A Crowd–AI Collaborative Approach to Address Demographic Bias for Student Performance Prediction in Online Education

Ruohan Zong, Yang Zhang, Frank Stinar, Lanyu Shang, Huimin Zeng, Nigel Bosch, Dong Wang

School of Information Sciences, University of Illinois Urbana-Champaign
{rzong2, yzhangnd, fstinar2, lshang3, huiminz3, pnb, dwang24}@illinois.edu

Abstract

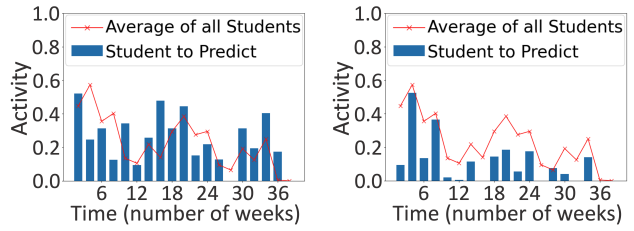
Recent advances in artificial intelligence (AI) and crowdsourcing have shown success in enhancing learning experiences and outcomes in online education. This paper studies a student performance prediction problem where the objective is to predict students' outcomes in online courses based on their behavioral data. In particular, we focus on addressing the limitation of current student performance prediction solutions that often make inaccurate predictions for students from underrepresented demographic groups due to the lack of training data and differences in behavioral patterns across groups. We develop *DebiasEdu*, a crowd–AI collaborative debias framework that melds the AI and crowd intelligence through 1) a novel gradient-based bias identification mechanism and 2) a bias-aware crowdsourcing interface and bias calibration design to achieve an accurate and fair student performance prediction. Evaluation results on two online courses demonstrate that DebiasEdu consistently outperforms state-of-the-art AI, fair AI, and crowd–AI baselines by achieving an optimized student performance prediction in terms of both accuracy and fairness.

Introduction

Emerging AI and crowdsourcing techniques have been utilized to enhance online education activities (e.g., assignment and exam assessment, interactions on online learning platforms, and personalized learning) (Chen et al. 2021; Wambsganss et al. 2022; Troussas, Krouska, and Sgouropoulou 2020). One of the critical problems in online education is student performance prediction (Albreiki, Zaki, and Alashwal 2021; Waheed et al. 2020; Qiu et al. 2022), which aims to predict a student's final performance result in a course (e.g., *Fail*, *Pass*, and *Distinction*) based on the behavioral data of students (Adnan et al. 2021; Qiu et al. 2022; Li et al. 2020). The prediction results can provide feedback to improve a student's metacognitive ability (Boud, Lawson, and Thompson 2015) and assist educational institutions in designing effective mechanisms to improve academic outcomes and avoid dropout (Rastrollo-Guerrero, Gómez-Pulido, and Durán-Domínguez 2020). Moreover, in the large-scale online education (e.g., Coursera and MOOC), automatic performance prediction is a critical feedback to enhance both learning and teaching since it is difficult for instructors to pay a close attention to each student's performance in such courses (Xu, Yuan, and Liu 2020). To accurately predict a student's performance, we focus on online activity data (e.g., reviewing course materials, completing quizzes, and engaging in collaborative activities) as it provides valuable insights into the level of engagement and effort of a student from various aspects (e.g., learning new contents, testing acquired knowledge, and discussing with peers) (Kuzilek, Hlosta, and Zdrahal 2017). In addition, the activity data is beneficial compared to other types of data (e.g., assignment or quiz scores) since it is easy to collect and measure at any time in online education, which enables early prediction and timely assistance for students (Adnan et al. 2021).

AI solutions that utilize machine learning models and deep neural networks have been developed to address the student performance prediction problem (Wasif et al. 2019; Hasan et al. 2020; Waheed et al. 2020; Qiu et al. 2022). However, these solutions primarily focus on achieving high prediction accuracy but pay less attention to demographic biases, where underrepresented students often receive less accurate prediction results due to the lack of training data for those students and differences in their behavioral patterns (Baker and Hawn 2022). For example, in a student performance prediction application, a deep sequential neural network-based model (Li et al. 2020) can achieve a 0.62 accuracy on the underrepresented age group (i.e., age greater than 35) that includes 24.4% students, while achieving a 0.75 accuracy on other students in the class. Figure 1 visualizes the online activities every two weeks of *underrepresented* students with incorrect prediction results from AI models. In particular, Figure 1(a) shows a student with an actual result of *Pass* but is predicted as a *Distinction* (even better than pass) by AI. Figure 1(b) shows a student with a *Fail* result but is predicted as a *Pass* by AI. If we provide such inaccurate prediction results to the underrepresented students, it could mislead them to have an inaccurate self-assessment and thus complete too few activities in the course (or waste time on too many activities), which could further exacerbate their potential disadvantages in online education (Kizilcec and Lee 2020).

Fair AI solutions have been developed to address the demographic bias problem (Zafar et al. 2019; Bendekgey and



(a) Truth: *Pass* (AI: *Distinction*) (b) Truth: *Fail* (AI: *Pass*)

Figure 1: Examples of Biased AI Prediction.

Sudderth 2021; Kini et al. 2021; Jiang and Pardos 2021). These solutions often address the demographic bias by increasing the weights of underrepresented samples during training (e.g., sample re-weighting) (Kini et al. 2021) or integrating fairness regularization into the training objective (e.g., fairness constraints and adversarial learning) (Zafar et al. 2019; Jiang and Pardos 2021). However, these solutions often achieve results with improved fairness at the cost of reduced overall accuracy due to the tradeoff between fairness and accuracy of data-driven models (Berk et al. 2017), which is also observed in the online education domain (Stinar and Bosch 2022). Motivated by the limitations of the above fair AI solutions, we propose to leverage human intelligence from crowdsourcing platforms to address the demographic bias of AI. We refer to such intelligence as *crowd intelligence*, which has been leveraged to improve AI models in different scenarios (e.g., image classification (Sener and Savarese 2018), facial analysis (Scheuerman et al. 2020), and disaster damage assessment (Zhang et al. 2019a)). In our work, crowd workers predict a student's final performance based on their high-level understandings of a student's activities (e.g., a student's general trend of activities compared to the average, consistent hard work, and extra hard work before the final), which is difficult for the AI models to learn given the lack of training data from the underrepresented groups in online education applications. For example, it should be a relatively easy task for humans to predict a *Pass* result for Figure 1(a) and a *Fail* result for Figure 1(b) based on their prior knowledge (including common sense), which does not require additional *training data* compared to the AI models. For an empirical effectiveness examination of using crowd intelligence in the student performance prediction problem, please refer to the *Crowdsourcing Settings and Pilot Study* in the *Evaluation* section. Motivated by the aforementioned cognitive power of crowd intelligence, we develop a *bias-aware crowd–AI collaborative* student performance prediction framework that jointly explores the collective strength of AI and crowd intelligence to effectively predict a student's performance while addressing the demographic bias in the prediction results. However, two critical technical challenges exist, which are elaborated as follows.

The first challenge lies in how to effectively *identify biased AI results from different demographic groups to achieve an optimized tradeoff between accuracy and fairness*. In particular, it is challenging to select a subset of samples where the AI models are likely to make inaccurate predictions due to the lack of training data and different behavioral patterns in underrepresented groups. One possible solution to overcome this challenge is to leverage current active learning methods (Abdar et al. 2021; Xie et al. 2022; Ren et al. 2021) to select data samples that are difficult for the AI models to predict based on uncertainty measurements or loss function designs. However, these active learning methods often focus on improving the prediction accuracy and tend to select fewer samples from underrepresented groups that contribute less to overall prediction accuracy. Hence, these methods can result in improved accuracy at the cost of reduced fairness. There have also been deep learning solutions that select biased AI results to boost the overall fairness by predicting the impact of *training* samples on the model fairness (Kou et al. 2021; Anahideh, Asudeh, and Thirumuruganathan 2022). Nevertheless, these solutions cannot accurately identify the biased results from the *testing* samples in the absence of the ground-truth labels.

