Exploring Visualization for Fairness in AI Education

Xinyuan Yan* Youjia Zhou† Arul Mishra‡ Himanshu Mishra§ Bei Wang¶

University of Utah

ABSTRACT

AI systems are becoming omnipresent in our daily lives, but they can sometimes be a source of bias for disadvantaged groups. Lack of fairness in AI systems is not just an engineering issue that influences public policy, it also has important implications for business ethics and corporate social responsibility. To educate nontechnical students at the business school, we have developed educational modules on fairness in AI that convey the importance of making not just accurate but also equitable business decisions. We introduce an educational module with six interactive components that illustrate how to detect, quantify, and mitigate biases in a logistic regression model. When such a module was deployed in a "Fair Algorithms for Business" course, it was shown to increase students' engagement and understanding. We further conducted a user study with 413 participants to examine whether adding visualizations and interactions (or not) could lead to an increased understanding of fairness concepts.

Index Terms: Human-centered computing—Visualization—Visualization design and evaluation methods; Social and professional topics—Computing education.

1 INTRODUCTION

Artificial Intelligence (AI) systems are becoming increasingly ubiquitous, deployed in various domains such as healthcare, market pricing, and college admission. Despite the benefits they may bring, these systems can also reflect, inject, or amplify societal biases into decisions [35]. Some examples include hiring systems that tend to favor applicants from certain demographic groups [3], and image recognition systems that are weak in recognizing darker skins [52].

Substantial technical progress has been made to enhance fairness in Machine Learning (ML), a subset of AI. These contributions primarily focus on the mathematical definition and measurement of fairness [14], prompting the development of various bias mitigation methods [35], which further sparks the discussion of the trade-off between accuracy and fairness [33]. Meanwhile, a recent proliferation in the development of fairness toolkits [2,5,16] makes these algorithms more accessible, empowering users to audit potential biases in ML models through libraries or visualization systems. Inspired by these advances, computer science (CS) education increasingly incorporates AI fairness into the curriculum [19,34], with the aim of teaching students to design responsible and equitable technologies.

However, many curricula around fairness in AI mainly cater to users with technical backgrounds (such as CS and engineering students). Educating nontechnical students on AI fairness is equally important. Passi and Vorvoreanu [40] showed that there is a risk of overreliance on AI, particularly among individuals with limited AI literacy; for instance, individuals may unquestioningly accept an AI system's recommended decision even when it gives consistently different predictions for different demographic groups. Such overreliance could not only jeopardize our own rights given the widespread deployment of AI systems today, but also pose potential threats to the benefits of others. For instance, students in business school who are training to be managers may regularly employ AI systems during critical decision-making, including hiring, admissions, and loan approvals. Thus, they need to know that ML models should be not only accurate in their predictions but also unbiased to all groups. Furthermore, even if users are aware of AI systems' bias, they may develop a high reliance on fairness metrics and debiasing methods. Although numerous fairness metrics exist, there is no consensus about which one is the best to use, and some criteria about fairness metrics are challenging to satisfy simultaneously [13]. Therefore, users should be aware of the trade-off among fairness metrics and be mindful in selecting appropriate metrics that best fit their decision environment. Similarly, different bias mitigation methods come with different assumptions and limitations [35]. Users should be able to compare multiple debiasing methods and, more importantly, understand the trade-off between fairness and accuracy [33].

Moreover, the impact of visualization on educating fairness in AI concepts has not been widely explored. Despite pedagogical research demonstrating that properly designed visuals could reinforce students' understanding of complex concepts [11,45,47], existing courses for AI fairness in CS education still fall short in taking advantage of visualization [19]. A recent study by Mashhadi et al. [34] explored the values of visualization in teaching CS students. It examined six open-source fairness tools that enable students to study algorithmic biases.

In light of these concerns, we aim to educate nontechnical students in higher education—in particular, business school students—about fairness in AI through visualization. We hypothesize that the utilization of visualization could increase students' learning outcomes. Specifically, we developed an educational module that features six types of interactive components in close collaboration with two professors from the School of Business. These components cover different aspects of fairness in AI, including ML pipeline explanations via logistic regression, fairness metrics, and bias mitigation methods. Our visualization design has three objectives. First, we enable users to perform what-if analysis by interactively changing model inputs to better understand the algorithms and the impacts of bias mitigation. Second, our design supports direct comparisons among multiple fairness metrics and mitigation methods, as well as highlights the trade-off between accuracy and fairness. Third, we seek to balance technical complexity and understandability for nontechnical users by hiding or simplifying algorithm details whenever appropriate. The educational module has been incorporated into a course—"Fair Algorithms for Business"— an elective available to graduate and undergraduate business students.

We conducted a user study involving 413 participants to examine how varying visualization formats impact education about fairness in AI, focusing specifically on learning the fairness metric of *disparate impact*. We developed learning materials using three degrees of visualization support. First, *texts and images* utilize texts and images of confusion matrices to explain the computation of fairness metrics. Second, *static visualization* combines delicately designed confusion matrices with a static visualization of fairness metrics. Third, *interactive visualization* enables interactivity via linked views

*e-mail: xyan@cs.utah.edu
†e-mail: zhouyj96180@gmail.com
‡e-mail: arul.mishra@utah.edu
§e-mail: himanshu.mishra@utah.edu
¶e-mail: beiwang@sci.utah.edu

between confusion matrices and fairness metrics, that is, users may change the values of confusion matrices to observe their impacts on fairness metric computation.

Our findings suggested that both static and interactive visualization could significantly enhance accuracy gain on recall questions compared to texts and images. We also found that static visualization demonstrated the most significant improvement in accuracy gain as reading time increased. Although participants provided positive feedback on interactive visualization, only a small percentage (13%) of them actually interacted with the visualization components. Participants with higher visual learning ability were more likely to recommend learning materials with interactive visualization, whereas those with lower visual learning ability were more likely to recommend static visualization. In summary, our work has explored the use of visualization for fairness in AI education, by providing:

- A characterization of tasks, challenges, and guidelines for creating visualization tools to teach nontechnical students about AI fairness concepts;
- An educational module that examines the use of visualization in detecting, quantifying, and mitigating biases in ML models;
- A user study involving 413 participants to evaluate the effectiveness of using varying visualization formats to teach fairness concepts and distill lessons that may inform the research of visualization in AI fairness education.

2 Related work

2.1 An Overview of Fairness in AI

Recent research on AI fairness has centered around fairness metrics and bias mitigation; see [35, 42] for surveys.

