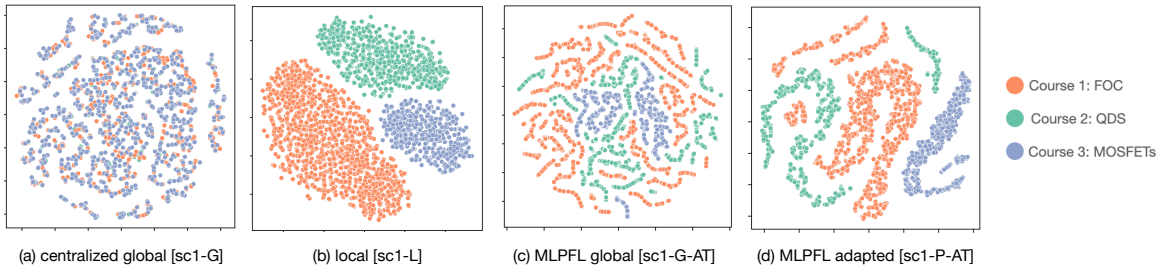


Fig. 5. The embedding vectors of students' knowledge state learned by the centralized method (`[sc1-G]`), local model (`[sc1-L]`), federated global model of MLPFL (`[sc1-G-AT]`), and the personalized model of MLPFL (`[sc1-P-AT]`) according to different courses in scenario I.

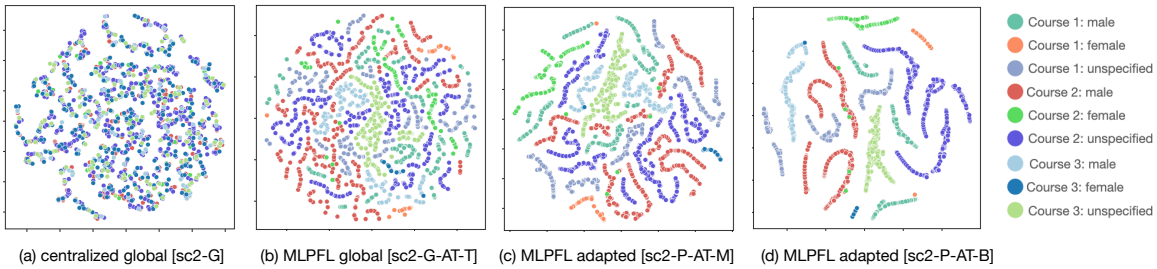


Fig. 6. Students' knowledge state representations from the centralized method (`[sc2-G]`) and different hierarchies of our MLPFL in scenario II. The organized and even clustered patterns learned by `[sc2-P-AT-B]` are consistent with its design to learn unique representations for each subgroup.

A/B testing procedures, where the learning outcomes of an adaptive learning system with and without the integration of our solution are compared.

VI. Discussion on Limitations

We finally discuss some limitations of our study which can motivate future work. One limitation is its dependency on the availability of demographic data, which students may not always be willing to provide. Even in the three courses we studied, a substantial portion of students chose not to specify their gender or age (see Table XII). Such limitations can affect MLPFL's ability to accurately personalize learning analytics for all students; indeed, we saw in Table X that a key disadvantage of scenario I (which does not employ demographic information) is that it can actually increase performance biases slightly. Nonetheless, our results have shown that even when a significant amount of demographic data is missing, MLPFL in scenario II is able to obtain substantial improvements in prediction quality and performance variance across student subgroups. Related to this, another limitation of our study is that our grouping of demographic variables was largely based on intuition from the available data, and may not be appropriate for all course populations. In particular, we modeled the geographic variable based on continent since many countries did not have more than a few samples, and we chose three groups for the age variable somewhat arbitrarily. Future work could investigate a technique for rigorously optimizing the partitioning of a subgroup variable according to performance and bias mitigation objectives.

VII. Conclusion

In this paper, we developed a Multi-Layer Personalized Federated Learning (MLPFL) framework for mitigating biases in student modeling stemming from data availability. Our approach is based on meta learning, adapting local models for various student subgroups from a common global model designed to promote personalization capability. We applied our framework to two student modeling tasks: (i) knowledge tracing and (ii) outcome prediction, modeling student activities when they interact with an online learning platform. Our proposed method considers personalization in a hierarchical manner: first by course (if there are multiple courses), and then by student demographic subgroup within each course (if this information is provided to the model). Evaluation on three online course datasets showed that our approach surpasses baseline methods by enhancing prediction accuracy across courses and student subgroups, both in terms of mean and variance. Moreover, well-organized student embeddings learned by our method were seen to be correlated with improved student modeling.