The second challenge lies in how to effectively *address demographic bias using potentially biased crowd intelligence*. Recent efforts in crowd–AI collaboration (Sener and Savarese 2018; Yoo and Kweon 2019; Zhang et al. 2022) have been made to address this challenge. These approaches often utilize crowd intelligence to improve prediction accuracy and fairness by troubleshooting failure cases of AI models under the assumption that the crowd can provide accurate and fair responses. However, cognitive bias of crowd workers (Draws et al. 2021) may have a negative impact on their annotation performance (Hettiachchi et al. 2020). For example, crowd workers may have the confirmation bias of being conservative in predicting a *Distinction* result due to their preexisting beliefs that *Distinction* is assigned to a really small percentage of students. Another example of cognitive bias is the anchoring effect, where crowd workers can be overly influenced by the first few examples they see. Hence, the generated crowd feedback can possibly mislead the AI models to learn inaccurate information in their predictions.

To address the above challenges, we develop *DebiasEdu*, a crowd–AI collaborative debias framework that effectively integrates AI and crowd intelligence to achieve accurate and fair student performance prediction in online education. To address the first challenge, we design a gradient-based AI bias identification module that analyzes the gradient variation of training data to select biased AI results. To overcome the second challenge, we develop a novel bias-aware crowdsourcing interface and a crowd–AI fusion mechanism to address the demographic bias of AI and the cognitive bias of the crowd. To the best of our knowledge, the DebiasEdu is the first crowd–AI collaborative framework to address the algorithmic demographic bias in online education. We evaluate the DebiasEdu using a Social Science course and a STEM (Science, Technology, Engineering, and Math) course on a widely used online learning platform Open University (Kuzilek, Hlosta, and Zdrahal 2017). Evaluation results demonstrate that our DebiasEdu consistently outperforms state-of-the-art AI, fair AI, and crowd–AI baselines in terms of student performance prediction accuracy and fairness. Our main contributions are summarized as follows:

- We develop a crowd–AI collaborative debias framework,

DebiasEdu, to achieve an accurate and fair student performance prediction using their behavioral data, which can be used to provide feedback to students and enhance their metacognitive abilities.

- We develop a novel bias identification mechanism and a crowdsourcing interface design to address two important technical challenges: 1) identifying biased AI results to achieve an optimized tradeoff between accuracy and fairness and 2) addressing demographic bias using potentially biased crowd intelligence.
- We perform extensive experiments to evaluate the DebiasEdu on two real online courses from a widely used online learning platform. Evaluation results demonstrate significant performance gains of DebiasEdu compared to state-of-the-art baselines in both accuracy and fairness.

Related Work

AI and Crowdsourcing for Online Education. Several efforts have been made to improve learning experiences and outcomes in online education with the recent advances in AI and crowdsourcing. For example, Abdi, Khosravi, and Sadiq (2020) designs a crowdsourcing-based learning system to assess students' knowledge state by tracing their performance on crowdsourcing knowledge assessment tasks. Prihar et al. (2021) utilizes crowdsourced tutoring to increase students' next-problem accuracy in online learning and develops a method to rank the tutoring effectiveness of different crowd workers. Wambsganss et al. (2022) develops a deep-learning-based student argumentation self-evaluation system that leverages nudging theory techniques to help students write convincing texts. Qadir (2023) analyzes how to use large language models to benefit students (e.g., customized explanations) while minimizing negative impacts (e.g., misinformation). However, current AI and crowdsourcing approaches often ignore the algorithmic demographic bias in online education to ensure fairness.

Algorithmic Demographic Bias. There is a growing trend of analyzing and addressing algorithmic demographic bias in computational applications, such as face recognition, text classification, and educational decision-making. For example, Scheuerman et al. (2020) examines the limitations of race and gender annotation and proposes solutions to reduce the demographic bias in facial datasets. Madaio et al. (2022) conducts structured interviews with AI practitioners to identify the challenges and needs in addressing the demographic bias in text classification. Moreover, algorithmic demographic bias has been examined in education. For example, Yang et al. (2021) analyzes the potential misuse of AI due to algorithmic bias that can inhibit human rights and lead to demographic inequality in education (e.g., exaggerating demographic bias in AI decision-making). Baker and Hawn (2022) studies the algorithmic bias in education and proposes potential ways to mitigate the bias (e.g., improving fairness in data collection and incentivizing bias analysis in model evaluation). To the best of our knowledge, this paper is the first crowd–AI collaborative framework to address the algorithmic demographic bias in online education.

Crowd–AI Collaboration. Our paper is also closely related to current advances in crowd–AI collaborative approaches to integrate the strengths of AI and crowd intelligence in various applications (e.g., object detection, human pose estimation, image annotation, and neural architecture search). For instance, Sener and Savarese (2018) develops a deep active learning scheme that identifies a core set of representative samples and leverages the crowd on the selected samples to improve object detection accuracy. Yoo and Kweon (2019) builds a crowd–AI framework that leverages a task-agnostic loss design to efficiently integrate AI and the crowd in human pose estimation. Kobayashi, Wakabayashi, and Morishima (2021) proposes a crowd–AI scheme that utilizes a divide-and-conquer task assignment strategy to optimize the image annotation performance. Zhang et al. (2022) designs a crowd-guided framework for neural architecture search through estimation-theory-based crowd–AI integration. However, these crowd–AI collaborative approaches often ignore the demographic bias of AI and the cognitive bias of the crowd and cannot ensure the fairness of the system.

Problem Formulation

In this section, we formally present our bias-aware crowd–AI collaborative student performance prediction problem. We first introduce a few key concepts and notations to define the input and output of student performance prediction.

Let $\boldsymbol{D} = \{D_1, D_2, ..., D_M\}$ represent the *demographic attribute* of students (e.g., age, gender, race, or highest education), where the demographic attribute can be classified into M different categories. For example, the highest education attribute can be categorized into high school, undergraduate, master, and doctoral. Among demographic attribute categories, we define *underrepresented groups* $\boldsymbol{U}$ to be groups of traditionally underrepresented students (e.g., female students in STEM courses).

Definition 1 Demographic Data ($\boldsymbol{X^D}$): We define $\boldsymbol{X^D} = \{X_1^D, X_2^D, ..., X_I^D\}$ to be the aforementioned demographic attribute categories of all students in a course, where I is the total number of students. In particular, X_i^D for $i \in \{1, 2, ..., I\}$ represents the demographic attribute category of the i^{th} student, where $X_i^D \in \{D_1, D_2, ..., D_M\}$.

Demographic data are widely used in student performance prediction (Kuzilek, Hlosta, and Zdrahal 2017; Adnan et al. 2021; Waheed et al. 2020; Fancsali et al. 2018) due to its criticality in predicting each student's performance accurately (Li et al. 2021; Sabnis, Yu, and Kizilcec 2022), though the use of such data is often debated (Baker and Hawn 2022). Moreover, the demographic bias often stems from the lack of data and difference in behavior of underrepresented groups instead of the demographic information itself (Baker and Hawn 2022), as explained in the *Introduction* section.

It is important to note that in our work, demographic data is not leveraged to make judgements on any students, which is different from the controversy of using demographics for decision-making (e.g., recidivism prediction (Chouldechova 2017), AI-assisted recruiting (Hunkenschroer and Luetge 2022), and automated grade assignment (Baker and Hawn 2022)). Instead, demographic data is only utilized to identify

and address the demographic bias and improve the prediction accuracy on underrepresented students. The improved prediction results can then be used to provide feedback to help underrepresented students enhance their metacognitive ability. In addition, to ensure privacy, the demographic data is anonymized, making it impossible to link the demographic information to any individual (Kuzilek, Hlosta, and Zdrahal 2017).

Definition 2 Behavioral Data ($\boldsymbol{X^B}$): We define $\boldsymbol{X^B} = \{X_1^B, X_2^B, ..., X_I^B\}$ as the study behavior (e.g., late assignment records, online activities, and previous performance in other courses) of all students in a course, where X_i^B for $i \in \{1, 2, ..., I\}$ is the study behavior of the i^{th} student.