Fairness metrics. Different statistical definitions of fairness have been introduced to measure fairness in model predictions [14]. One popular strategy is to first define subpopulations using attributes like gender and race, often referred to as protected or sensitive attributes, and then compare disparities of model performances among these subpopulations. In this line of work, commonly used fairness metrics include statistical parity difference [9], equal opportunity difference [20], average odds difference [20], etc. However, it has been mathematically proven that these metrics cannot be simultaneously satisfied except in certain highly constrained cases [13]. Thus, the selection of a fairness metric relies on the decision context [35].

Bias mitigation. Numerous debiasing methods, mostly technical in nature, have been proposed to mitigate biases during different stages of an ML pipeline [35]. A pre-processing debiasing method aims to reduce the bias in the dataset by either assigning proper weights to data points [25] or transforming the representation of the dataset to obfuscate protected attributes [7, 55]. An in-processing debiasing method trains a model to maximize accuracy and reduce discrimination in ways that weaken the reliance between protected attributes and output [27, 56]. A post-processing debiasing method modifies the prediction to make it fairer, such as by adjusting the decision threshold [26]. The performance of these methods varies across different cases [42] and could impact model accuracy [33].

In this work, we employ visualization to help nontechnical users, in particular, business school students, understand commonly used fairness metrics and bias mitigation methods, as well as perform comparisons to comprehend fairness-accuracy trade-offs.

2.2 Visual Analytics for Fairness in AI

A number of visualization tools have been developed recently in various domains to enhance the accessibility of fairness algorithms. Many of these tools focus on identifying and quantifying biases in ML model predictions before deployment. Aequitas [44] and Google What-If Tool [51] calculate the fairness of model predictions concerning a single protected attribute. Other tools support the exploration of intersectional bias, where groups are defined by multiple protected attributes. For example, FairVis [6] and DiscriLens [49] empower users to discover and compare biased intersectional groups via either user-selected attributes or automated recommendations. Silva [54] identifies potentially biased attributes through causal relationships. RMExplorer [29] measures fairness across various patient subpopulations in disease risk models. FairRankVis [53] introduces a framework for exploring the bias in ranking decisions at both group and individual levels. Beyond bias measurement, some other tools support bias mitigation. FairSight [2] aims to understand, measure, diagnose, and mitigate biases across the entire pipeline of ML, including data, model, and outcome. D-BIAS [16] audits and mitigates biases of tabular data through causality networks and incorporates humans in the loop.

Fairness tools also demonstrate utility in the fields of Natural Language Processing and Computer Vision. For example, ShortcutLens [24] helps domain experts uncover spurious biases in Natural Language Understanding (NLU) benchmark datasets. VERB [43] explains how various debiasing methods influence the geometric structure of word embeddings. WordBias [15] identifies intersectional biases in word embeddings using Parallel Coordinates Plots. In Computer Vision, DASH [30] discovers and mitigates spurious correlations in the training data of image classification via data augmentation. SliceTeller [57] is used to detect biases in the outcome of image classification models.

Nonetheless, these visualization tools are primarily intended for research or commercial use, requiring users to have prior knowledge in programming, ML, or specific domains. Besides, these tools are specialized in distinct tasks that emphasize specific facets of AI fairness, through developing powerful interactive systems but typically with steep learning curves [32]. Our work aims to educate nontechnical users in higher education on various facets of AI fairness through carefully crafted visualizations.

The interactive demo provided by AI Fairness 360 [5] is most relevant to ours, which presents the bias measurement, mitigation, and comparison to the public. However, the IBM's tool conceals all technical details and is perceived to be overly simplified by Lee and Singh [32], ranking as the least effective by Mashhadi et al. [34].

2.3 Roles of Interactive Visualization in Education

The effectiveness of interactive visualization in pedagogy has been studied for some time, as surveyed in [10]. The mainstream research focuses on Algorithm Visualization (AV), which depicts the dynamic process of algorithms through interaction, visualization, and animation [11, 12]. Some literature reports mixed results [23, 38, 39] on evaluating the educational effectiveness of AV. Existing works also demonstrate the benefits of using AV [47, 48], especially in increasing the engagement level of learners [45, 46]. Moreover, a large body of work has studied the key features of educationally effective AV, such as accompanying visualization with textual explanations [48] and keeping visualization short and focused [46]. Drawing from these lessons, we derive guidelines for designing interactive visualizations tailored for fairness in AI education (Sec. 4.1).

Mashhadi et al. [34] assessed the the efficacy of six open-source fairness tools, including Aequitas, Dalex, FairLearn, Fairness 360, Responsibly, and What-if-tool, in teaching algorithmic fairness to CS students. They performed a qualitative comparison of these tools via focus groups, in terms of transparency (i.e. showing the inner-workings of bias mitigation techniques), dataset integration (i.e., flexibility with importing custom datasets), and interactivity (i.e., how student learning may be impacted by the tools' visualization styles and presentation of fairness criteria). Tools with options of interactivity were briefly mentioned to be overall favored by the focus groups in their study. However, they did not investigate in detail how interactive visualization impacts learning fairness concepts.

Similar to [34], our work is also motivated by the benefits, challenges, and opportunities of integrating visualization into fairness in AI education. However, although Mashhadi et al. obtained qualitative feedback from a class of undergraduate CS students, their obtained insights are not completely generalizable to nontechnical users. More importantly, whereas Mashhadi et al. assessed the pedagogical values of various fairness tools, we examine how varying visualization formats, i.e., adding visualizations and interactions (or not), could impact the performance of learning AI fairness concepts by nontechnical users.

3 Task Analysis and Design Rationale

Our goal is to design an educational module using visualization to educate nontechnical users—in particular, undergraduates and MBA students in the business school—about AI fairness concepts. We assume that these users possess a basic understanding of statistics but may lack knowledge in ML and programming.

To design our tool, we conducted weekly meetings with two professors (referred to as P1 and P2 throughout the paper) from the School of Business for approximately six months. They are our coauthors and have rich experience in teaching ML classes to business school students. Each meeting lasted an hour, and meeting notes were shared among all participants. We also recorded meetings on Zoom during in-depth discussions on the design requirements and expert feedback.

3.1 Task Analysis

Our discussions first focused on the key concepts necessary for a six-week introductory course on AI fairness, and identified the challenges beginners might encounter in learning the course material. We summarized three tasks (**T1**-**T3**) that our tool aims to support.

T1. Reveal the mechanism and biases of ML models. Users need to understand how ML works before their introduction to fairness concepts. The tool should assist in explaining ML models from three aspects, including *basic concepts*, such as training and test data, model training, and model evaluation (e.g., confusion matrix); *typical ML workflows*, in particular, training and evaluation processes; and the *influence of input data on the model prediction*, for instance, how the accuracy would change when modifying the training and test data. With a basic understanding of ML mechanisms, users may further understand and detect biases in model predictions based on fairness-related concepts, such as protected attributes and privileged/unprivileged groups. The challenge is to help nontechnical users learn numerous ML concepts in a short time frame.

