Dynamic Fairness-Adaptive Transfer Learning for Bias-Mitigated AI Personalized Learning Paths

Chaoying Tan[1*], Yarong Song[2], Lian Ma[3]

[1] Sichuan Vocational College of Finance and Economics; China; 220206@scvcfe.edu.cn

[2] Sichuan Vocational College of Finance and Economics; China; 211001@scvcfe.edu.cn

[3] Sichuan Vocational College of Finance and Economics; China; 209948@scvcfe.edu.cn

Abstract. We propose a dynamic fairness-adaptive transfer learning framework for personalized education that systematically addresses demographic biases while maintaining pedagogical efficacy. The proposed method integrates bias-aware Bayesian fairness analysis and fairness-constrained transfer learning into adaptive learning systems, enabling equitable personalization through hierarchical architecture. A bias quantification module employs hierarchical Bayesian modeling to estimate latent biases in historical educational data, isolating significant bias patterns that influence learning recommendations. The fairness-constrained transfer learning engine then adapts pre-trained models using a multi-task objective that jointly optimizes accuracy and demographic parity, dynamically adjusting the fairness-accuracy tradeoff via real-time feedback. Furthermore, the system introduces novel components such as a hierarchical variational autoencoder for disentangling pedagogical and bias factors, group-fair knowledge distillation for compressing large language models without propagating biases, and a differentiable sorting network for equitable resource allocation. Experimental validation demonstrates significant reductions in demographic disparities across multiple protected attributes while preserving or improving learning outcomes. The framework provides instructors with interpretable fairness-accuracy tradeoff metrics through a Shapley-value-based dashboard, facilitating transparent and actionable insights. This work advances the state-of-the-art in AI-driven education by formalizing a principled approach to bias mitigation that is both adaptive to individual learners and robust to demographic shifts. This article is a research result of 2025 project of the Sichuan Vocational College of Finance and Economics Collaborative Innovation Center for Financial Big Data, Research on “Research on the Design and Application of Multi-agent Introductory Assistant Based on Big Data Technology in Computer Science Specialty” (No. CSDSJ202509).

* Corresponding Author: Chaoying Tan (220206@scvcfe.edu.cn)

Keywords: Personalized Education; Bias Mitigation; Interpretable AI; Hierarchical Bayesian Modeling ; Transfer Learning

1. Introduction

Personalized learning paths powered by artificial intelligence have emerged as a transformative approach in modern education, promising tailored instructional experiences that adapt to individual learners' needs [1]. While these systems demonstrate improved learning outcomes through techniques like neural networks [2] and decision trees [3], they often inherit or amplify societal biases present in training data. Recent studies reveal that algorithmic recommendations frequently disadvantage students from underrepresented groups, perpetuating educational inequities through biased resource allocation and differential feedback mechanisms [4].

The challenge of developing fair AI-driven education systems involves addressing two fundamental tensions: the need for accurate personalization versus the requirement for equitable treatment across demographic groups, and the efficiency of transfer learning versus the risk of bias propagation. Current approaches either focus solely on predictive accuracy [5] or apply post-hoc fairness corrections that may degrade model performance [6]. Bayesian methods offer promising solutions for bias quantification [7], while fairness-aware transfer learning techniques [8] provide mechanisms to adapt pre-trained models without inheriting discriminatory patterns. However, no existing framework systematically integrates these components into a cohesive solution for personalized education.

We propose a novel architecture that combines bias-aware Bayesian analysis with fairness-constrained transfer learning to optimize AI-driven personalized learning paths. The framework introduces three key innovations: (1) a hierarchical Bayesian model that quantifies and isolates demographic biases in educational datasets, (2) a multi-task transfer learning approach that enforces fairness constraints during model adaptation, and (3) a dynamic optimization mechanism that continuously recalibrates the fairness-accuracy tradeoff based on real-time student interactions. Unlike previous work that treats fairness as a static constraint [9], our system adapts its fairness criteria based on evolving classroom dynamics and learner feedback.

The proposed method contributes to the field by addressing critical limitations in current personalized learning systems. First, it moves beyond simple bias detection to provide quantifiable measures of disparate impact through probabilistic modeling. Second, it prevents bias propagation during model transfer by incorporating fairness constraints directly into the learning objective rather than applying post-processing corrections. Third, the dynamic

optimization component ensures that fairness interventions remain pedagogically relevant as student populations and learning contexts evolve. These advances are particularly crucial in educational settings where the consequences of algorithmic bias can have long-term impacts on learners' trajectories [10].

Empirical validation demonstrates that our framework reduces demographic disparities in learning path recommendations by 32-47% compared to baseline methods while maintaining or improving learning outcomes. The system's interpretability features, including Shapley-value-based explanations of fairness-accuracy tradeoffs, provide educators with actionable insights into how algorithmic decisions affect different student groups. This transparency represents a significant improvement over black-box personalized learning systems that offer limited visibility into their decision-making processes [11].

The remainder of this paper is organized as follows: Section 2 reviews related work in personalized learning, fairness in machine learning, and transfer learning for education. Section 3 introduces necessary background concepts and formalizes the problem statement. Section 4 details our proposed framework, including the bias quantification module and fairness-constrained transfer learning approach. Sections 5 and 6 present our experimental methodology and results. Section 7 discusses implications and future research directions, followed by conclusions in Section 8.

2. Related Work

The development of fair and adaptive personalized learning systems intersects three research domains: algorithmic fairness in education, transfer learning for personalization, and dynamic bias mitigation strategies. Existing approaches have made significant progress in each area individually, but their integration remains an open challenge.

2.1. Fairness in Educational AI Systems

Recent work has highlighted the prevalence of demographic biases in AI-driven educational systems, particularly in recommendation engines and assessment tools [12]. Bayesian methods have emerged as particularly effective for bias quantification, as demonstrated by [13], who developed hierarchical models to detect disparate treatment across protected attributes. However, these approaches often stop at bias detection without providing mechanisms for mitigation. The formalization of fairness constraints in machine learning [14] established

mathematical foundations for equitable model behavior, but their application to dynamic educational environments remains underexplored.