The behavioral data in our problem setting is measured by activities (e.g., clickstream data on different activities) on the online learning platform per day during the semester, which is widely used in student performance prediction frameworks (Adnan et al. 2021; Qiu et al. 2022; Karimi et al. 2020; Chu et al. 2021). The online activities indicate a student's effort and engagement from various aspects (e.g., reviewing course materials, completing course quizzes, participating in topic forums and collaborative activities) in an online course (Kuzilek, Hlosta, and Zdrahal 2017). Specifically, we leverage clickstream data to measure online activities as it is both easily collectible and effective in representing a student's learning trajectory across different activities in online education (Park et al. 2017; Chu et al. 2021). Moreover, clickstream data is demonstrated to effectively capture a student's behavioral pattern (e.g., conflict, confusion, and motivation in decision-making), which provides important information in predicting a student's performance (Rhim and Gweon 2022; Park et al. 2017).

Definition 3 Student Final Performance ($\boldsymbol{L}$): We define $\boldsymbol{L} = \{L_1, L_2, ..., L_I\}$ to be the final grade (e.g, *Fail*, *Pass*, and *Distinction*) of all students in a course. In particular, L_i for $i \in \{1, 2, ..., I\}$ represents the final performance of the i^{th} student, which is assigned by the course instructor.

Definition 4 Framework Prediction ($\widehat{\boldsymbol{L}}$): We define $\widehat{\boldsymbol{L}} = \{\widehat{L_1}, \widehat{L_2}, ..., \widehat{L_I}\}$ as the prediction of our framework, where $\widehat{L_i}$ represents the framework prediction for the i^{th} student.

The overall objective of our bias-aware crowd–AI collaborative framework is to explore AI and crowd intelligence to achieve both accurate and fair prediction results of student performance (i.e., maximizing the prediction accuracy while minimizing the demographic bias) as follows

$$\begin{aligned} &\arg\max_{\widehat{L_i}} \Pr(\widehat{L_i} = L_i \mid \boldsymbol{X^B}, \boldsymbol{X^D}), \quad \forall 1 \le i \le I \\ &\arg\min_{\widehat{L_i}} \mathrm{F}(\widehat{L_i}, L_i \mid \boldsymbol{X^B}, \boldsymbol{X^D}), \quad \forall 1 \le i \le I \end{aligned} \tag{1}$$

where $\mathrm{F}(\cdot, \cdot)$ denotes the fairness metric to measure the performance disparity among different demographic groups given the demographic attribute $\boldsymbol{D}$. The problem is challenging in terms of 1) identifying biased AI results to achieve an optimized tradeoff between accuracy and fairness and 2) addressing demographic bias using potentially biased crowd intelligence.

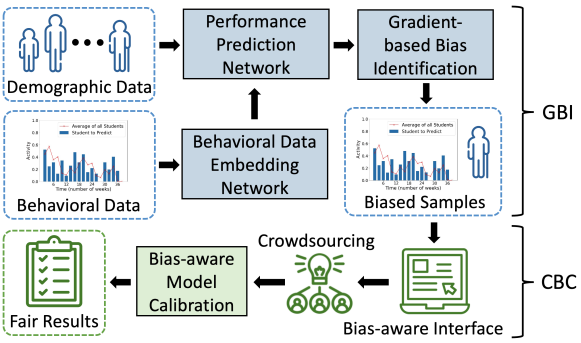


Figure 2: Overview of the DebiasEdu framework.

Solution

The DebiasEdu is a bias-aware crowd–AI collaborative approach that integrates AI and crowd intelligence to achieve accurate and fair student performance prediction. The overview of DebiasEdu is presented in Figure 2. In particular, the DebiasEdu consists of two key modules:

- *Gradient-based Bias Identification (GBI)*: it analyzes the variation in gradients of training samples to identify biased AI results from different demographic groups.
- *Crowd-guided Bias Calibration (CBC)*: it creates a bias-aware crowdsourcing interface design and a crowd-guided calibration model to address the demographic bias of AI and the cognitive bias of the crowd.

Gradient-based Bias Identification (GBI)

To effectively predict student final performance $\boldsymbol{L}$ using inputs of behavioral data $\boldsymbol{X^B}$ and demographic attributes $\boldsymbol{X^D}$, we first design two key networks as follows.

To extract useful information from the behavioral data $\boldsymbol{X^B}$ for student performance prediction (e.g., consistent hard work, extra hard work before the final), we design a *behavioral data embedding network* $f(\cdot)$ as follows:

$$E_k^B = f(X_k^B), \quad \forall 1 \le k \le K \tag{2}$$

where E_k^B represents the generated embedding of the behavioral data X_k^B for the k^{th} student. K is the total number of students in the training set. In particular, we utilize the long short-term memory (LSTM) model as the behavioral data embedding network $f(\cdot)$ in our setting, which has been shown to be effective in extracting information from sequential data (Li et al. 2020; Hassan et al. 2019).

After feature embedding, we build a student *performance prediction network* $g(\cdot, \cdot)$ that leverages the generated behavioral embedding and the demographic information to predict a student's final result as:

$$\widehat{L_k^{AI}} = g(E_k^B, X_k^D), \quad \forall 1 \le k \le K \tag{3}$$

where $\widehat{L_k^{AI}}$ is the *AI prediction* for the k^{th} student's final performance. In particular, the performance prediction network $g(\cdot, \cdot)$ is a multilayer perceptron consisting of a sequence of fully connected feedforward neural network layers to predict

a student's performance by comprehensively examining the embedded behavioral data.

To guide the behavioral data embedding network $f(\cdot)$ to effectively capture useful behavior pattern information (e.g., consistent work throughout the semester) and train the performance prediction network $g(\cdot,\cdot)$ to accurately predict a student's final performance result, we define the objective function $\mathcal{L}_{AI}$ for the AI model as follows:

$$\mathcal{L}_{AI} = \mathcal{L}_{\text{CE}}\left(g\left(f(X_k^B), X_k^D\right), L_k\right), \quad \forall 1 \leq k \leq K \tag{4}$$

where $\mathcal{L}_{\text{CE}}$ is the cross entropy loss to measure classification accuracy. L_k is the ground truth label of the k^{th} student's final performance on the training set (Definition 3).

Given the designed AI model, the key focus of the GBI module is identifying biased AI results from the testing set for crowd intelligence to improve framework prediction fairness. We first define the set of these AI results as follows:

Definition 5 Crowdsourcing Subset ($\boldsymbol{S}$): We select a subset of students on the testing set where the AI model is likely to generate inaccurate predictions for crowd workers to improve. We focus particularly on selecting from underrepresented groups $\boldsymbol{U}$ since these students are more likely to receive incorrect predictions due to the lack of training data and difference in behavioral pattern (e.g., older students often need to complete more activities to achieve the same result compared to younger students). We formally define the crowdsourcing subset to include the behavioral and demographic data for the selected J students as $\boldsymbol{S} = \{\{X_1^B, X_1^D\}, ..., \{X_J^B, X_J^D\}\}$, where $J = \alpha I$.

We refer to the demographic data and behavioral data (Definition 1 and 2) of students as *samples* in the rest of the solution. It is observed that the AI prediction network is more likely to predict incorrectly for the samples with gradients varying significantly during the training process (Ren et al. 2018). These samples exhibiting more variant gradients are more likely to belong to underrepresented groups. This is because underrepresented samples, with different input data characteristics (e.g., behavioral patterns) compared to the non-underrepresented samples, pose greater challenges for deep neural networks to learn to predict accurately (Ren et al. 2018). Therefore, we define these samples whose gradients vary significantly during training as the *biased training samples*. Our objective is to identify biased training samples from different demographic groups inversely proportional to the number of students in each group (e.g., more samples from worse-performing underrepresented groups).

To identify biased training samples using gradient variation, we first define the *training sample gradient* $\boldsymbol{\nabla} = \{\nabla_1, \nabla_2, ..., \nabla_K\}$ to be the gradients of training samples with respect to the objective function $\mathcal{L}_{AI}$ as follows:

$$\nabla_k = \text{E}\left[\frac{\partial \mathcal{L}_{AI}}{\partial\{X_k^B, X_k^D\}}\right], \quad \forall 1 \leq k \leq K \tag{5}$$

where $\text{E}[\cdot]$ denotes the expectation and ∂ denotes the partial derivative. The training sample gradient can be computed by the chain rule using derivatives of each neural network layer.