T2. Illustrate different fairness metrics. Given the varied fairness metrics, it is important for the user to understand which metric best fits the task at hand rather than assume that any metric can be used in any situation [35]. The visualization tool should help the user understand how to choose a metric, how to compute fairness of the model output using the metric, and then determine whether the fairness score is within range (or not). Many existing fairness tools choose to explain the values and computations using plain text, which was shown to be overwhelming for nontechnical users [32].

T3. Understand bias mitigation and fairness-accuracy trade-off. As described in Sec. 2, three categories of debiasing methods (preprocessing, in-processing, and post-processing) have been proposed to enhance the fairness of ML models. For each debiasing method, the tool should help users understand which part of a ML pipeline it works on, how biases are mitigated, and what impact it has on the predictions. In addition, the tool should help users explore the tradeoffs between the fairness metrics and accuracy, as pursuing a higher level of fairness often leads to the compromise of accuracy. As there are numerous debiasing methods with different mechanisms, it would be time-consuming and unscalable to design visualization for each one of them. Therefore, the tool should focus on a few representative methods.

3.2 Design Rationale

Next, we went through an iterative process to design the static and interactive visualization components. We generated a number of design alternatives for each visualization component using sketches, prototypes, or existing fairness toolkits. These designs were then reviewed by the entire project team. They were polished, refactored, or discarded after discussion. Coupled with a literature review, we distilled five design rationales (**R1**-**R5**) to enhance the teaching of AI fairness concepts.

R1. Keep visualization simple and consistent. Although complex visual designs may be inevitable in systems developed for analyzing intricate data [49], in the context of education, P1 and P2 emphasized that visualizations and interactions should be simple and intuitive, echoing prior research [46] that advocates for short and focused interactive visualizations to enhance classroom effectiveness. Since students' attention span is limited during a class, they should concentrate more on understanding the core concepts rather than interpreting and learning the use of visualization. Additionally, given the diverse fairness concepts, maintaining consistency in visual encodings and interactions wherever appropriate could alleviate the burden of interpretation [50].

R2. Enable visual comparison. It is important for students to recognize differences among various fairness metrics, bias mitigation algorithms, and the trade-offs between fairness and accuracy, which could not only reduce the overreliance on fairness algorithms but also provide guidance on their usage in specific circumstances. Visualization proves to be an effective way to convey differences by using comparison techniques, such as juxtaposition, superposition, and explicit encodings [17]. Thus, visual comparison of fair algorithms should be provided for increased pedagogical value.

R3. Support what-if analysis. Previous research [48] showed that allowing students to change the input data and observe its impact on the results, commonly known as *what-if analysis*, promotes active learning. Recent visualization systems [51] also embraced this approach for probing models. Particularly, to facilitate comprehension of AI fairness concepts, interactive visualizations should empower students to witness how alterations to input data can impact model predictions, subsequently influencing fairness metrics, as commented by P1 and P2.

R4. Visualize appropriate degrees of technical details. It is challenging for nontechnical students to grasp intricate ML concepts. Therefore, the visualization should abstract and simplify technical details wherever necessary, such as model optimization, and focus on simple and interpretable ML models, such as linear and logistic regressions. Technical discussions on a complex or opaque model would detract from the main purpose of learning about AI fairness.

R5. Integrate visualization with texts and images. P1 and P2 addressed the need that visualizations should be integrated with learning materials, including text and images, to reinforce the comprehension of key fairness concepts and facilitate both in-class demonstration and self-exploration. However, Lee and Singh [32] found that user interface design could potentially overwhelm the user with information. It would be important to understand *how* the pedagogical efficacy of interactive visualization can be improved by integrating *the appropriate amount of* textual content that explains the algorithms and interprets the visualization and interaction without creating cognitive overload.

4 Interactive Visualization Components

We developed our educational module as a mixture of texts, images, and interactive visualizations, where texts and images are provided by instructors P1 and P2 to assist in explaining fairness concepts

(**R5**). Specifically, we designed six types of interactive components (**C1**-**C6**), with each focusing on one specific functionality (**R1**). They are used independently or collectively to explain different aspects of fairness in AI as outlined in **T1**-**T3**.

- **C1** displays a ML workflow via a logistic regression model;
- **C2** supports the customization of training and test data;
- **C3** enables the training and the evaluation of an ML model;
- **C4** explains and compares different fairness metrics;
- **C5** visualizes the results of pre-processing debiasing methods;
- **C6** visualizes the results of post-processing debiasing methods.

We use a loan example (Sec. 4.1) and a German Credit dataset [8] (Sec. 4.2-Sec. 4.4), to explain how these interactive components are used to educate nontechnical users on fairness in AI and how our design considerations align with the design rationales (**R1**-**R5**). We highlight key functionalities of these components and discuss additional features in the supplementary material.

4.1 Making Predictions Using ML Models

For nontechnical users, a good beginning for learning AI fairness is to understand how ML models make predictions. We design **C1** to illustrate a basic ML workflow using logistic regression, due to its simplicity and extensive use in practice. Here, a logistic regression is used to predict a binary variable (valued 0 or 1) via a logistic function. It assigns each data point a probability and makes a prediction of 1 when the probability is above a threshold (e.g., 0.5).

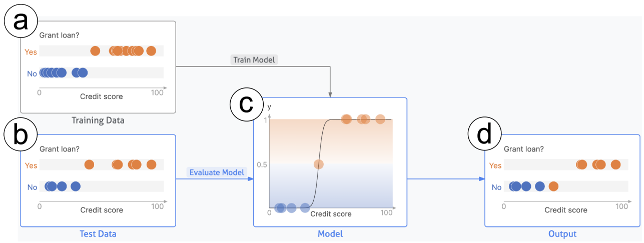


Figure 1: **C1** visualizes the workflow of a logistic regression.

We first visualize logistic regression applied to a toy dataset consisting of 30 data points, as illustrated in Fig. 1. Each data point represents a borrower and contains a feature called the *credit_score* and a binary label *loan_decision* that indicates whether the loan application was approved (1) or denied (0). We fit a logistic regression with a predictor variable (*credit_score*) and a predicted variable (*loan_decision*). We describe the overall design of **C1**, its training and evaluation processes, and the influence of input data on the prediction, addressing the three facets of learning in **T1**.