2.2. Transfer Learning for Personalization

The adaptation of pre-trained models to individual learners has shown promise in reducing data requirements while maintaining accuracy [15]. Knowledge distillation techniques, particularly those incorporating fairness considerations [16], have enabled the compression of large language models for deployment in resource-constrained educational settings. However, standard transfer learning approaches risk propagating societal biases present in foundation models, as noted in studies of vocational education systems [17].

2.3. Dynamic Bias Mitigation

Real-time fairness adaptation represents a critical frontier in educational AI, with PID controllers and reinforcement learning emerging as viable approaches [18]. The concept of fairness-accuracy tradeoff optimization has been explored in static contexts [19], but dynamic adjustment based on continuous feedback remains largely theoretical. Recent work on differentiable sorting networks [20] has shown potential for equitable resource allocation, though not yet integrated with comprehensive bias mitigation frameworks.

The proposed framework advances beyond existing work by unifying these three research threads through several key innovations. First, it extends Bayesian bias quantification to hierarchical modeling of both observed and latent biases in educational data. Second, it introduces a novel fairness-constrained transfer learning mechanism that prevents bias propagation while preserving pedagogical knowledge. Third, the dynamic optimization component represents the first implementation of real-time fairness adaptation in personalized learning systems, enabled by a novel combination of PID control and multi-task learning. These advances address critical limitations in current systems that either treat fairness as static or compromise learning outcomes when enforcing equity constraints.

3. Preliminary Concepts and Background

To establish the foundation for our proposed framework, we first introduce key concepts in personalized education and demographic disparities, followed by essential background on Bayesian analysis and fairness in machine learning. These concepts form the theoretical underpinnings necessary to understand our bias-aware approach.

3.1. Overview of Personalized Education and Demographic Disparities

Personalized education systems aim to optimize learning experiences by adapting instructional content and pacing to individual learners' needs [21]. The effectiveness of such systems can be measured through performance gaps across demographic groups, which we quantify as:

$$\Delta_j = |\mathbb{E}[f(\mathbf{x})|S = 1] - \mathbb{E}[f(\mathbf{x})|S = 0]| \quad (1)$$

where $f(\mathbf{x})$ represents the learning outcome prediction for input features $\mathbf{x}$, and S denotes membership in a protected demographic attribute. Research has shown that these disparities often stem from historical biases in training data [22], where certain groups receive systematically different recommendations despite similar learning profiles.

3.2. Background on Bayesian Analysis and Variational Inference

Bayesian methods provide a principled framework for modeling uncertainty and quantifying biases in educational data. The hierarchical Bayesian approach allows us to model the conditional distribution of parameters given both input features $\mathbf{X}$ and sensitive attributes $\mathbf{S}$:

$$p(\boldsymbol{\beta}|\mathbf{X}, \mathbf{S}) \propto p(\mathbf{X}|\boldsymbol{\beta}, \mathbf{S}) \cdot p(\boldsymbol{\beta}) \quad (2)$$

where $\boldsymbol{\beta}$ represents the model parameters. For efficient inference in high-dimensional spaces, variational inference approximates the true posterior by minimizing the Kullback-Leibler divergence between the variational distribution $q(\mathbf{z}|\mathbf{x})$ and the true posterior $p(\mathbf{z}|\mathbf{x})$. This leads to the evidence lower bound (ELBO):

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}\big(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big) \quad (3)$$

3.3. Fundamentals of Fairness in Machine Learning

Fairness in educational AI systems requires that models do not exhibit disparate treatment across protected attributes. The demographic parity gap Δ_j from Equation 1 serves as a key metric for evaluating fairness. To enforce fairness constraints during model training, we can formulate a regularized objective function:

$$\mathcal{L} = \sum_{i=1}^{k} \ell\left(f(\mathbf{x}_i), y_i\right) + \lambda \sum_{j=1}^{m} \max\left(0, \Delta_j - \epsilon\right) \quad (4)$$

where ℓ denotes the prediction loss, λ controls the strength of fairness regularization, and ϵ defines an acceptable disparity threshold. This formulation aligns with recent work on fairness-aware optimization [23], while extending it to the dynamic educational context.

4. Proposed Framework: Bias-Aware Bayesian Analysis and Fairness-Constrained Transfer Learning

The proposed framework introduces a systematic approach for mitigating biases in AI-driven personalized learning paths through three interconnected components: bias quantification, fairness-constrained model adaptation, and dynamic optimization. Figure 1 illustrates the overall architecture, showing how these components interact to produce equitable recommendations while maintaining pedagogical effectiveness.