Definition 6 Gradient Variance ($\boldsymbol{V}$): We define $\boldsymbol{V} = \{V_1, V_2, ..., V_K\}$ to be the variance of sample gradient $\boldsymbol{\nabla}$:

$$V_k = \text{Var}\left[\frac{\partial \mathcal{L}_{AI}}{\partial\{X_k^B, X_k^D\}}\right], \quad \forall 1 \leq k \leq K \tag{6}$$

where $\text{Var}[\cdot]$ denotes the variance. In particular, the variance of sample gradient can be approximated by the average gradient in different epochs (Chang, Learned-Miller, and McCallum 2017), where the first several epochs are eliminated due to unstable performance at the beginning of training.

We select a subset of training samples with the top α largest gradient variances in the training set (i.e., *variant gradient subset*), where α is selected empirically based on the tradeoff between the algorithmic fairness and the crowdsourcing budget. However, it remains challenging to identify a subset of samples with gradients varying significantly from the *testing* set since there are no ground truth annotations available to even train a model and compute gradients. Therefore, we select the testing samples that share a similar behavioral pattern as the training samples in the selected variant gradient subset. This idea is motivated by the fact that an AI model generates similar predictions and gradients for input samples with similar characteristics (e.g., behavioral patterns) (Charpiat et al. 2019). We introduce the measurement to identify the crowdsourcing subset $\boldsymbol{S}$ of demographically biased testing samples as follows:

Definition 7 Bias Measurement ($\boldsymbol{B}$): We define $\boldsymbol{B} = \{B_1, B_2, ..., B_I\}$ to be the bias measurements of all studied testing samples. In particular, the bias measurement B_i for the i^{th} student is formally defined as follows:

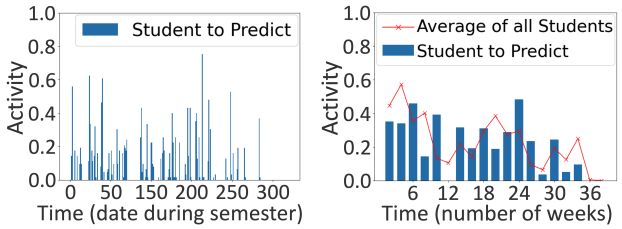
$$B_i = \sum_{k=1}^{K}\left(\left\|X_k^B - X_i^B\right\|_2 + \left\|X_k^D - X_i^D\right\|_2\right), \forall 1 \leq i \leq I \tag{7}$$

where $\|\cdot\|_2$ denotes the L2-norm of a vector. In particular, a lower value of the bias measurement B_i indicates a larger bias (i.e., a higher similarity with the variant gradient subset) for the i^{th} student in the testing set.

Based on the bias measurement for all testing samples, we then select the samples with top α lowest B_i from the testing set to generate the crowdsourcing subset $\boldsymbol{S}$ that primarily contains underrepresented students who are likely to receive inaccurate AI predictions. In the next subsection, we discuss how to use crowd intelligence to address the identified bias.

Crowd-guided Bias Calibration (CBC)

Given the selected crowdsourcing subset $\boldsymbol{S}$ from the GBI module discussed above, we then design a crowdsourcing interface and a model calibration mechanism to address the identified demographic bias while mitigating the negative impact of cognitive bias from crowd workers as illustrated in the *Introduction* section. In particular, we leverage crowd intelligence to work on the student performance prediction task and mitigate demographic bias. For an in-depth effectiveness analysis of using the crowd in student performance prediction, please refer to the *Crowdsourcing Settings and Pilot Study* in the *Evaluation* section.



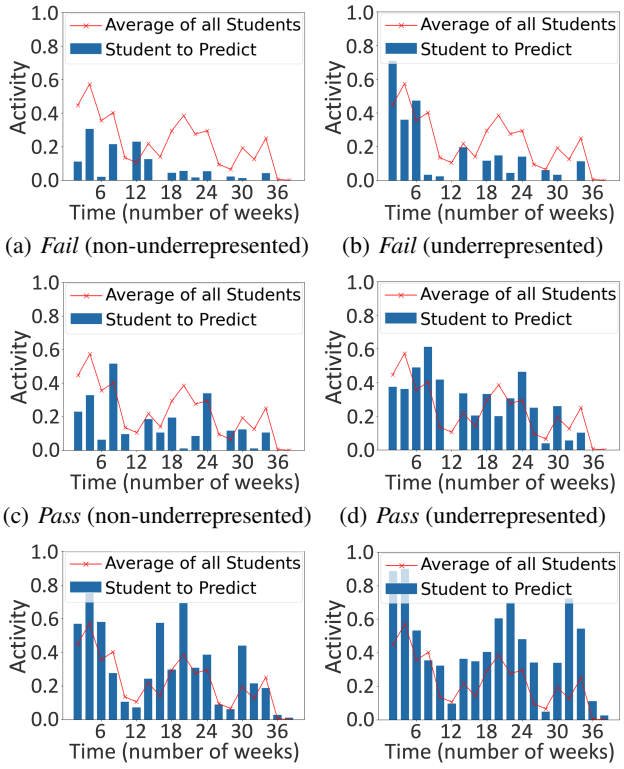
(a) Visualization of the original per-day behavioral data. (b) Behavioral data visualization design for the crowd.

Figure 3: Examples of the original behavioral data and our behavioral data visualization design.

We first design the visualization of behavioral data since student performance prediction based on behavioral data is not a trivial task for crowd workers. In particular, humans are often not good at analyzing the raw data (e.g., dozens or hundreds of numbers) compared to AI models, which motivates our design of a clear visualization of the high-level behavioral patterns (e.g., the general trend of activities and consistent hard work) to the crowd workers. The behavioral data X_i^B (Definition 2) for the i^{th} student are their activities on the online learning platform *per day* during the semester, which is shown in Figure 3(a). However, crowd workers often do not need such detailed information to predict student performance accurately. In particular, we observe that even those students who achieve *Distinction* results in many classes do not spend time on every course every day, highlighting the fact that accumulative activities within a certain time period can be more informative to help crowd workers to predict accurately. Therefore, we present the accumulative activities on a *bi-weekly* basis of a student to crowd workers using the blue bars shown in Figure 3(b). In addition to the bar plot of the behavioral data, we also add the *average activities* of all students in a course to help crowd workers make their predictions.

However, crowd workers are observed to have cognitive bias (Draws et al. 2021), which can lead to inaccurate crowd prediction in student performance prediction (Hettiachchi et al. 2020). Therefore, our next question is how to design a crowdsourcing interface to address the cognitive bias of the crowd. In particular, confirmation bias is a key cognitive bias of humans observed in prediction tasks (Draws et al. 2021; Abrahamyan et al. 2016). We define it as follows:

Definition 8 Confirmation Bias: Confirmation bias of the crowd refers to the fact that crowd workers can be overly influenced by their preexisting beliefs. For example, crowd workers can be conservative in predicting a *Distinction* result if they believe *Distinction* is assigned to a really small percentage of students. The confirmation bias is more obvious when a crowd worker is only presented with the behavioral data visualization of a specific student, since they need to predict completely based on their preexisting beliefs if no additional information (e.g., annotation examples provided by the task administrator as reference) is provided to the crowd workers regarding the labeling tasks.



(a) *Fail* (non-underrepresented) (b) *Fail* (underrepresented)

(c) *Pass* (non-underrepresented) (d) *Pass* (underrepresented)

(e) *Distinct* (non-underrepresented) (f) *Distinct* (underrepresented)

Figure 4: Anchoring examples for students in the non-underrepresented and underrepresented age group.

We design an approach to address the confirmation bias of the crowd by leveraging the anchoring effect of human cognition. We first define the anchoring effect as follows:

Definition 9 Anchoring Effect: Anchoring effect refers to the fact that crowd workers can be influenced by the first few examples they see and then use these examples as the anchor for the subsequent prediction.

We can leverage this cognitive characteristic of the crowd to train the crowd workers to calibrate their preexisting prediction criteria with only a few *anchoring examples* for each student performance category (e.g., *Fail*, *Pass*, and *Distinction*). For instance, anchoring examples selected from the training set of a STEM course are shown in Figure 4(a), 4(c), and 4(e). Note that we cannot simply train the AI model with such anchoring examples since AI models often rely on a large number of data samples for effectively predictions. Even the few-shot learning methods still depend on large-scale datasets to pre-train data representations, which are not available in our problem setting.