References

[1] Y.-W. Chu, S. Hosseinalipour, E. Tenorio, L. Cruz, K. Douglas, A. Lan, and C. Brinton, "Mitigating biases in student performance prediction via attention-based personalized federated learning," *ACM Int. Conf. Info. Knwl. Mgmt. (CIKM)*, 2022.

[2] O. B. Adedoyin and E. Soykan, "Covid-19 pandemic and online learning: the challenges and opportunities," *Interactive Learning Environments*, pp. 1–13, 2020.

[3] J. Zhang, X. Shi, I. King, and D. Yeung, "Dynamic key-value memory networks for knowledge tracing," *The World Wide Web Conf.*, 2017.

[4] C.-Y. Chou, S.-F. Tseng, W.-C. Chih, Z.-H. Chen, P.-Y. Chao, K. R. Lai, C.-L. Chan, L.-C. Yu, and Y.-L. Lin, "Open student models of core competencies at the curriculum level: Using learning analytics for student reflection," *IEEE Transactions on Emerging Topics in Computing*, vol. 5, pp. 32–44, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:17787963

[5] K. VanLehn, "Student modeling," *Found. Intell. Tut. Syst.*, vol. 55, p. 78, 1988.

[6] W. J. van der Linden and R. K. Hambleton, *Handbook of modern item response theory*. Springer Science and Business Media, 2013.

[7] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User modeling and user-adapted interaction*, vol. 4, no. 4, pp. 253–278, 1994.

[8] T.-Y. Yang, R. S. Baker, C. Studer, N. Heffernan, and A. S. Lan, "Active learning for student affect detection," in *Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019, Montréal, Canada, July 2-5, 2019. International Educational Data Mining Society (IEDMS) 2019*. Université du Québec; Polytechnique Montréal, 2019, pp. 208–217.

[9] W.-L. Chan and D.-Y. Yeung, "Clickstream knowledge tracing: Modeling how students answer interactive online questions," in *Int. Learn. Analytics and Knowl. Conf.*, 2021, pp. 99–109.

[10] R. R. Sahay, S. Nicoll, M. Zhang, T.-Y. Yang, C. Joe-Wong, K. A. Douglas, and C. G. Brinton, "Predicting learning interactions in social learning networks: A deep learning enabled approach," *IEEE/ACM Transactions on Networking*, vol. 31, pp. 2086–2100, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:255415930

[11] Y. Fan, L. Shepherd, E. Slavich, D. Waters, M. Stone, R. Abel, and E. Johnston, "Gender and cultural bias in student evaluations: Why representation matters," *PloS ONE*, vol. 14, no. 2, p. e0209749, 2019.

[12] R. F. Kizilcec and H. Lee, "Algorithmic fairness in education," *arXiv preprint arXiv:2007.05443*, 2020.

[13] S. Yao and B. Huang, "Beyond parity: Fairness objectives for collaborative filtering," *NIPS*, 2017.

[14] L. Paquette, J. Ocumpaugh, Z. Li, A. Andres, and R. Baker, "Who's learning? using demographics in edm research." *Journal of Educational Data Mining*, vol. 12, no. 3, pp. 1–30, 2020.

[15] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conf. Fairness, Accountability and Transparency*, 2018, pp. 77–91.

[16] P. Lahoti, K. P. Gummadi, and G. Weikum, "ifair: Learning individually fair data representations for algorithmic decision making," in *Int. Conf. Data Engrg.* IEEE, 2019, pp. 1334–1345.

[17] J. Wu, Z. Huang, Q. Liu, D. Lian, H. Wang, E. Chen, H. Ma, and S. Wang, "Federated deep knowledge tracing," *ACM Int. Conf. Web Search and Data Mining*, 2021.

[18] J. Konecný, H. B. McMahan, F. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *ArXiv*, vol. abs/1610.05492, 2016.

[19] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Int. Conf. Artif. Intell. and Stats.*, 2017.