Overall design. The interface of **C1** visualizes the process of training and evaluating a logistic regression model (Fig. 1). It consists of four panels: training data panel (a), test data panel (b), model panel (c), and prediction panel (d). Arrows between panels highlight the ML workflow. In each panel, a node represents a data point (a borrower), colored by its *loan_decision* label, approval (1) in orange and denial (0) in blue. In panels (a), (b), and (d), the *x*-axis represents the input (predictor) variable (*credit_score*), whereas the *y*-axis represents the output (predicted) variable (*loan_decision*). Panel (c) depicts the learned relationship between the input and the output, where the *x*-axis represents the input variable, and the *y*-axis encodes the predicted probabilities of the logistic function. The background is colored by the probabilities.

For the training and test data in panels (a) and (b), respectively, nodes are placed based on their ground truth label (top row: orange nodes, bottom row: blue nodes). In the prediction panel (d), nodes are placed based on the model predictions.

ML workflow. For nontechnical students, we simplify the explanation of a ML pipeline by focusing on its training and evaluation processes. To start the training process, users click the "Train Model" button, and then the fitted curve of probabilities will be displayed in the model panel (c); see Fig. 1. To start the evaluation process, users click the "Evaluate Model" button, and then the predicted probabilities and the predicted labels are displayed in the model (c) and prediction panels (d), respectively. In the prediction panel (d), blue nodes on the top row and orange nodes on the bottom row indicate incorrect predictions. Various visual designs are introduced to help users better understand the process, such as highlighting the active panel that is being updated and displaying a text reminder to guide users toward the next step in the workflow.

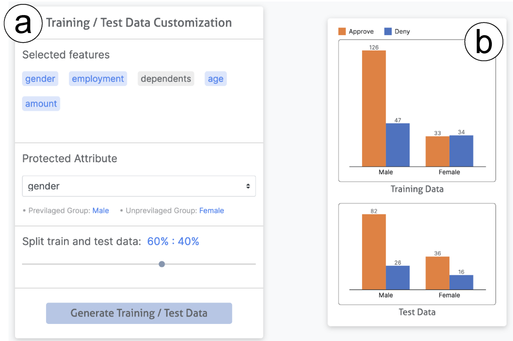
What-if analysis. We also design interactions for answering the following what-if question: how would a model prediction change if we modify the training or test data? Three types of interactions are provided in panels (a)-(b) to support modifications of the training/test data: dragging a node horizontally will change its input variable value; dragging a node away from a row will delete it; and double clicking any empty space on a row will add a new data point to the training/test data. Once the training/test data is modified, users can click on the buttons to retrain/reevaluate the model.

Design considerations. We aim to reduce visual and technical complexities by considering a single input variable, adopting easy-to-use interactions for what-if analysis such as drag-and-drop and clicking, and intentionally omitting certain technical details such as model validation (**R1**, **R4**). An alternative design for the model view is to show the training process by animating the curve. However, this design requires the introduction of iterative optimization that may overwhelm the students, where the main goal is for students to understand that training a model is essentially learning a mathematical relationship between the input and output variables.

4.2 Biases in Model Predictions

After explaining a logistic regression, we explain biases in model predictions (**T1**). We use a sample of the German Credit dataset [8] that contains 400 borrowers. Each borrower data point contains a binary label (*loan_decision*) and five additional features: *gender*, *age*, *employment* status (employed/unemployed), the number of *dependents* the borrower has, and the *amount* of loan requested.

Biases arise when the input data is divided into several subgroups by a *protected attribute*, such as race and gender, and the model predicts differently for different subgroups. Given a protected attribute, we identify two subgroups, the *privileged group* and the *unprivileged group*, in which the privileged group receives a more favorable outcome than the unprivileged group. Our goal is to help users identify and explore biases in model predictions via interactive visualization. We thus introduce interactive components **C2** and **C3**.


Training / Test Data Customization

Selected features

gender employment dependents age amount

Protected Attribute

gender

• Previlaged Group: Male • Unprevilaged Group: Female

Split train and test data: 60% : 40%

Figure 2: **C2** supports the customization of training and test data.

Customize the training and test data. We present **C2** to enable the customization of training and test data; see Fig. 2. In the control panel (a), users can select a subset of features as input variables, choose a protected attribute (e.g. *gender*) to define privileged and unprivileged groups, and specify the ratio for splitting the training and test data. When users click the "Generate Training/Test Data" button, the distribution of training and test data is displayed in the bar chart panel (b) based on the splitting. In Fig. 2, approval (orange) and denial (blue) labels are grouped by the protected attribute. Male and female borrowers are identified as privileged and unprivileged groups, respectively, based on the protected attribute *gender*.

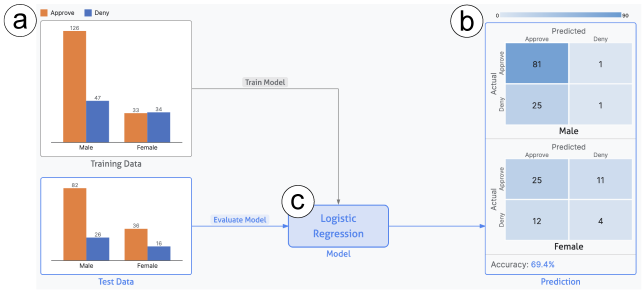


Figure 3: **C3** enables the training and evaluation of a ML model.

Visualize model performance on subgroups. After generating input data with **C2**, users could explore the model performance on different subgroups with **C3**. Similar to **C1**, **C3** demonstrates a ML workflow, but emphasizes the input data and the model prediction of subgroups defined by a protected attribute. As shown in Fig. 3, the input panel (a) visualizes the training and test data grouped by the protected attribute, and is consistent with the output of **C2** (Fig. 2b). The prediction panel (b) shows the confusion matrix for each subgroup and displays the prediction accuracy. We simplify the model panel (c) to a text box and hide its inner working. Similar to **C1**, **C3** highlights the active panels during the training/evaluation process and displays a text reminder when applicable.

Design considerations. For simplicity, we consider only binary protected attributes and treat the model as a black box, assuming that students have already obtained an overview of the ML pipeline from **C1**. We use grouped bar charts to visualize the input data due to their effectiveness in displaying and comparing data distributions [22]. These bar charts highlight the sources of bias, such as the inherent bias or the over-/under-representation in the training data. Additionally, we choose the confusion matrix to display predictions because it is a core concept in ML that forms the basis for computing various fairness metrics, and facilitates detailed subgroup comparisons.

4.3 Fairness Metrics

After users have identified the biases in model predictions by comparing confusion matrices of subgroups, we explain and compare four popular fairness metrics using **C4**: *statistical parity difference* (SPD), *disparate impact* (DI), *equal opportunity differences* (EOD), and *average odds difference* (AOD) (addressing **T2**). These metrics are also featured in AEquitas [44] and AI Fairness 360 [5].