In addition, while crowd workers achieve better overall accuracy and fairness compared to the AI model on the selected crowdsourcing subset $\boldsymbol{S}$ (Definition 5), the crowd prediction accuracy of *underrepresented groups* can still be worse than the accuracy of *non-underrepresented groups*. Given the difference in behavioral patterns among different

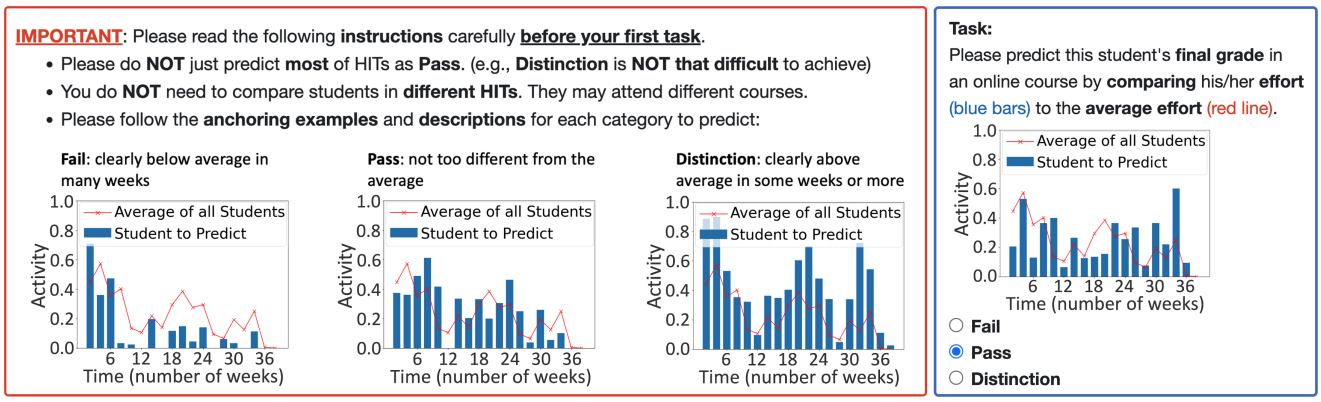


Figure 5: Crowdsourcing interface instruction and task design to address the demographic and confirmation bias. For the task of each student, we present the anchoring examples of the demographic group to which this student belongs in the instructions.

demographic groups, the demographic bias can be further addressed by showing anchoring examples of *each demographic group* to crowd workers. For instance, anchoring examples for students in the underrepresented age group (i.e., age $\geq$ 35) of the STEM course are shown in Figure 4(b), 4(d), and 4(f). We can clearly observe the behavioral difference between these underrepresented examples and the anchoring examples selected from the non-underrepresented group shown in Figure 4(a), 4(c), and 4(e): underrepresented students need to complete much more activities to achieve the same result compared to non-underrepresented students. Observed behavioral difference demonstrates that the crowd prediction accuracy and fairness can be further enhanced by presenting accurate anchoring examples from each demographic group to help crowd workers form accurate performance criteria. Our pilot studies show an 18.9% performance improvement when using the anchoring examples from each demographic group compared to not using such anchoring examples (refer to the *Crowdsourcing Settings and Pilot Study* in *Evaluation* section for detailed settings).

We present our crowdsourcing interface design for student performance prediction in Figure 5. For the prediction task of each student, we present the anchoring examples and corresponding descriptions of the demographic group this student belongs to. For instance, in Figure 5, since the sample student in the prediction task belongs to the underrepresented age group, our interface also presents the anchoring examples for this group (Figure 4(b), 4(d), and 4(f)) in the instructions. The objective of this interface design is to 1) address the confirmation bias by presenting anchoring examples of each performance category and 2) reduce the bias caused by the difference in behavior patterns among demographic groups by showing demographic group-wise anchoring examples.

We collect crowd predictions from the crowdsourcing platform using our bias-aware interface design shown in Figure 5. Only samples in the crowdsourcing subset $\boldsymbol{S}$, which are likely to be underrepresented samples receiving inaccurate predictions from AI models, are predicted by the crowd to explore the tradeoff between improving algorithmic fairness and limiting crowdsourcing budget. We observe that crowd workers might have different levels of accuracy in terms of providing accurate responses. Hence, instead of directly applying the majority voting strategy to obtain the aggregated crowd labels that is known to be suboptimal when crowd workers have different reliability (Zhang et al. 2019b), we leverage an estimation theory-based truth discovery model (Wang et al. 2012) to jointly derive the truthful crowd labels of the queries as well as the reliability of the workers. Let $\widehat{L_j^C}$ for all $j \in [1, J]$ represent the *aggregated crowd prediction* of students in the crowdsourcing subset $\boldsymbol{S}$. We then design a crowd offloading strategy to effectively address the biased AI results using the aggregated crowd labels. In our strategy, for all J students, the truthful labels $\widehat{L_j^C}$ derived from the crowd are used to replace the AI predictions $\widehat{L_j^{AI}}$ of the testing samples in the crowdsourcing subset $\boldsymbol{S}$ to generate the final framework prediction $\widehat{\boldsymbol{L}}$ (Definition 4).

Evaluation

Student Performance Prediction Datasets

To evaluate the accuracy and fairness of the proposed DebiasEdu framework, we leverage the demographic, behavioral, and performance data collected from the online learning platform Open University (Kuzilek, Hlosta, and Zdrahal 2017). In particular, we take the age of students as the demographic attribute in our evaluation since potential disadvantages have been observed for underrepresented older students in online education (Kuzilek, Hlosta, and Zdrahal 2017). Following the common practice in fairness applications (Hardt, Price, and Srebro 2016), we categorize the age attribute into two demographic groups (i.e., age less than 35 and age greater than or equal to 35). The behavioral data is measured by the activities (i.e., clickstream data) on different web pages (e.g., course material, course quizzes, topic forums, and collaborative activities) on the online learning platform per day for each student. The ground truth label of a student's final performance is assigned by the course instructors into three different levels (i.e., *Fail*, *Pass*, and *Distinction*). We use two datasets for different types of courses to comprehensively evaluate our DebiasEdu framework. In

Datasets	STEM	Social Science
Total Number of Students	1,938	1,767
Percent of *Fail*	34.0%	36.4%
Percent of *Pass*	58.2%	51.3%
Percent of *Distinction*	7.8%	12.3%
Percent of $Age < 35$	75.6%	67.6%
Percent of $Age \geq 35$	24.4%	32.4%

Table 1: Student performance prediction dataset statistics.

particular, the first dataset is collected from a STEM course, and the second dataset is collected from a Social Science course. Statistics of the two datasets are shown in Table 1.

Crowdsourcing Settings and Pilot Study

We deploy our interface design shown in Figure 5 to collect the crowd prediction from Amazon Mechanical Turk (AMT), one of the largest crowdsourcing platforms that provides the access to a massive number of online crowd workers worldwide with reasonable costs. To ensure the crowdsourcing quality, we set the qualification requirement as follows: the crowd workers must have completed over 10,000 approved tasks with an overall approval rate greater than 95% before starting to work on our task. The inter-worker agreements of different crowd workers are 0.664 and 0.637 in terms of the Cohen's Kappa score (Kappa) on the STEM course dataset and the Social Science course dataset, respectively. A Kappa score greater than 0.6 typically indicates good agreement (Cohen 1960). We pay \$0.05 to a crowd worker for each prediction task. We follow the Institutional Review Board protocol approved for this project. In our evaluation, we set the percentage α of crowdsourcing samples as 15% and the number of crowd workers as 5.