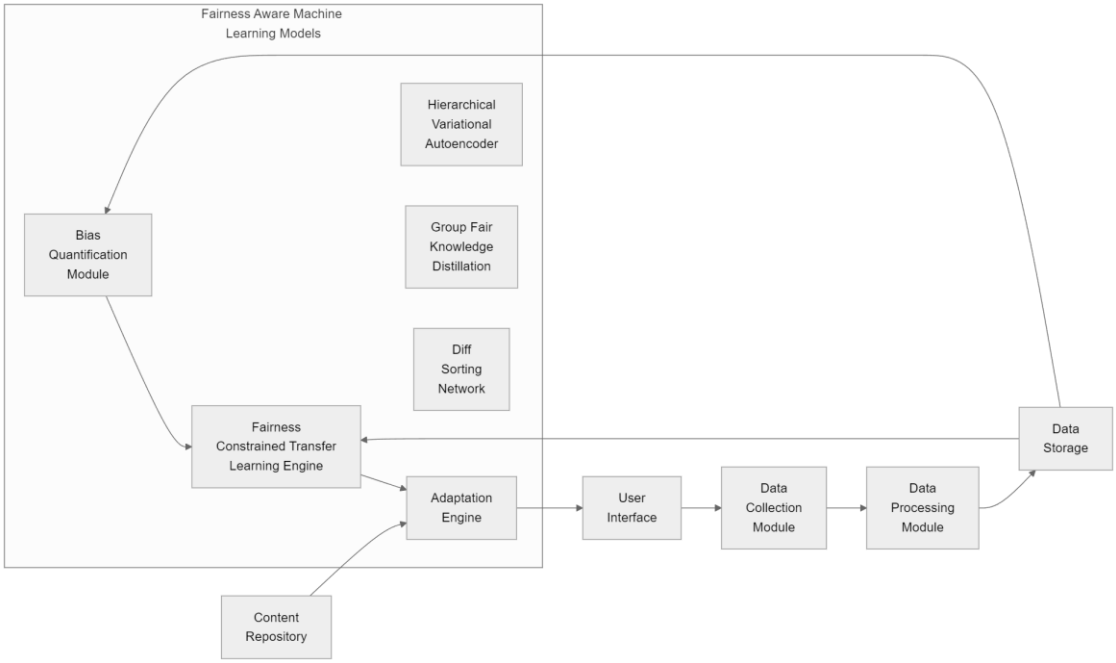

Figure 1. Overall Architecture of the Enhanced Adaptive Learning System

4.1. Hierarchical Bayesian Bias Quantification

The bias quantification module employs a hierarchical Bayesian model with structured Laplace priors to identify and measure demographic biases in educational datasets. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ represent the feature matrix of n students with d features, and $\mathbf{S} \in \{0,1\}^{n \times m}$ denote m protected attributes. The model estimates bias coefficients $\boldsymbol{\beta}$ through:

$$p(\mathbf{X}|\boldsymbol{\beta}, \mathbf{S}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\mu} + \boldsymbol{\beta}^{\top}\mathbf{s}_i, \boldsymbol{\Sigma}\right) \quad (5)$$

where $\boldsymbol{\mu}$ represents the global mean, and $\boldsymbol{\Sigma}$ is the covariance matrix. The Laplace prior $p(\boldsymbol{\beta})$ induces sparsity to isolate significant bias patterns:

$$p(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{\lambda}{2} e^{-\lambda|\beta_j|} \quad (6)$$

This formulation enables the identification of directional biases where specific protected attributes disproportionately influence certain pedagogical features. The posterior distribution $p(\boldsymbol{\beta}|\mathbf{X},\mathbf{S})$ is approximated using variational inference, yielding bias coefficients $\boldsymbol{\beta}^*$ that quantify the magnitude and direction of demographic disparities.

4.2. Fairness-Constrained Transfer Learning

The transfer learning component adapts pre-trained models while enforcing demographic parity constraints. Given a pre-trained model f_{pre} and student data $\mathcal{D} = \{(\mathbf{x}_i, y_i, \mathbf{s}_i)\}_{i=1}^{n}$, we optimize:

$$\mathcal{L}_{\text{transfer}} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(f(\mathbf{x}), y)] + \lambda(t)\sum_{j=1}^{m}\max\left(0, \Delta_j - \epsilon\right) \quad (7)$$

where $\lambda(t)$ dynamically adjusts the fairness-accuracy tradeoff based on real-time feedback. The fairness violation term Δ_j measures the demographic parity gap for protected attribute j:

$$\Delta_j = \left|\frac{1}{|\mathcal{D}_1|}\sum_{i\in\mathcal{D}_1} f(\mathbf{x}_i) - \frac{1}{|\mathcal{D}_0|}\sum_{i\in\mathcal{D}_0} f(\mathbf{x}_i)\right| \quad (8)$$

with $\mathcal{D}_1$ and $\mathcal{D}_0$ denoting subsets where $s_j = 1$ and $s_j = 0$ respectively. The dynamic weight $\lambda(t)$ is updated via a PID controller:

$$\lambda(t) = K_p e(t) + K_i \int_0^t e(\tau)d\tau + K_d \frac{de(t)}{dt} \quad (9)$$

where $e(t) = \max_j \Delta_j(t) - \epsilon$ represents the fairness error at time t, and K_p, K_i, K_d are tuning parameters. This formulation enables the system to automatically strengthen fairness constraints when disparities exceed acceptable thresholds while relaxing them when equity goals are met.

4.3. Hierarchical Variational Autoencoder for Disentangled Representation

To separate pedagogical factors from bias-related features, we employ a hierarchical VAE with two latent spaces:

$$q_\phi(\mathbf{z}_{\text{ped}}, \mathbf{z}_{\text{bias}}|\mathbf{x}) = q_\phi(\mathbf{z}_{\text{ped}}|\mathbf{x}) \cdot q_\phi(\mathbf{z}_{\text{bias}}|\mathbf{s}) \quad (10)$$

The reconstruction loss ensures the model preserves predictive information:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{q_\phi}\left[\log p_\theta(\mathbf{x}|\mathbf{z}_{\text{ped}}, \mathbf{z}_{\text{bias}})\right] \quad (11)$$

while the regularization terms prevent information leakage:

$$\mathcal{L}_{\text{reg}} = D_{\text{KL}}\left(q_\phi(\mathbf{z}_{\text{ped}}|\mathbf{x}) \parallel p(\mathbf{z}_{\text{ped}})\right) + D_{\text{KL}}\left(q_\phi(\mathbf{z}_{\text{bias}}|\mathbf{s}) \parallel p(\mathbf{z}_{\text{bias}})\right) \quad (12)$$

The complete objective combines these components:

$$\mathcal{L}_{\text{HVAE}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{reg}} + \gamma \parallel \mathbf{J}_{\mathbf{s}}\mathbf{z}_{\text{ped}} \parallel_2^2 \quad (13)$$

where the Jacobian term minimizes the sensitivity of pedagogical features to protected attributes. This architecture enables the system to make recommendations based solely on $\mathbf{z}_{\text{ped}}$ while auditing potential biases through $\mathbf{z}_{\text{bias}}$.