C4 contains three panels (Fig. 4): confusion matrix panel (a), calculation panel (b), and fairness metrics panel (c). Panel (a) displays the predictions grouped by the protected attribute, which is consistent with the output panel of **C3**. Panel (c) visualizes the four fairness metrics calculated from the confusion matrices. For each metric, solid vertical lines indicate its current value and dotted vertical lines represent its fair/baseline value. Values that are biased against the unprivileged group are shown in gray. The distance between the current and the baseline value indicates the bias level of the model predictions. When users select a fairness metric, its mathematical formula will be displayed in the calculation panel (b). The example in Fig. 4 calculates the SPD metric, which measures the disparity in the rate of positive outcomes between female and male groups.

Under the *static visualization* setting, users can view only panels (a), (b), and (c). Under the *interactive visualization* setting, users can modify the values within the confusion matrices, and the fairness metrics get updated accordingly. As shown in Fig. 4, when users update the number of True Positive points in the female group (indicated by a black arrow) from 24 to 40, the values of all four fairness metrics move closer to the baseline. In particular, the SPD value increases from −0.14 to −0.10.

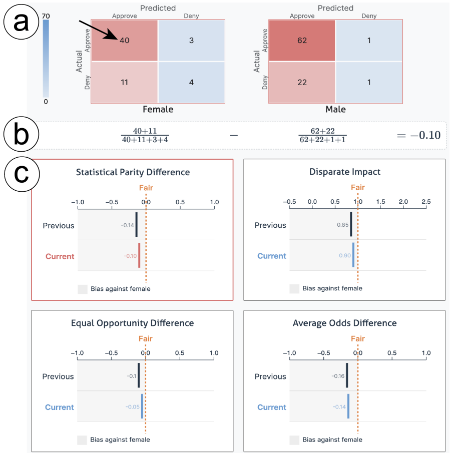


Figure 4: **C4** explains/compares fairness metrics. Updating values in confusion matrices will update the fairness metrics accordingly.

Design considerations. An alternative design we explored for displaying the metric values is the bar charts used in the demo of AI Fairness 360 [5], where each bar originates from 0. However, we rejected this approach due to its potential for visual bias. For example, the bar length fails to reflect fairness degrees for the *disparate impact* metric, which has a baseline value of 1. Besides, we placed each metric in a separate panel and arranged different metrics in a juxtaposition [17] for effective comparison (**R3**).

4.4 Bias Mitigation

Finally, we introduce various methods suggested by P1 and P2 to mitigate the biases observed in model predictions (**T3**). We support three pre-processing methods [7, 25, 55], two in-processing methods [27, 56], and one post-processing method [26].

Given these diverse debiasing methods, it is impractical and unscalable to design visualizations for each method. Instead, we design a uniform visualization for each category, and we visualize only the input and output of a debiasing method and hide their inner workings from nontechnical users (addressing **R4**). This design choice also increases the scalability of our interactive components.

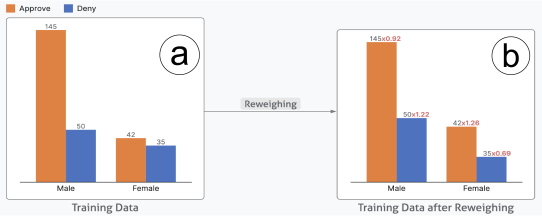


Figure 5: **C5** visualizes pre-processing debiasing methods.

Pre-processing methods aim to mitigate biases by modifying the input data, such as assigning appropriate weights to individuals, or transforming the input data to reduce discrimination. We present **C5** for users to explore the pre-processing methods. As shown in Fig. 5, its interface consists of two panels: the input (a) and the output panels (b) visualize the training data grouped by the protected attribute, before and after applying a pre-processing method. In this example, we use the reweighing method, which assigns different weights to individuals from different subgroups such that all subgroups have a similar proportion of positive data points. When users click the "Reweighing" button, the modified data together with the weights applied to each subgroup will be displayed in the panel (b).

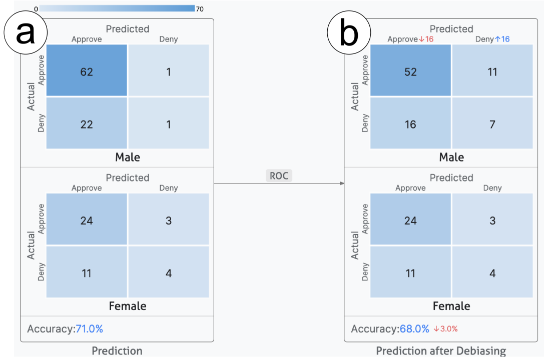


Figure 6: **C6** visualizes post-processing debiasing methods.

In-processing methods use ML models that take fairness into account, typically by adding a fairness term when optimizing the model. We utilize **C3** to visualize these methods by changing the model in the backend. See the supplementary material for details.

Post-processing methods improve the fairness of predictions by directly changing the predicted results, for example, by modifying the decision threshold. We propose **C6** (Fig. 6) to visualize post-processing methods. Similar to **C5**, **C6** consists of an input panel (a) and an output panel (b). Model predictions are grouped by the protected attribute before and after applying the debiasing method. In this example, we apply the reject option classification (ROC) method, and consequently the outcome for 16 males changes from being approved to being denied. However, the accuracy for the debiased prediction decreases by 3%.

Furthermore, these debiasing components can be integrated with **C3** to display changes in fairness metrics before and after debiasing. Simultaneously, the output panel of **C4** and the prediction panel of **C6** showcase changes in accuracy in the debiased model predictions. Users can gain insights into fairness-accuracy trade-offs by comparing these changes (**T3**, **R2**).

Design considerations. We drew upon prior design choices for visualizing bias mitigation, such as the grouped bar charts in **C5** and confusion matrices in **C6**. These choices not only ensures consistency that reduces the interpretation burden (**R1**), but also makes it clear where the technique is applied (**T3**). In addition, we adopted visual strategies of juxtaposition and explicit encodings [17] to enhance the comparison of debiasing effects (**R2**). For instance, in **C5** and **C6**, we positioned panels showing outcomes before and after debiasing side by side, and used arrows in the prediction panel to indicate changes in model performance.