We first demonstrate the effectiveness of utilizing general crowd workers without educational domain knowledge in the student performance prediction task using both quantitative and qualitative analysis. In particular, we conduct a pilot study using the crowdsourcing subset (Definition 5) on the STEM course that includes 50 sample students, which are predicted by crowd workers in our experiments. We recruit both general crowd workers and educational practitioners (i.e., crowd workers who engage in educational activities as job responsibilities) to predict the final grades of these students using the same crowdsourcing task design (Figure 5). Our objective is to study if educational domain knowledge is required to conduct this task by comparing the crowdsourcing performance of general crowd workers and educational workers. The educational workers are selected on AMT using the premium qualification of job function[1]. We set the number of crowd workers per student to be 5. Our crowdsourcing experiments involved the participation of 69 educational workers and 113 general workers. The difference in the number of crowd workers is related to the fact that there are more general crowd workers available on AMT compared to educational workers. The collected predictions from educational workers are aggregated by the estimation theory-based truth discovery model introduced in the Solution section for each student. We utilize the same estimation theory-based aggregation model for general crowd workers in this study to ensure a fair comparison. By comparing the aggregated prediction results, we observe that general crowd workers and educational workers achieve an agreement of 0.746 in terms of the Kappa score. The notable consensus demonstrates that our student performance prediction task can be completed by general crowd workers without educational domain knowledge. In addition, for the recruited educational practitioners, we further ask them the following question: "Based on your work experience in education, do you think completing this task requires educational domain knowledge? If you think it is required, please provide explanations of what domain knowledge is required." Collected results indicate that 95.7% of educational workers involved in the study believe that no educational domain knowledge is required to effectively conduct the student performance prediction task. Specifically, some educational workers justify their conclusions by acknowledging the clarity of our prediction task, such as "I don't think it is required as the graphical representation makes it easy to predict". To conclude, the quantitative prediction comparison and qualitative inquiry results consistently demonstrate the effectiveness of recruiting general crowd workers to work on the student performance prediction task.

To further verify the effectiveness of our crowdsourcing task design, we formulate the following question to ask the recruited educational workers after they finish the prediction tasks: "Based on your work experience in education, do you feel comfortable predicting a student's final grade in an online course based on online activities (e.g., measured by clickstream data)?" A noteworthy 87.0% of educational practitioners felt comfortable conducting this task. This substantial percentage serves as evidence of the effectiveness of our task design since it is important to note that no measurements can 100% effectively predict a student's final grade except for the final grade itself. Particularly, we receive responses from educational workers that endorse our task design based on their professional domain experience, such as "I feel comfortable to predict a student's final grade because I do this work in my current job." The inquiry results confirm the effectiveness of our crowdsourcing task design in predicting a student's performance.

Baselines and Evaluation Settings

In our evaluation, we compare our DebiasEdu with a rich set of state-of-the-art AI, Fair AI, and Crowd–AI baselines.

1) AI Baselines: **ANN** (Waheed et al. 2020) utilizes a deep neural network to predict a student's performance based on a set of handcrafted features (e.g., clicks in a course, clicks on the assignment web page). **BCEP** (Qiu et al. 2022) classifies and fuses different types of online behavior (e.g., consistent hard work) to predict a student's performance. **SPDN** (Li et al. 2020) utilizes an LSTM-based feature extraction network and a convolutional feature fusion network to predict a student's performance from online learning records.

2) Fair AI Baselines: **JMLR19** (Zafar et al. 2019) integrates fairness measurements (e.g., false positive parity) as

[1]https://requester.mturk.com/pricing

Category	Algorithm	STEM Course				Social Science Course			
		Accuracy	F1-Score	Kappa	MCC	Accuracy	F1-Score	Kappa	MCC
AI	ANN	0.6059	0.6177	0.3142	0.3164	0.5500	0.5567	0.2230	0.2238
	BCEP	0.7322	0.7039	0.4733	0.4786	0.7086	0.6659	0.4405	0.4600
	SPDN	0.7118	0.7162	0.5011	0.5083	0.7059	0.6864	0.4430	0.4691
Fair AI	JMLR19	0.6500	0.6706	0.4282	0.4432	0.6588	0.6734	0.4122	0.4183
	NeurIPS21	0.7000	0.7179	0.4870	0.4927	0.6882	0.6959	0.4516	0.4586
	VS	0.6676	0.6833	0.4424	0.4514	0.5265	0.5727	0.3059	0.3392
Crowd–AI	StreamCollab	0.7147	0.7169	0.4949	0.4981	0.6824	0.6868	0.4418	0.4482
	DeepActive	0.7382	0.7228	0.5063	0.5089	0.6324	0.6468	0.3593	0.3692
	LearningLoss	0.6882	0.6717	0.4144	0.4352	0.6882	0.6717	0.4144	0.4352
Ours	**DebiasEdu**	**0.8294**	**0.8283**	**0.6844**	**0.6850**	**0.7647**	**0.7556**	**0.5676**	**0.5818**

Table 2: Evaluation results of student performance prediction *Accuracy* on the STEM and Social Science course datasets.

constraints during training to achieve fair performance prediction. **NeurIPS21** (Bendekgey and Sudderth 2021) utilizes data re-weighting and fairness constraints (e.g., equal opportunity) to achieve robust fairness in student performance prediction. **VS** (Kini et al. 2021) is a vector-scaling-based optimization approach that utilizes multiplicative and logit adjustments for fair group-sensitive classification.

3) Crowd–AI Baselines: **StreamCollab** (Zhang et al. 2021) is a crowd–AI system that leverages uncertainty quantification and crowd knowledge fusion for effective student performance prediction. **DeepActive** (Sener and Savarese 2018) is a deep active learning framework that identifies a core set of samples and integrates the crowd on them to improve prediction accuracy. **LearningLoss** (Shukla and Ahmed 2021) is a crowd–AI framework that leverages a task-agnostic loss design to efficiently integrate AI and the crowd for accurate student performance prediction.

For a fair comparison, we use the same inputs for all compared schemes: 1) the demographic attribute of age for all students, 2) the behavioral data of online activities per day for all students, and 3) the crowd prediction collected from the crowdsourcing platform for students in the crowdsourcing subset. Our DebiasEdu and all baselines are implemented using PyTorch libraries and trained on NVIDIA RTX 6000 GPUs. We use the Adam optimizer with a learning rate of 1×10^{-3} to train all compared models. We set the batch size to 20 and train the models for over 200 epochs.

To evaluate the model accuracy, we leverage four representative metrics for multi-class classification (Parker 2011): 1) Accuracy, 2) F1-Score, 3) Cohen's Kappa Score (Kappa), and 4) Matthews Correlation Coefficient (MCC). We include Kappa and MCC since our datasets are imbalanced as shown in Table 1 and these metrics have been demonstrated to be reliable in evaluating prediction accuracy given imbalanced data (Chicco and Jurman 2020). Higher values of these accuracy metrics indicate better performance. To evaluate the model fairness, we utilize four commonly used fairness metrics (Hardt, Price, and Srebro 2016; Zafar et al. 2017): 1) True Positive Parity (T.P. Parity) (i.e., Equal Opportunity), 2) False Positive Parity (F.P. Parity), 3) Equalized Odds (Eq. Odds), and 4) Accuracy Parity (Acc. Parity). Lower values of the fairness metrics indicate less bias and better fairness.

Evaluation Results

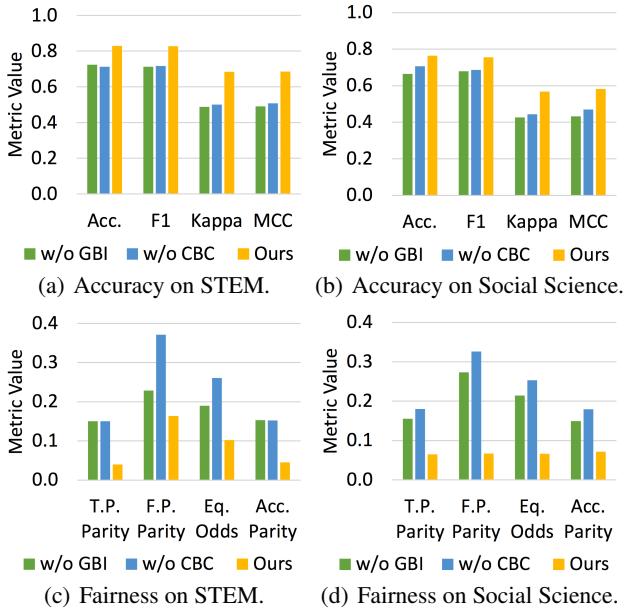
Accuracy Comparisons First, we evaluate the accuracy of all compared approaches in student performance prediction on the STEM course and Social Science course datasets. Evaluation results are shown in Table 2. We observe that our DebiasEdu consistently outperforms all baselines on all accuracy metrics. For example, on the STEM course dataset, the performance gains of our DebiasEdu compared to the best-performing baseline DeepActive on Accuracy, F1-Score, Kappa, and MCC are 12.3%, 14.6%, 35.2%, and 34.6%, respectively. Such performance gains can be attributed to the fact that our DebiasEdu framework develops a novel gradient-based module to identify the demographic bias of AI and designs a debias-driven crowd–AI collaboration module to address the identified bias and improve the overall student performance prediction accuracy.