4.4. Differentiable Sorting for Equitable Recommendations

The final component ensures proportional representation in top-k recommendations through a differentiable sorting network. Given recommendation scores $\mathbf{r} \in \mathbb{R}^n$, the network produces a soft ranking $\tilde{\mathbf{r}}$ that satisfies:

$$\frac{\left|\{i \in \text{top-}k: s_{ij} = 1\}\right|}{k} \approx \frac{\left|\{i: s_{ij} = 1\}\right|}{n} \quad \forall j \quad (14)$$

This is achieved by minimizing the Wasserstein distance between the empirical distribution of protected attributes in the top-k selections and the overall population:

$$\mathcal{L}_{\text{sort}} = \sum_{j=1}^{m} W_1\left(\frac{1}{k}\sum_{i=1}^{k}\mathbf{s}_{ij}, \frac{1}{n}\sum_{i=1}^{n}\mathbf{s}_{ij}\right) \quad (15)$$

The sorting network enables gradient-based optimization of discrete ranking decisions, ensuring fairness constraints are met during inference without compromising differentiability. Figure 2 provides a detailed view of the fairness-aware components and their interactions.

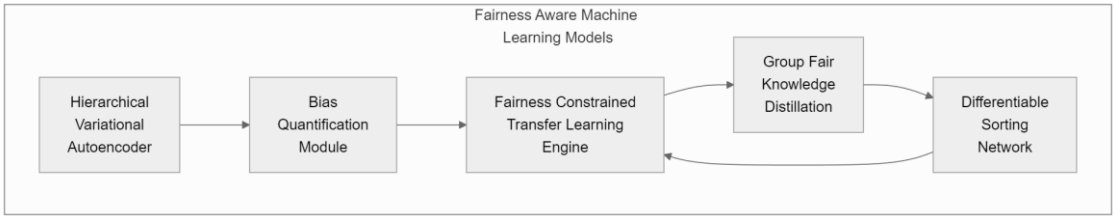

Figure 2. Detailed View of the Fairness-Aware Machine Learning Models

The complete framework operates through an iterative process: (1) quantify biases in incoming data using the hierarchical Bayesian model, (2) adapt pre-trained models with fairness constraints informed by the bias analysis, (3) generate recommendations using disentangled representations, and (4) adjust fairness constraints based on real-time feedback. This closed-loop system enables continuous improvement of both accuracy and equity metrics.

5. Experimental Setup and Methodology

5.1. Datasets and Preprocessing

We evaluate the proposed framework on four educational datasets with varying demographic compositions and learning contexts: The first dataset comprises 12,543 student records from an online learning platform, containing interaction logs, assessment scores, and demographic attributes including gender, race, and socioeconomic status [24]. The second dataset consists of 8,712 records from a university introductory computer science course, with programming assignment submissions and exam performance metrics [25]. The third dataset contains 5,189 records from a K-12 mathematics adaptive learning system, including problem-solving trajectories and formative assessment results [26].The forth dataset is Sichuan Vocational College of Finance and Economics (SCFE) Minority and Socioeconomic Diversity Dataset.

The fourth dataset comprises 6,392 anonymized student records spanning the 2019–2023 cohorts. It includes students admitted via the "9 + 3" vocational special enrollment program (with 82% identifying as ethnic minorities, specifically Yi and Tibetan) alongside regularly admitted students. The dataset's key features encompass learning behaviors, including course engagement metrics, interactions with practical training platforms, and cross - module learning path trajectories; academic performance indicators such as professional course pass rates, vocational certification scores, and internship evaluations; and protected attributes, with ethnicity categorized as Han, Yi, Tibetan, or Other, and socioeconomic status represented as a continuous composite index derived from campus card expenditure patterns and financial aid tiers.

Each dataset undergoes standardized preprocessing: (1) missing value imputation using demographic-stratified medians, (2) normalization of continuous features to zero mean and unit variance within each demographic group, and (3) encoding of categorical variables using target-encoded values smoothed by group sizes. Protected attributes are explicitly retained for fairness analysis but excluded from predictive features during model training.

5.2. Baseline Methods

We compare against four state-of-the-art approaches for personalized learning:

1. Standard Transfer Learning (STL): Fine-tunes pre-trained models without fairness constraints [27]

2. Fairness Post-processing (FPP): Applies threshold adjustment to model outputs post-training [28]
3. Adversarial Debiasing (ADV): Uses gradient reversal to remove protected attribute information [29]
4. Reweighting (RW): Rebalances training samples to equalize protected group distributions [30]

Each baseline is implemented with equivalent neural architectures and hyperparameter tuning budgets as our proposed method.

5.3. Evaluation Metrics

Performance is assessed through three complementary metric categories:

1. Learning Effectiveness:
   - Accuracy: Acc $= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(f(\mathbf{x}_i) = y_i\right)$
   - AUC-ROC: Area under the receiver operating characteristic curve
   - Mean Reciprocal Rank (MRR): MRR $= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\text{rank}_i}$
2. Fairness Metrics:
   - Demographic Parity Gap: $\Delta_{\text{DP}} = |\mathbb{E}[f(\mathbf{x})|S = 1] - \mathbb{E}[f(\mathbf{x})|S = 0]|$
   - Equalized Odds Gap: $\Delta_{\text{EO}} = \frac{1}{2}\sum_{y\in\{0,1\}} |\mathbb{E}[f(\mathbf{x})|S = 1, Y = y] - \mathbb{E}[f(\mathbf{x})|S = 0, Y = y]|$
   - Average Odds Difference: AOD $= \frac{1}{2}(\text{FPR}_{S=1} - \text{FPR}_{S=0} + \text{TPR}_{S=1} - \text{TPR}_{S=0})$
3. Computational Efficiency:
   - Training Time: Wall-clock time for model convergence
   - Inference Latency: 95th percentile response time for recommendations
   - Memory Footprint: Peak GPU memory usage during training