5 In-Class Qualitative Evaluation

We integrated the six types of interactive components into an educational module for a six-week summer 2022 course (1.5 credits) taught by professors P1 and P2 in the School of Business, *Fair Algorithms for Business*. We examined the effectiveness of teaching fairness in AI using visualization. Course topics included *Machine Learning and Fairness* (**M1**), *Fairness Metrics and Sources of Bias* (**M2**), *Debiasing Methods* (**M3**), *Bias in Texts and Images* (**M4**), and *Final Project* (**M5**). We embedded interactive components in web documents with texts and images provided by the instructors as the course material. We used examples described in Sec. 4 for the first three topics (**M1**-**M3**). and a sample of the Bank Marketing dataset [36] for the assignment (**M5**). Six graduate and undergraduate students from the School of Business attended this elective course. By manipulating these interactive components, the students explored the trade-off between fairness metrics and accuracy, and learned about pre-processing, in-processing, and post-processing debiasing methods. At the conclusion of the course, we collected feedback from the instructors and students, through interviews and surveys, respectively, to understand the effectiveness of using visualization in learning about fairness in AI.

Instructor Feedback. Instructors P1 and P2 served dual roles in this project. As part of a project team, they went through an iterative process in providing design requirements for the visualization components. As instructors, they provided feedback on the effectiveness of these modules during deployment based on instructor-student interactions in class. P1 and P2 commented that the tool was able to handle web traffic from multiple students simultaneously. The material was organized in a proper sequence that went from understanding a ML model, to detecting bias in the model prediction, and then to applying debiasing methods. In particular, the interactive component **C4** was highly valued by the instructors as it allowed students to alter input and receive output on the fly that could be assessed for accuracy and bias. This feature increased student engagement since it gave them the freedom to try different values and understand the consequences. On the other hand, instructors noted that certain concepts were missing from the educational module, which may be necessary for students to comprehend the debiasing methods, such as mutual information and gradient descent. They suggested designing additional visualizations to explain these concepts. To avoid bias in instructor feedback, it would be useful to obtain feedback from independent instructors using our educational modules, which is intended for future work.

Student Survey. Three students (S1-S3) anonymously shared their user experiences and improvement suggestions for each component, through free-text responses in an online survey. Overall, they found the interactive components to be beneficial for their learning. S2 commented that "(the interactive components) help a lot. Intuitive visual is a really very great method for understanding." Additionally, our design allowed students to perform what-if analyses by altering the model input, which received positive feedback from S1: "It is a good simulator, you can see that things change as you make adjustments." In particular, the fairness metric visualization (**C4**) received high praise from S1 and S2, and it was considered to be "very intuitive" and "highly successful". On the other hand, S1 initially found the use of the confusion matrix in **C3** and **C6** to be confusing, but "once they are placed in the larger simulator in the final project, they make more sense." S1 expressed a desire for more details on the logistic regression: "the math is being done behind the scenes, and so I did not understand what was happening there." S3 suggested to "provide code separately if we want a more in-depth look at the model."

6 User Study Design

A user study was conducted outside the classroom with a large sample size of 413 participants. Using three different visualization formats, we examined whether adding visualization leads to an increased understanding of fairness concepts.

Past studies have suggested mixed results about the effect of visualization in education [39, 47]. In particular, Hundhausen et al. [23] found that *how* students use Algorithm Visualization (AV) technology has a greater impact on the effectiveness than *what* AV shows them. Grissom et al. [18] identified a taxonomy encompassing six forms of learner engagement with visualization. Among them, "no viewing" refers to instruction without any form of accompanying AV; "viewing" refers to a passive form of engagement where users watch different visual representations of the algorithm being studied; and "changing" entails modifying the visualization.

Utilizing past findings, our user study followed a one-factor, three-level, between-participant experimental design. This design uses three visualization formats: texts with images of confusion matrices, static visualization, and interactive visualization. These formats fall under the categories of "no viewing", "viewing", and "changing", respectively, according to the taxonomy [38]. In particular, interactive visualization aligns well with the "changing" category as it allows users to change input data, thus modifying the visualization.

6.1 Experimental Conditions

We utilized the fairness metric of disparate impact as the learning objective in the user study due to its simplicity and popularity. Moreover, our in-class evaluation indicated that the fairness metric view was especially favored by instructors and students when demonstrated in class. Although how a visualization is used can be more important than what is shown [23], we investigated whether a favorably-reviewed visualization could benefit learning.

We designed the learning material of the user study together with P1 and P2 following a similar workflow in the course. We first introduced the background on bias in ML and the confusion matrix. We then explained the concept of disparate impact and its computation with a loan example. The three experimental conditions provided the same information on fairness, with only the visualization formats being varied:

- **Texts and Images** (*TextImg*) condition utilizes texts and images of confusion matrices to explain the computation.
- **Static Visualization** (*StaticVis*) condition uses the static **C4** to explain computations, with the fairness metrics panel displaying only disparate impact (c.f., Fig. 4).
- **Interactive Visualization** (*InterVis*) condition uses the interactive **C4** and allows participants to change the values of confusion matrices to observe the change of metric values.

6.2 Participants

413 student participants from the School of Business participant pool took part in the study for partial course credit. They were randomly assigned to one of the three between-participant conditions: *TextImg*, *StaticVis*, or *InterVis*. Of all the participants, we excluded in our analysis 31 participants who reported completing a major or minor in Computer Science, since we focused on how visualization affected learning for nontechnical users. The remaining 382 participants consisted of 378 undergraduates, 3 master's students, and 1 doctoral student (based on participants' self-reports). The sample sizes for *TextImg*, *StaticVis*, and *InterVis* are 134, 126, and 122, respectively.

6.3 Procedure

The user study was deployed on a customized survey website that could render the visualization as well as collect user data such as response time. Participants were initially shown a welcome screen followed by an informed consent document. The study began after obtaining the informed consent of the participants.

Participants first completed a *background questionnaire* regarding their degree (bachelor's, master's, and doctoral) and major (economics, social sciences, humanities and arts, policy, natural science, and engineering). Then they took a *pre-test* that evaluated their ability to understand whether the decision of a ML model was biased or not, and their awareness of the concept of disparate impact. The average accuracy across all participants was 16.76%, indicating limited exposure to ML and fairness concepts.

Participants were then provided *reading material* that introduced a loan approval case followed by the fairness metric of disparate impact. Specifically, we explained the interpretation of the visualization and/or the utilization of interaction in the reading material of *StaticVis* and *InterVis* conditions. Participants were randomly assigned to one of the three visualization conditions that vary by learning formats. They were allowed as much time as they needed to review the learning material.