Fairness Comparisons Second, we compare the fairness of our DebiasEdu and the compared baselines on the two datasets. The evaluation results are presented in Table 3. We note that the DebiasEdu achieves consistent performance gains compared to all baselines on both datasets by reaching the lowest prediction bias. For instance, on the Social Science course dataset, the decreases in T.P. Parity, F.P. Parity, Eq. Odds, and Acc. Parity of our DebiasEdu compared to the best-performing baseline BCEP are 55.2%, 75.3%, 68.3%, and 50.8%, respectively. The significant improvements in fairness demonstrate that our DebiaEdu approach successfully identifies and addresses the demographic bias in student performance prediction by carefully modeling the AI bias by gradient variation and designing a novel crowdsourcing interface to reduce the crowd cognitive bias.

Ablation Study Next, we conduct an ablation study to evaluate the contribution of the two key modules (i.e., GBI and CBC) of our DebiasEdu framework. We present the accuracy and fairness evaluation results when removing each of the two modules in DebiasEdu. In particular, to remove the GBI module, we randomly select 15% of samples from the testing set for crowd prediction and model calibration.

Category	Algorithm	STEM Course				Social Science Course			
		T.P. Parity	F.P. Parity	Eq. Odds	Acc. Parity	T.P. Parity	F.P. Parity	Eq. Odds	Acc. Parity
AI	ANN	0.2150	0.2056	0.2103	0.2129	0.2150	0.2824	0.2487	0.2158
	BCEP	0.1400	0.3253	0.2301	0.1377	0.1450	0.2720	0.2085	0.1453
	SPDN	0.1550	0.3714	0.2607	0.1520	0.1800	0.3259	0.2529	0.1795
Fair AI	JMLR19	0.1500	0.2526	0.2013	0.1470	0.1950	0.3675	0.2813	0.1973
	NeurIPS21	0.1700	0.1786	0.1743	0.1611	0.2850	0.1919	0.2385	0.2833
	VS	0.1350	0.2316	0.1833	0.1388	0.2300	0.3629	0.2965	0.2277
Crowd–AI	StreamCollab	0.1450	0.4909	0.3130	0.1347	0.2350	0.1290	0.1820	0.2339
	DeepActive	0.1250	0.2678	0.1964	0.1334	0.1750	0.2720	0.2235	0.1719
	LearningLoss	0.1050	0.2717	0.1884	0.1096	0.1600	0.3324	0.2462	0.1648
Ours	**DebiasEdu**	**0.0400**	**0.1635**	**0.1017**	**0.0451**	**0.0650**	**0.0672**	**0.0661**	**0.0716**

Table 3: Evaluation results of student performance prediction *Fairness* on the STEM and Social Science course datasets.



(a) Accuracy on STEM. (b) Accuracy on Social Science.

(c) Fairness on STEM. (d) Fairness on Social Science.

Figure 6: Ablation study on the two datasets.

The sampling rate is the same as the one used in our framework to ensure a fair comparison in terms of crowdsourcing budget. To remove the CBC module, we utilize the crowd prediction on the crowdsourcing subset to retrain the AI model. The accuracy and fairness evaluation results on two datasets are shown in Figure 6. The evaluation results demonstrate that both the GBI and CBC modules make critical contributions to the DebiasEdu framework in terms of both prediction accuracy and fairness.

Discussion of Benefits for Students

The accurate and fair student performance prediction results can be utilized to provide feedback to students, thereby enhancing their metacognitive abilities (Boud, Lawson, and Thompson 2015). Figure 7 shows the sample feedback design for students in our pilot testing. First, we incorporate a self-estimation page where students are prompted to estimate their final grades and specify their desired grades. Second, we design a model prediction and suggestion page that offers 1) predicted final grades from our framework and 2) suggestions to help students refine their self-estimation and achieve the desired final grades given current completed activities. Qualitative results from initial pilot testing suggests that self-estimation of learning performance relative to an accurate AI prediction leads to students thinking critically about their own knowledge. Specifically, the results involve participants trying to decipher why their self-estimation differs from AI predictions by assessing their own knowledge.

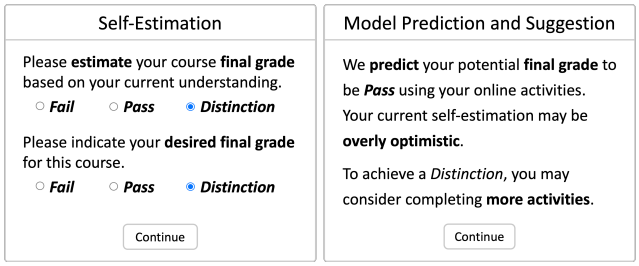


Figure 7: Sample feedback design for students in an online course based on our prediction results.

Conclusion

In this paper, we develop the DebiasEdu to address the demographic bias in student performance prediction for online education. In particular, we design a crowd–AI collaborative framework that effectively melds AI and crowd intelligence to accurately predict student performance in online education and jointly minimize the demographic bias of AI and the cognitive bias of the crowd. Evaluation results on two online courses demonstrate that DebiasEdu achieves consistent performance gains compared to all state-of-the-art baselines in terms of both prediction accuracy and fairness. We believe our DebiasEdu provides useful insights to address the demographic bias problem in other online education applications (e.g., exam assessment and cheating detection).

Acknowledgments

This research is supported in part by the National Science Foundation under Grant No. IIS-2202481, CHE-2105005, IIS-2008228, CNS-1845639, CNS-1831669. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R.; et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76: 243–297.

Abdi, S.; Khosravi, H.; and Sadiq, S. 2020. Modelling learners in crowdsourcing educational systems. In *International Conference on Artificial Intelligence in Education*, 3–9. Springer.

Abrahamyan, A.; Silva, L. L.; Dakin, S. C.; Carandini, M.; and Gardner, J. L. 2016. Adaptable history biases in human perceptual decisions. *Proceedings of the National Academy of Sciences*, 113(25): E3548–E3557.

Adnan, M.; Habib, A.; Ashraf, J.; Mussadiq, S.; Raza, A. A.; Abid, M.; Bashir, M.; and Khan, S. U. 2021. Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *Ieee Access*, 9: 7519–7539.

Albreiki, B.; Zaki, N.; and Alashwal, H. 2021. A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences*, 11(9): 552.

Anahideh, H.; Asudeh, A.; and Thirumuruganathan, S. 2022. Fair active learning. *Expert Systems with Applications*, 199: 116981.

Baker, R. S.; and Hawn, A. 2022. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32: 1052–1092.

Bendekgey, H. C.; and Sudderth, E. 2021. Scalable and stable surrogates for flexible classifiers with fairness constraints. *Advances in Neural Information Processing Systems*, 34: 30023–30036.

Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Boud, D.; Lawson, R.; and Thompson, D. G. 2015. The calibration of student judgement through self-assessment: disruptive effects of assessment patterns. *Higher education research & development*, 34(1): 45–59.

Chang, H.-S.; Learned-Miller, E.; and McCallum, A. 2017. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30.

Charpiat, G.; Girard, N.; Felardos, L.; and Tarabalka, Y. 2019. Input similarity from the neural network perspective. *Advances in Neural Information Processing Systems*, 32.

Chen, X.; Chen, S.; Wang, X.; and Huang, Y. 2021. " I was afraid, but now I enjoy being a streamer!" Understanding the Challenges and Prospects of Using Live Streaming for Online Education. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3): 1–32.

Chicco, D.; and Jurman, G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, 21: 1–13.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153–163.