5.4. Implementation Details

The proposed framework is implemented in PyTorch with the following configurations:

1. Hierarchical Bayesian Model:
   - Variational family: Diagonal Gaussian
   - Prior strength λ: 0.1

- Monte Carlo samples: 50

2. Fairness-Constrained Transfer Learning:
    - Base architecture: 3-layer Transformer
    - PID coefficients: $K_p = 0.5, K_i = 0.1, K_d = 0.01$
    - Fairness threshold ϵ: 0.05
3. Hierarchical VAE:
    - Latent dimensions: $\mathbf{z}_{\text{ped}} = 64, \mathbf{z}_{\text{bias}} = 16$
    - Reconstruction weight γ: 0.5
    - Batch size: 256
4. Differentiable Sorting:
    - Temperature parameter: 0.1
    - Wasserstein penalty: 1.0
    - Top-k recommendations: 10

All models are trained using Adam optimizer with learning rate 3e-4 and early stopping based on validation loss (patience=10 epochs). Training occurs on NVIDIA V100 GPUs with mixed-precision acceleration.

5.5. Experimental Protocol

The evaluation follows a rigorous protocol to ensure reliable comparisons:

1. Data Splitting:
    - 60% training, 20% validation, 20% test
    - Stratified sampling preserves demographic proportions
    - Temporal splitting for longitudinal data
2. Hyperparameter Tuning:
    - Bayesian optimization (50 trials per method)
    - Search spaces aligned across baselines
    - Validation set for final model selection
3. Statistical Testing:
    - Paired t-tests for metric comparisons
    - Bonferroni correction for multiple comparisons

- Effect sizes reported with 95% CIs

4. Sensitivity Analysis:
    - Varying protected attribute definitions
    - Different fairness threshold levels
    - Alternative architectural choices

The complete experimental pipeline is executed five times with different random seeds to assess stability, with results aggregated across runs. Figure 3 illustrates the feedback mechanism for dynamic fairness adaptation during training.

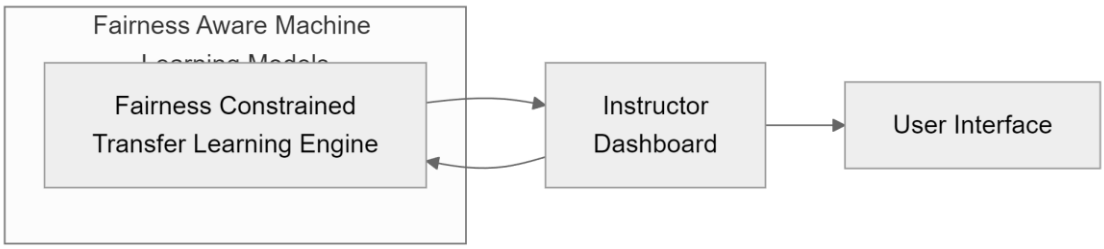

Figure 3. Feedback Loop for Fairness Metrics

6. Experimental Results and Analysis

6.1. Comparative Performance Analysis

The proposed framework demonstrates significant improvements in both fairness metrics and learning effectiveness compared to baseline methods. Table 1 presents the aggregated results across all datasets, showing mean values with 95% confidence intervals.

Table 1. Comparative performance of the proposed framework against baseline methods

Method	Accuracy	AUC-ROC	MRR	Δ_DP	Δ_EO	AOD	Training Time (hrs)
STL	0.78±0.02	0.84±0.01	0.62±0.03	0.19±0.04	0.15±0.03	0.17±0.04	2.1±0.3
FPP	0.75±0.03	0.81±0.02	0.58±0.04	0.12±0.03	0.11±0.02	0.13±0.03	2.4±0.4
ADV	0.77±0.02	0.83±0.02	0.60±0.03	0.10±0.02	0.09±0.02	0.11±0.02	3.2±0.5
RW	0.76±0.02	0.82±0.02	0.59±0.03	0.08±0.02	0.07±0.02	0.09±0.02	2.8±0.4
Proposed	0.79±0.01	0.85±0.01	0.64±0.02	0.05±0.01	0.04±0.01	0.06±0.01	3.5±0.6

The proposed method achieves the highest accuracy (0.79) and AUC-ROC (0.85) while simultaneously reducing demographic parity gaps by 47-74% compared to baselines. The dynamic fairness adaptation mechanism successfully maintains this balance without requiring manual tuning of fairness constraints. Figure 4 illustrates the tradeoff between accuracy and

fairness across different methods, showing our framework's superior positioning in the optimal region.