Subsequently, participants were directed to questions in a *post-test*, which included two sets of questions designed based on the Bloom's taxonomy [28]. Bloom's taxonomy structures a learner's depth of understanding into six levels of objectives (remember, understand, apply, analyze, evaluate, and create), and the two sets of questions fall into the first two levels *remember* and *understand*, referred to as *recall* and *comprehension* in our study. The first set of *recall questions* was the same as those in the pre-test. The aim was to examine whether different visualization conditions improved their understanding of bias in an ML model. The second set of *comprehension questions* was designed to further examine how well the participants understood the concepts under different visualization conditions. The questions were more nuanced and in-depth. For the *InterVis* group, interaction operations were disabled when answering comprehension questions because participants could simply enter the values and the interactive components would provide the answers.

Finally, to examine how participants' experience differed across the three conditions, we asked them three *impression questions* about their impression of the learning material on a 5-point Likert scale. They were also provided with an open-ended question to collect their subjective feedback about the learning experience. The study concluded with a validated *visual learning style questionnaire* that consists of 22 questions on a 6-point Likert scale [1]. All questions can be found in the supplemental material.

7 User Study Results

Via the backend of the survey website, the user study collected participants' responses to pre-test and post-test questions, time spent on each webpage, whether participants referred to the learning material while answering the post-test questions, open-ended responses, and responses to the impression questions and learning style questions.

The user study utilized a one-factor, three-level, between-participant experimental design. We used the analysis of variance (ANOVA) and linear regression to analyze the data. The three conditions—*TextImg*, *StaticVis*, and *InterVis*—served as the independent variables. The dependent variables included accuracy in the recall test, accuracy in the comprehension test, and responses to the impression questions.

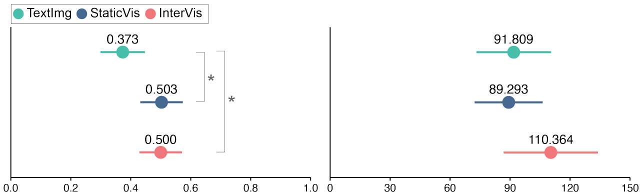
We first examined the influence of the three conditions on each dependent variable by performing a one-way ANOVA, and then conducted the pairwise comparison using Tukey's HSD Test if significant differences were found. Second, we tested whether the time spent on the learning material *moderated* the influence of the three conditions and each dependent variable using regression analysis. Third, we tested how the visual learning style *moderated* the relationship between the three conditions and each independent variable. The visual learning score was calculated based on 11 visual-related questions in the visual learning style questionnaire. The reported statistical results were validated using StatCheck [21]. For qualitative analysis, we summarized the subjective feedback of users in Sec. 7.4. Notably, we found that only 16 of 122 participants in the *InterVis* group used interaction, which is similar to a phenomenon observed by Mosca et al. [37]. We then analyzed their subjective feedback in detail. We also analyzed the time that participants spent answering questions and revisiting material without significant findings (see the supplemental material for completeness).

7.1 Recall Questions

Accuracy gain. The recall questions were designed to measure whether reviewing the learning material improved participants' accuracy in answering the same set of pre-test questions. Accuracy gain—the difference in accuracy before versus after reviewing the material—was used as the dependent variable; see Fig. 7 (a). A one-way ANOVA test revealed a significant difference across the three conditions, with an F-value $F(2,379) = 4.134$ and a p-value $p = 0.017$.

To further examine the difference, we used the Tukey's HSD test for multiple comparisons and found that the *StaticVis* condition had a significantly higher mean accuracy gain than the *TextImg* condition, with $p = 0.033$, and a 95% confidence interval $CI = [0.008, 0.251]$. The *InterVis* condition had a significantly higher mean accuracy gain than *TextImg* with $p = 0.040$ and a $95\% CI =$

$[0.004, 0.249]$. However, there was no significant difference in the mean accuracy gain between the *StaticVis* and *InterVis* conditions ($p = 0.900$, $95\%CI = [-0.122, 0.127]$).



(a) Accuracy Gain (b) Reading Time (sec)

Figure 7: (a) Accuracy gain of recall test and (b) reading time before and during recall test. Error bars show 95% confidence intervals. * indicates a significant difference between conditions ($p < 0.05$).

Influence of reading time. We then investigated whether the *reading time* influenced the relationship between the three conditions and accuracy gain. The reading time was calculated as the time participants spent in reading the learning material, before and during the recall test. We noticed that several participants had a very long reading time, possibly due to their temporary absence during the study. To reduce data bias, we removed seven significant outliers using the Mahalanobis distance ($p < 0.001$). The results are shown in Fig. 7 (b). While *InterVis* participants showed the longest average reading time, the one-way ANOVA result ($F(2,372) = 1.259$, $p = 0.285$) indicates no significant difference across three conditions.

Next, we tested the influence of reading time on the accuracy gain across the three conditions. We ran a linear regression using *TextImg* and *StaticVis* as the reference condition, respectively, to enable all pairwise comparisons. The analysis revealed a significant interaction between *TextImg* × *StaticVis* and reading time ($p = 0.0498$), and between *InterVis* × *StaticVis* and reading time ($p = 0.014$). This result implies that *StaticVis* achieved a significantly higher accuracy gain than *TextImg* and *InterVis* when participants spent longer reading time. See the supplementary material for additional details.

Influence of visual learning ability. Finally, we examined whether visual learning ability would affect the relationship between three conditions and accuracy gain by employing linear regression (twice) with reference conditions being *TextImg* and *StaticVis*, respectively. The p-values of condition pairs were *TextImg* × *StaticVis*: 0.282; *TextImg* × *InterVis*: 0.976; *InterVis* × *StaticVis*: 0.304. In short, our result revealed no significant interactions between visual learning ability and conditions.

Highlighted results. Overall, our findings suggested that our static and interactive visualization design could significantly enhance accuracy gain on recall questions compared to texts and images. We also found that reading time did not differ significantly across three conditions, but static visualization demonstrated the most significant improvement in accuracy gain as reading time increased. Furthermore, visual learning ability did not significantly impact the relationship between the three conditions and accuracy gain.

7.2 Comprehension Questions

The comprehension questions were intended to evaluate participants' comprehension of the learning material, which required a deeper understanding than the recall questions. We calculated the accuracy of the comprehension test as the performance metric. A one-way ANOVA indicated that there was no significant difference in the mean accuracy across the three conditions ($F(2,379) = 0.660$, $p = 0.518$). We also analyzed the influence of reading time and visual learning ability on accuracy but found no significant results. Details of the analysis are provided in the supplementary material.

7.3 Impression Questions

We collected participants' ratings of the learning material based on three impression questions using a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). The questions focused on whether the visualizations were effective (Q1), engaging (Q2), and recommended (Q3). We provide a detailed analysis in the supplementary material. In short, we observed no significant differences in participants' ratings of each impression question among three conditions, and reading time did not significantly impact the relationship between conditions and ratings.