[20] S. Ji, S. Pan, G. Long, X. Li, J. Jiang, and Z. Huang, "Learning private neural language modeling with attentive aggregation," *Int. Joint Conf. Neural Netw.*, pp. 1–8, 2019.

[21] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei, and Y. Zhang, "Personalized cross-silo federated learning on non-iid data," in *AAAI Conf. Artif. Intell.*, vol. 35, no. 9, 2021, pp. 7865–7873.

[22] P. Kairouz, H. B. McMahan, ..., H. Yu, and S. Zhao, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. abs/1912.04977, 2021.

[23] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning: A meta-learning approach," *NIPS*, vol. abs/2002.07948, 2020.

[24] C. Perez, "Invisible women: Exposing data bias in a world designed for men," 2020.

[25] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. E. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernández, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, and S. Staab, "Bias in data-driven artificial intelligence systems—an introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, 2020.

[26] B. MacNamee, P. Cunningham, S. Byrne, and O. I. Corrigan, "The problem of bias in training data in regression problems in medical decision support," *Artificial intelligence in medicine*, vol. 24 1, pp. 51–70, 2002. [Online]. Available: https://api.semanticscholar.org/CorpusID:3176364

[27] L. Dixon, J. Li, J. S. Sorensen, N. Thain, and L. Vasserman, "Measuring and mitigating unintended bias in text classification," *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:54997157

[28] M. Kolla and A. Savadamuthu, "The impact of racial distribution in training data on face recognition bias: A closer look," *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 313–322, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:254044307

[29] J. Gardner, C. A. Brooks, and R. Baker, "Evaluating the fairness of predictive student models through slicing analysis," *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:67870704

[30] C. Kung and R. Yu, "Interpretable models do not compromise accuracy or fairness in predicting college success," *Proceedings of the Seventh ACM Conference on Learning @ Scale*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220885559

[31] J. Barrett, A. Day, and Y. Gal, "Improving model fairness with time-augmented bayesian knowledge tracing," in *International Conference on Learning Analytics and Knowledge*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268321508

[32] H. Anderson, A. Boodhwani, and R. Baker, "Assessing the fairness of graduation predictions," in *Educational Data Mining*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:163158880

[33] C. G. Brinton, S. Buccapatnam, M. Chiang, and H. Poor, "Mining MOOC clickstreams: Video-watching behavior vs. in-video quiz performance," *IEEE Trans. Signal Process.*, vol. 64, pp. 3677–3692, 2016.

[34] Y.-W. Chu, E. Tenorio, L. Cruz, K. A. Douglas, A. S. Lan, and C. G. Brinton, "Click-based student performance prediction: A clustering guxided meta-learning approach," *IEEE Int. Conf. Big Data*, pp. 1389–1398, 2021.

[35] C. Neill, S. Cotner, M. Driessen, and C. J. Ballen, "Structured learning environments are required to promote equitable participation," *Chemistry Education Research and Practice*, 2019.

[36] S. M. Aguillon, G.-F. Siegmund, R. H. Petipas, A. G. Drake, S. Cotner, and C. J. Ballen, "Gender differences in student participation in an active-learning classroom," *CBE Life Sci. Edu.*, vol. 19, 2020.

[37] J. McBroom, I. Koprinska, and K. Yacef, "How does student behaviour change approaching dropout? a study of gender and school year differences," in *EDM*, 2020.

[38] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," in *Adv. Neural Inf. Process. Syst.*, Dec. 2015, pp. 505–513.

[39] A. F. Botelho, R. S. Baker, and N. T. Heffernan, "Improving sensor-free affect detection using deep learning," in *Int. Conf. Artif. Intell. Edu.*, 2017, pp. 40–51.

[40] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *NIPS*, vol. abs/1706.03762, 2017.

[41] M. Tatsuoka, F. Lord, M. R. Novick, and A. Birnbaum, "Statistical theories of mental test scores." *Journal of the American Statistical Association*, vol. 66, p. 651, 1968.