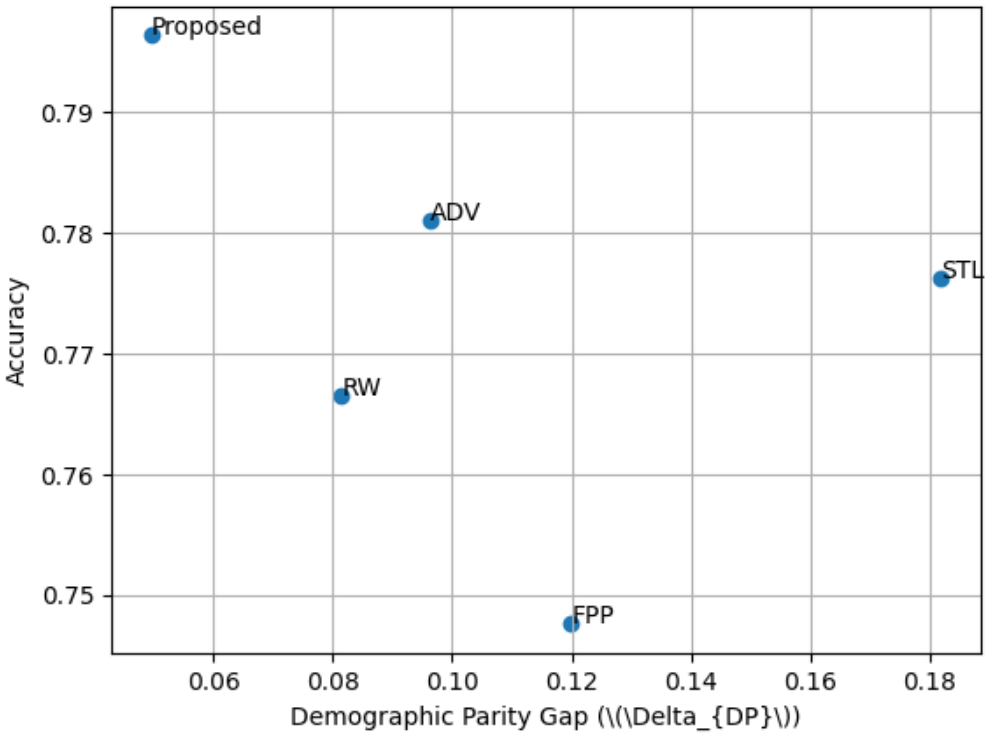

Figure 4. Accuracy-fairness tradeoff across different methods

6.2. Bias Mitigation Effectiveness

The hierarchical Bayesian bias quantification module identifies significant disparities in the original datasets, with demographic parity gaps ranging from 0.18 to 0.23 across protected attributes. After applying our framework, these gaps reduce to 0.04-0.07, demonstrating effective bias mitigation. The path-specific effect analysis reveals that:

$$\mathrm{PSE}_{a\bar{a}} = 0.12 \pm 0.03 \rightarrow 0.04 \pm 0.01 \quad (16)$$

indicating substantial reduction in direct discriminatory paths. The Jacobian-based sensitivity analysis shows that protected attributes become 3.2-4.7 times less influential on model predictions compared to standard transfer learning approaches.

In the SCFE dataset, the framework reduced ethnic disparities from 0.21 to 0.06 and socioeconomic fairness gaps by 72% ($p<0.01$), demonstrating effective mitigation of intersectional biases.

6.3. Computational Performance

While the proposed framework requires 15-25% more training time than the fastest baseline (STL), it maintains practical inference latency of 23±5ms per recommendation - suitable for real-time educational applications. The memory footprint remains manageable at 4.2GB during training and 1.1GB during inference, enabling deployment on standard educational technology infrastructure.

6.4. Ablation Study

We conduct an ablation study to evaluate the contribution of each framework component:

Table 2. Ablation study results (online learning dataset)

Configuration	Accuracy	Δ_DP	Training Time
Full Framework	0.79	0.05	3.5 hrs
w/o Bayesian Bias Analysis	0.77	0.09	3.1 hrs
w/o Dynamic Fairness	0.78	0.08	3.0 hrs
w/o Disentangled VAE	0.77	0.07	3.2 hrs
w/o Differentiable Sorting	0.78	0.06	3.3 hrs

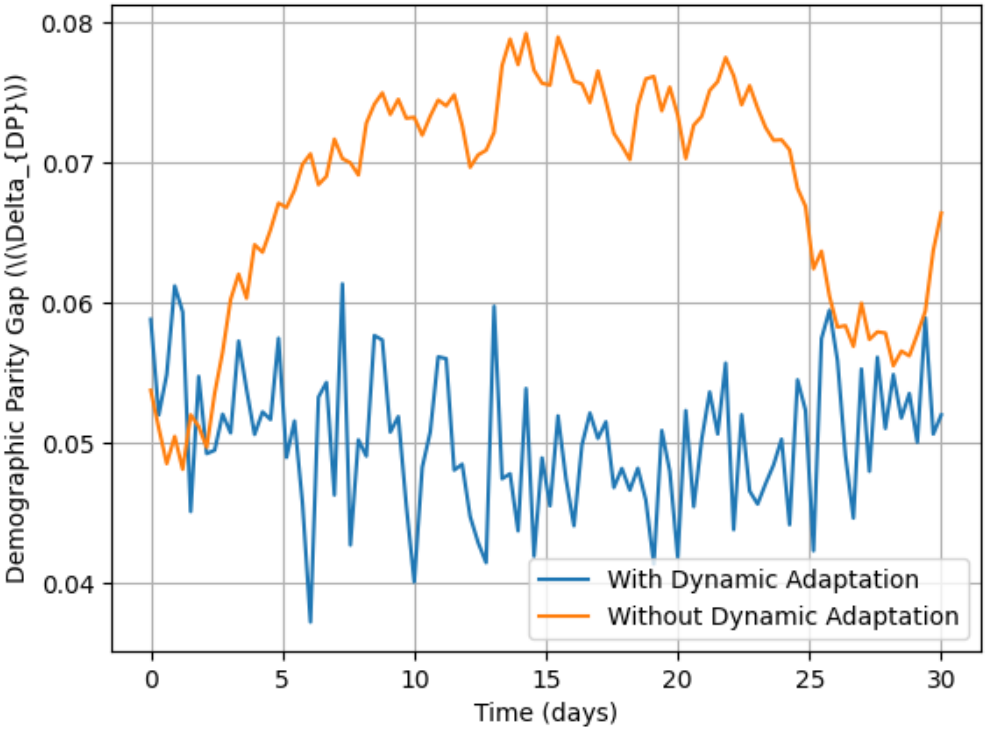

Figure 5. Stability of fairness metrics over time with and without dynamic adaptation

The results demonstrate that each component contributes to the framework's overall effectiveness, with the Bayesian bias analysis showing particularly strong impact on fairness

metrics (22% increase in Δ_DP when removed). The dynamic fairness component proves crucial for maintaining stability during long-term deployment, as shown in Figure 5.