Chu, Y.-W.; Tenorio, E.; Cruz, L.; Douglas, K.; Lan, A. S.; and Brinton, C. G. 2021. Click-based student performance prediction: A clustering guided meta-learning approach. In *2021 IEEE International Conference on Big Data (Big Data)*, 1389–1398. IEEE.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20: 37–46.

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 48–59.

Fancsali, S. E.; Zheng, G.; Tan, Y.; Ritter, S.; Berman, S. R.; and Galyardt, A. 2018. Using embedded formative assessment to predict state summative test scores. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 161–170.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hasan, R.; Palaniappan, S.; Mahmood, S.; Abbas, A.; Sarker, K. U.; and Sattar, M. U. 2020. Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences*, 10(11): 3894.

Hassan, S.-U.; Waheed, H.; Aljohani, N. R.; Ali, M.; Ventura, S.; and Herrera, F. 2019. Virtual learning environment to predict withdrawal by leveraging deep learning. *International Journal of Intelligent Systems*, 34(8): 1935–1952.

Hettiachchi, D.; Van Berkel, N.; Kostakos, V.; and Goncalves, J. 2020. CrowdCog: A Cognitive skill based system for heterogeneous task assignment and recommendation in crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–22.

Hunkenschroer, A. L.; and Luetge, C. 2022. Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics*, 178(4): 977–1007.

Jiang, W.; and Pardos, Z. A. 2021. Towards equity and algorithmic fairness in student grade prediction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 608–617.

Karimi, H.; Derr, T.; Huang, J.; and Tang, J. 2020. Online Academic Course Performance Prediction Using Relational Graph Convolutional Neural Network. *International Educational Data Mining Society*.

Kini, G. R.; Paraskevas, O.; Oymak, S.; and Thrampoulidis, C. 2021. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34: 18970–18983.

Kizilcec, R. F.; and Lee, H. 2020. Algorithmic fairness in education. *arXiv preprint arXiv:2007.05443*.

Kobayashi, M.; Wakabayashi, K.; and Morishima, A. 2021. Human+ AI Crowd Task Assignment Considering Result Quality Requirements. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 97–107.

Kou, Z.; Zhang, Y.; Shang, L.; and Wang, D. 2021. Faircrowd: Fair human face dataset sampling via batch-level crowdsourcing bias inference. In *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, 1–10. IEEE.

Kuzilek, J.; Hlosta, M.; and Zdrahal, Z. 2017. Open university learning analytics dataset. *Scientific data*, 4(1): 1–8.

Li, J.; Zhang, M.; Li, Y.; Huang, F.; and Shao, W. 2021. Predicting students' attitudes toward collaboration: Evidence from structural equation model trees and forests. *Frontiers in Psychology*, 12: 604291.

Li, X.; Zhu, X.; Zhu, X.; Ji, Y.; and Tang, X. 2020. Student academic performance prediction using deep multi-source behavior sequential network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 567–579. Springer.

Madaio, M.; Egede, L.; Subramonyam, H.; Wortman Vaughan, J.; and Wallach, H. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1): 1–26.

Park, J.; Denaro, K.; Rodriguez, F.; Smyth, P.; and Warschauer, M. 2017. Detecting changes in student behavior from clickstream data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 21–30.

Parker, C. 2011. An analysis of performance measures for binary classifiers. In *2011 IEEE 11th International Conference on Data Mining*, 517–526. IEEE.

Prihar, E.; Patikorn, T.; Botelho, A.; Sales, A.; and Heffernan, N. 2021. Toward Personalizing Students' Education with Crowdsourced Tutoring. In *Proceedings of the Eighth ACM Conference on Learning@ Scale*, 37–45.

Qadir, J. 2023. Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In *2023 IEEE Global Engineering Education Conference (EDUCON)*, 1–9. IEEE.

Qiu, F.; Zhang, G.; Sheng, X.; Jiang, L.; Zhu, L.; Xiang, Q.; Jiang, B.; and Chen, P.-k. 2022. Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*, 12(1): 1–15.

Rastrollo-Guerrero, J. L.; Gómez-Pulido, J. A.; and Durán-Domínguez, A. 2020. Analyzing and predicting students' performance by means of machine learning: A review. *Applied sciences*, 10(3): 1042.

Ren, M.; Zeng, W.; Yang, B.; and Urtasun, R. 2018. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, 4334–4343. PMLR.

Ren, P.; Xiao, Y.; Chang, X.; Huang, P.-Y.; Li, Z.; Gupta, B. B.; Chen, X.; and Wang, X. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9): 1–40.

Rhim, J.; and Gweon, G. 2022. Understanding the Relationship Between Students' Learning Outcome and Behavioral Patterns using Touch Trajectories. In *Proceedings of the Ninth ACM Conference on Learning@ Scale*, 441–445.

Sabnis, S.; Yu, R.; and Kizilcec, R. F. 2022. Large-Scale Student Data Reveal Sociodemographic Gaps in Procrastination Behavior. In *Proceedings of the Ninth ACM Conference on Learning@ Scale*, 133–141.

Scheuerman, M. K.; Wade, K.; Lustig, C.; and Brubaker, J. R. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW1): 1–35.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*.

Shukla, M.; and Ahmed, S. 2021. A mathematical analysis of learning loss for active learning in regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3320–3328.

Stinar, F.; and Bosch, N. 2022. Algorithmic unfairness mitigation in student models: When fairer methods lead to unintended results. In *Proceedings of the 15th International Conference on Educational Data Mining*, 606.

Troussas, C.; Krouska, A.; and Sgouropoulou, C. 2020. Collaboration and fuzzy-modeled personalization for mobile game-based learning in higher education. *Computers & Education*, 144: 103698.

Waheed, H.; Hassan, S.-U.; Aljohani, N. R.; Hardman, J.; Alelyani, S.; and Nawaz, R. 2020. Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human behavior*, 104: 106189.

Wambsganss, T.; Janson, A.; Käser, T.; and Leimeister, J. M. 2022. Improving students argumentation learning with adaptive self-evaluation nudging. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–31.

Wang, D.; Kaplan, L.; Le, H.; and Abdelzaher, T. 2012. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proceedings of the 11th international conference on Information Processing in Sensor Networks*, 233–244.

Wasif, M.; Waheed, H.; Aljohani, N. R.; and Hassan, S.-U. 2019. Understanding student learning behavior and predicting their performance. In *Cognitive Computing in Technology-Enhanced Learning*, 1–28. IGI Global.

Xie, B.; Yuan, L.; Li, S.; Liu, C. H.; Cheng, X.; and Wang, G. 2022. Active learning for domain adaptation: An energy-based approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8708–8716.

Xu, Z.; Yuan, H.; and Liu, Q. 2020. Student performance prediction based on blended learning. *IEEE Transactions on Education*, 64(1): 66–73.

Yang, S. J.; Ogata, H.; Matsui, T.; and Chen, N.-S. 2021. Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2: 100008.

Yoo, D.; and Kweon, I. S. 2019. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 93–102.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. P. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1): 2737–2778.

Zafar, M. B.; Valera, I.; Rodriguez, M.; Gummadi, K.; and Weller, A. 2017. From parity to preference-based notions of fairness in classification. *Advances in Neural Information Processing Systems*, 30.

Zhang, D.; Zhang, Y.; Li, Q.; Plummer, T.; and Wang, D. 2019a. Crowdlearn: A crowd-ai hybrid system for deep learning-based damage assessment applications. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 1221–1232. IEEE.

Zhang, X.; Wu, Y.; Huang, L.; Ji, H.; and Cao, G. 2019b. Expertise-aware truth analysis and task allocation in mobile crowdsourcing. *IEEE Transactions on Mobile Computing*, 20(3): 1001–1016.

Zhang, Y.; Shang, L.; Zong, R.; Wang, Z.; Kou, Z.; and Wang, D. 2021. StreamCollab: A Streaming Crowd-AI Collaborative System to Smart Urban Infrastructure Monitoring in Social Sensing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 179–190.

Zhang, Y.; Zong, R.; Kou, Z.; Shang, L.; and Wang, D. 2022. CrowdNAS: A Crowd-guided Neural Architecture Searching Approach to Disaster Damage Assessment. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–29.