[42] W. Chen, C. G. Brinton, D. Cao, A. Mason-Singh, C. Lu, and M. Chiang, "Early detection prediction of learning outcomes in online short-courses via learning behaviors," *IEEE Transactions on Learning Technologies*, vol. 12, pp. 44–58, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:56785882

[43] F. Okubo, T. Shiino, T. Minematsu, Y. Taniguchi, and A. Shimada, "Adaptive learning support system based on automatic recommendation of personalized review materials," *IEEE Transactions on Learning Technologies*, vol. 16, pp. 92–105, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256802500

[44] C. G. Brinton, R. Rill, S. Ha, M. Chiang, R. Smith, and W. Ju, "Individualization for education at scale: Miic design and preliminary evaluation," *IEEE Transactions on Learning Technologies*, vol. 8, pp. 136–148, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:18653175

[45] I. Eegdeman, I. Cornelisz, C. van Klaveren, and M. Meeter, "Computer or teacher: Who predicts dropout best?" in *Frontiers in Education*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253841551

Yun-Wei Chu is currently working toward the PhD degree with the Department of Electrical and Computer Engineering, Purdue University, IN, USA. His current research interests include federated learning and meta-learning.

Seyyedali Hosseinalipour (S'17, M'20) received the Ph.D. degree in Electrical Engineering from North Carolina State University, NC, USA 2020. He was the recipient of the ECE Doctoral Scholar of the Year Award (2020) and ECE Distinguished Dissertation Award (2021) at North Carolina State University. He has served as the TPC Co-Chair of workshops related to federated learning and fog computing at several conferences such as IEEE INFOCOM, IEEE GLOBECOM, IEEE ICC, and IEEE MSN. He is currently an Assistant Professor of Electrical Engineering Department at University at Buffalo—SUNY. His research interests include the analysis of modern wireless networks and communication systems, distributed machine learning, and network optimization.

Elizabeth Tenorio is a data engineer with expertise in building data-driven applications across diverse domains including education, health, cybersecurity, and robotics. In her past roles as software engineer and data scientist, she developed apps for evidence-based cancer care at Vermonster, a content topic analysis platform at Zoomi AI, a drug discovery database at PMS-ICBG, and cyber behavior analytics tools at Forcepoint X-Labs. She serves as a data consultant for Purdue University and a senior data engineer at iRobot. Dr. Tenorio received her PhD in Biology from Kobe University and completed research fellowships in microbiology and bioinformatics at UCLA and Tufts Medical Center.

Laura Melissa Cruz Castro is an instructional assistant professor at the University of Florida. Her research interests focus on computational thinking at scale and data science education. She looks at K-12, higher education, professional, and community settings for both research interests. She holds a bachelor's degree in statistics from Universidad Nacional de Colombia, an M.S. in computer engineering, and Ph.D. in engineering education from Purdue University.

Kerrie A. Douglas (Member, IEEE) received the Ph.D. degree in educational studies with concentration on measurement and evaluation from Purdue University, West Lafayette, IN, USA, in 2012. She is an Associate Professor of Engineering Education with Purdue University and Deputy-Director of a large US Department of Defense- funded workforce development consortium for microelectronics, Scalable Asymmetric Lifecycle Engagement (SCALE). Dr. Douglas is a 2021 US NSF CAREER award recipient for her research on increasing the fairness of assessments in engineering education.

Andrew (Shiting) Lan is an assistant professor in the Manning College of Information and Computer Sciences, University of Massachusetts Amherst. His research focuses on the development of artificial intelligence (AI) and natural language processing (NLP) methods to enable scalable and effective personalized learning in education. He has worked on many research topics in this domain, including learner modeling, personalization policy learning, automated question and feedback generation, and educational process data analysis. His work resulted in several top prizes at public educational data mining challenges, including ones organized by the US Department of Education.

Christopher G. Brinton (S'08, M'16, SM'20) is the Elmore Rising Star Assistant Professor of Electrical and Computer Engineering at Purdue University. His research interest is at the intersection of networked systems and machine learning, specifically in distributed machine learning, fog/edge network intelligence, and data-driven network optimization. Dr. Brinton is a recipient of the NSF CAREER Award, ONR Young Investigator Program (YIP) Award, DARPA Young Faculty Award (YFA), AFOSR YIP Award, and Intel Rising Star Faculty Award. He currently serves as an Associate Editor for IEEE/ACM Transactions on Networking. Prior to joining Purdue, Dr. Brinton was the Associate Director of the EDGE Lab and a Lecturer of Electrical Engineering at Princeton University. Dr. Brinton received the PhD (with honors) and MS Degrees from Princeton in 2016 and 2013, respectively, both in Electrical Engineering.