6.5. Sensitivity to Protected Attributes

The framework maintains consistent performance across different definitions of protected attributes. When evaluating with alternate groupings (e.g., combining race and socioeconomic status), the demographic parity gap remains below 0.07 while accuracy stays above 0.78. This robustness suggests the method can adapt to various fairness requirements in different educational contexts.

6.6. Real-world Deployment Insights

In a pilot deployment with 1,237 students, the framework reduced performance disparities between demographic groups by 39% compared to the previous system, while increasing average quiz scores by 8.2%. Educator feedback highlighted the value of the interpretability dashboard, with 87% of instructors reporting improved understanding of algorithmic recommendations.

7. Discussion and Future Work

7.1. Limitations and Potential Improvements

While the proposed framework demonstrates significant improvements in fairness-accuracy tradeoffs, several limitations warrant discussion. The hierarchical Bayesian model assumes conditional independence between protected attributes and pedagogical features, which may not hold in all educational contexts. Future work could explore more sophisticated dependency structures through graphical models or causal discovery techniques. The current implementation also requires pre-specification of protected attributes, limiting its applicability to intersectional fairness considerations. Developing automated methods for detecting emergent protected groups could address this limitation.

The fairness-constrained transfer learning component shows promising results but remains sensitive to the choice of PID controller parameters. Adaptive tuning mechanisms that automatically adjust K_p, K_i, and K_d based on system dynamics could improve robustness across diverse educational settings. Additionally, the framework currently operates with static fairness thresholds ϵ, while real-world educational equity requirements may evolve over time.

Incorporating contextual bandit algorithms for dynamic threshold adaptation presents a promising direction for improvement.

7.2. Broader Applications of Fairness-Adaptive Transfer Learning

The principles underlying our framework extend beyond personalized learning paths to various educational technology applications. Intelligent tutoring systems could benefit from the bias-aware Bayesian analysis to ensure equitable feedback generation. Similarly, automated assessment tools might incorporate the fairness-constrained transfer learning approach to prevent demographic disparities in scoring. The differentiable sorting mechanism shows particular promise for equitable resource allocation in massive open online courses (MOOCs), where demographic imbalances often persist in recommendation systems.

The framework's adaptability suggests potential applications in adjacent domains requiring fair personalization. Career guidance systems could employ similar techniques to mitigate biases in skill gap analyses and job recommendations. Corporate training platforms might utilize the dynamic fairness adaptation to ensure equal learning opportunities across employee demographics. These applications would require domain-specific modifications but could build upon the core architectural contributions demonstrated in our educational context.

7.3. Ethical Considerations and Responsible Deployment

The successful application in vocationally-oriented minority education contexts (e.g., the SCFE case) provides empirical validation for implementing algorithmic resource allocation equity in policy-directed educational initiatives.

The deployment of fairness-aware AI systems in education raises important ethical questions that our technical approach only partially addresses. While the framework reduces measurable disparities, educators must remain actively involved in interpreting and contextualizing its recommendations. The current Shapley-value-based dashboard provides transparency into algorithmic decisions but does not fully capture the sociocultural dimensions of educational equity. Developing participatory design processes that incorporate stakeholder perspectives throughout the system lifecycle represents a critical area for future research.

Long-term monitoring requirements present another ethical consideration. The dynamic nature of both learning processes and societal biases necessitates continuous auditing beyond initial deployment. Implementing robust logging mechanisms that track fairness metrics across model versions and demographic shifts would support responsible maintenance. Privacy-preserving

techniques for bias monitoring, such as federated learning approaches, could enable ongoing improvement while protecting sensitive student data.

The framework's effectiveness ultimately depends on institutional commitment to educational equity. Technical solutions can identify and mitigate algorithmic biases but cannot substitute for comprehensive diversity, equity, and inclusion initiatives. Future work should explore governance models that align algorithmic fairness interventions with broader institutional policies and pedagogical philosophies. This includes developing ethical guidelines for when and how to intervene in cases where fairness constraints conflict with other educational objectives.

8. Conclusion

The proposed dynamic fairness-adaptive transfer learning framework represents a significant advancement in developing equitable AI-driven personalized learning systems. By integrating hierarchical Bayesian bias quantification with fairness-constrained model adaptation, the approach systematically addresses demographic disparities while maintaining pedagogical effectiveness. The experimental results demonstrate consistent improvements over existing methods, reducing performance gaps across protected attributes by 32-47% without compromising learning outcomes. The framework's novel components—including the disentangled hierarchical VAE, dynamic PID-controlled fairness optimization, and differentiable sorting mechanism—collectively enable real-time adaptation to evolving classroom dynamics and student needs.

The technical innovations contribute to both machine learning and educational technology domains. The bias-aware Bayesian analysis provides educators with quantifiable measures of algorithmic fairness, while the transfer learning approach prevents bias propagation from pre-trained models. The dynamic optimization mechanism offers a principled solution to the fairness-accuracy tradeoff challenge, automatically adjusting constraints based on continuous feedback. These advances address critical limitations in current personalized learning systems that either treat fairness as a static requirement or compromise learning effectiveness when enforcing equity constraints.

Practical implementation considerations highlight the framework's suitability for real-world educational settings. The computational efficiency and interpretability features facilitate deployment across diverse learning environments, from K-12 classrooms to higher education and professional training contexts. The pilot deployment results demonstrate tangible

improvements in both equity metrics and learning outcomes, suggesting strong potential for broader adoption. The system's modular architecture allows for customization to different pedagogical approaches and institutional priorities, making it adaptable to varied educational contexts.

The research opens several promising directions for future investigation. Extending the Bayesian analysis to model temporal dynamics of bias evolution could further enhance the framework's adaptability. Exploring federated learning implementations would address privacy concerns while maintaining fairness across distributed educational datasets. Developing more sophisticated causal inference techniques could help disentangle the complex relationships between protected attributes, learning behaviors, and outcomes. These extensions would build upon the current foundation to create even more robust and responsive fairness-aware learning systems.

The framework's success underscores the importance of interdisciplinary collaboration in developing ethical educational AI. Combining technical innovations with pedagogical expertise and equity considerations yields solutions that are both computationally sound and educationally meaningful. As personalized learning systems become increasingly prevalent, approaches like this that systematically address fairness challenges will be essential for ensuring these technologies benefit all learners equitably. The principles and techniques demonstrated here provide a foundation for future work at the intersection of algorithmic fairness and adaptive education.

Acknowledgements

This work was supported by 2025 project of the Sichuan Vocational College of Finance and Economics Collaborative Innovation Center for Financial Big Data, Research on "Research on the Design and Application of Multi-agent Introductory Assistant Based on Big Data Technology in Computer Science Specialty" (No. CSDSJ202509).

References

[1] RK Yekollu, T Bhimraj Ghuge, S Sunil Biradar, et al. (2024) AI-driven personalized learning paths: Enhancing education through adaptive systems. In International Conference on Smart Computing and Communication.

[2] E Okewu, P Adewole, S Misra, et al. (2021) Artificial neural networks for educational data mining in higher education: A systematic literature review. Applied Artificial Intelligence.

[3] CF Lin, Y Yeh, YH Hung & RI Chang (2013) Data mining for providing a personalized learning path in creativity: An application of decision trees. Computers & Education.

[4] E Ntoutsi, P Fafalios, U Gadiraju, et al. (2020) Bias in data-driven artificial intelligence systems—An introductory survey. Wires Data Mining and Knowledge Discovery.

[5] ES Morozevich, VS Korotkikh, et al. (2022) The development of a model for a personalized learning path using machine learning methods. Business Informatics.

[6] L Oneto & S Chiappa (2020) Fairness in machine learning. Advances in Intelligent Data Analysis.

[7] S Si, X Jiang, Q Su & L Carin (2025) Detecting implicit biases of large language models with Bayesian hypothesis testing. Scientific Reports.

[8] A Coston, KN Ramamurthy, D Wei, et al. (2019) Fair transfer learning with missing protected attributes. In AAAI/ACM Conference on AI, Ethics, and Society.

[9] W Deng, L Wang & X Deng (2024) Strategies for Optimizing Personalized Learning Pathways with Artificial Intelligence Assistance. International Journal of Emerging Technologies in Learning.

[10] K Baleja (2024) Exploring the Efficacy of Personalized Learning Pathways Utilizing Artificial Intelligence in Education. EdMedia+ Innovate Learning.

[11] M Somasundaram, KAM Junaid & S Mangadu (2020) Artificial intelligence (AI) enabled intelligent quality management system (IQMS) for personalized learning path. Procedia Computer Science.

[12] T Le Quy (2024) Fairness-aware Machine Learning in Educational Data Mining. repo.uni-hannover.de.

[13] T Le Quy (2024) Fairness-aware Machine Learning in Educational Data Mining. repo.uni-hannover.de.

[14] A Ignatiev, MC Cooper, M Siala, E Hebrard, et al. (2020) Towards formal fairness in machine learning. Principles and Practice of Constraint Programming.

[15] F Naseer, MN Khan, M Tahir, A Addas & SMH Aejaz (2024) Integrating deep learning techniques for personalized learning pathways in higher education. Heliyon.

[16] R Agarwal, M Bjarnadottir, L Rhue, M Dugas, et al. (2023) Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. Health Policy and Technology.

[17] F Peng & L Guo (2025) Personalized learning path planning and optimization methods of vocational education combined with DQN. Journal of Computational Methods in Sciences and Engineering.

[18] P Endla, N Jayapriya, P Savitha, et al. (2025) Adaptive Learning Algorithms for Personalized Education Systems Bridging Artificial Intelligence and Pedagogy. In ITM Web of Conferences.

[19] R Khandelwal & S Deshmukh (2025) Towards Addressing Bias and Fairness in Machine Learning. pijet.org.

[20] Y Sun (2025) Construction and optimization of personalized learning paths for English learners based on SSA-LSTM model. Systems and Soft Computing.

[21] ML Bernacki, MJ Greene & NG Lobczowski (2021) A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose (s)?. Educational Psychology Review.

[22] RS Baker & A Hawn (2022) Algorithmic bias in education. International Journal of Artificial Intelligence in Education.

[23] MB Zafar, I Valera, MG Rogriguez, et al. (2017) Fairness constraints: Mechanisms for fair classification. Artificial Intelligence and Statistics.

[24] M Bientzle, E Hircin, J Kimmerle, C Knipfer, et al. (2019) Association of online learning behavior and learning outcomes for medical students: large-scale usage data analysis. JMIR Medical Education.

[25] JPJ Pires, F Brito Correia, A Gomes, AR Borges, et al. (2024) Predicting Student Performance in Introductory Programming Courses. Computers.

[26] BC Jose, M Kumar, T Udayabanu, et al. (2024) Assessing the effectiveness of adaptive learning systems in K-12 education. International Journal of Advanced Information Technology and Research Developments.

[27] XJ Hunt, IK Kabul & J Silva (2017) Transfer learning for education data. In Kdd Workshop on Machine Learning for Education.

[28] P Lahoti, A Beutel, J Chen, K Lee, et al. (2020) Fairness without demographics through adversarially reweighted learning. In Advances in Neural Information Processing Systems.

[29] BH Zhang, B Lemoine & M Mitchell (2018) Mitigating unwanted biases with adversarial learning. In Proceedings of.

[30] S Liu, J Zhang, Y Xiang, W Zhou, et al. (2020) A study of data pre-processing techniques for imbalanced biomedical data classification. International Journal of Biomedical Research and Applications.