Highlighted results. However, regarding the recommendation question (Q3), we noted a significant interaction between visual learning ability and the *StaticVis* or *InterVis* condition ($p = 0.027$). This finding suggests that although visual learning ability did not significantly impact the accuracy in the post-tests as demonstrated in Sec. 7.1 and Sec. 7.2, participants with higher (resp., lower) visual learning ability are more likely to recommend learning material with interactive visualization (resp., static visualization).

7.4 Subjective Feedback

All participants provided subjective feedback about the learning material by answering the following open-ended question:

> *How has the information shown on the screen changed your understanding of the topic? In the space provided please let us know your thoughts in as much detail as possible. Please list the new concepts that you learned and which aspects of the learning environment helped you understand the concepts the most.*

To analyze the subjective feedback, we conducted a semantic analysis using the open-source library pysentimiento [41] to categorize each response as positive, neutral, or negative. Positive (resp., negative) responses describe how the learning material aids (resp., obstructs) their understanding of the concepts, whereas neutral responses either simply state what they have learned from the material or mention both positive and negative aspects of it. Table 1 shows the percentages of these three categories under each condition. The three conditions share similar percentages of negative responses, whereas the *StaticVis* condition exhibits a smaller amount of positive feedback than the other two conditions. Guided by the sentiment analysis, we carefully examined all responses and summarized key factors that either promote or impede the learning process.

Condition	Positive	Neutral	Negative
TextImg	32.8%	29.1%	38.1%
StaticVis	23.0%	38.9%	38.1%
InterVis	32.8%	33.6%	33.6%

Table 1: Semantic analysis of subjective feedback.

Visualization improves concept understanding. Under the *StaticVis* and *InterVis* conditions, we obtained 16 and 23 positive comments, respectively, in which participants explicitly stated that the visualization helped them understand the material better. One *StaticVis* participant said, "The graphics helped me learn the most." Another *InterVis* participant said, "I really enjoyed the visuals as I am a visual learner, and it helped me grasp the concepts being displayed." Even though we showed only a simple image of the confusion matrix under the *TextImg* condition, 18 participants expressed a preference for it. One participant noted, "I thought that the simplified relationships relating to the quadrants of the confusion matrix were useful for identifying what each one meant." However, we observed three responses under each condition that considered the visualizations confusing or unclear. Overall, visualization could serve as an effective way to improve the learning experience.

Interaction facilitates metric comprehension. We further analyzed feedback from 16 *InterVis* participants who interacted with the visualization based on our log data. Among them, seven participants explicitly expressed strong preferences toward the interaction design, which helped them better understand the mathematical definition of disparate impact. One participant noted, "What helped me learn the most had to be the interactable table which let me change the numbers around, and it really helped me understand which number's increasing or decreasing affect the disparate impact." Another participant commented, "The graphs helped, and I really liked the idea of the interactive graph that shows you how the calculation changes as you change the inputs." However, we also found a participant who disliked the visualization due to an unmatched learning style, who stated, "It was difficult since my brain does not do well with graphs." Whereas the use of interaction cannot be ensured, those who engage with visual interaction appear to benefit from the interactive features.

Factors hindering learning experience. We identified two factors from negative and neutral responses that hindered participants' learning experience. First, the lack of knowledge in ML or data science made it challenging to comprehend the material (e.g., mathematical formulas or technical concepts), reported by 7, 7, and 9 participants from the three respective conditions. A *TextImg* participant said, "The paragraphs were bulky and technical, which was not an effective mode of communication for someone with no previous knowledge of the topic." Second, the presentation of excessive information made the material less appealing to participants, reported by 11, 9, and 5 participants from the three respective conditions. An *InterVis* participant stated, "I found it challenging to want to read the information. There was a lot of information in large paragraphs." Although explaining ML concepts like model predictions is inevitable before fairness metrics, the introduced cognitive overload may degrade the use of visualization. Therefore, further experiments are needed to investigate how the participants' cognitive overload may influence the use of visualization.

8 Limitations and Future Work

Supporting diverse fairness tasks. In addition to *group fairness* in classification, it would be useful to extend our visualization to incorporate *individual fairness* and *intersectional fairness* into other ML tasks (e.g., regression and clustering).

Adjusting level-of-details of ML models. We visualized a simple logistic regression model and treated bias mitigation models as black boxes, to avoid overwhelming the nontechnical users. However, some students expressed a desire to dive deeper into the inner workings of ML models (Sec. 5). Instructor P1 further expressed that "even if we use a simple linear regression example, a student should be able to understand optimization." It would be useful to incorporate concepts that can be adapted for students requiring finer levels of details, such as loss functions, gradient descent, and mutual information. Following recent advances on explainable AI [4, 31], it would be interesting to introduce advanced models (e.g., neural networks and large language models) to nontechnical students.

Improving user study. Our user study also has limitations. First, as the user study was conducted outside the classroom without time and location restrictions, the uncontrolled study environment may lead to biased data collection, caused by distracting environments, users' temporary absences, and varied study times. Second, we observed that the comprehension question accuracy was low across three the conditions, which may have been caused by the excess of text presented in the learning material that might have overwhelmed participants. Such overload would lower learning interests and reduce the willingness to interpret the visualization and interactions, consequently affecting learning performances. Whereas previous research [48] has shown the benefits of combining text with visualization, our findings suggest that excessive text may lead to cognitive overload, a phenomenon also noted by Lee et al. [32]. It would be interesting to investigate how the ratio between visualization and text could influence user performance. Third, our study revealed that both static and interactive visualization improved student's performance in recall questions within a short time frame. It would be useful to assess an educational module's potential in enhancing long-term knowledge retention by conducting a subsequent study after a specified period.

Another limitation in our study is the limited use of interaction (13%), in line with prior research by Mosca et al. [37]. Although we analyzed the relevant feedback in detail, a quantitative assessment of the value of interactions requires further investigation, where participants are exclusively exposed to the *InterVis* condition and explicitly directed to utilize interactive features. The reason behind the low usage of the interaction cannot be inferred from the users' subjective feedback and remains unclear. More research is required to identify influencing factors, such as user omissions, lack of motivation, technical barriers, or inappropriate interaction design. Finally, our study focused only on visualizing the fairness metrics and a single type of interaction that modifies the input data. Future studies are needed to investigate the influence of different types of visualization and interaction on learning performance.

Acknowledgments

This project was partially funded by a seed grant from the University of Utah and the Utah Board of Higher Education's Deep Technology Initiative grant on "Bringing Fairness in AI to the Forefront of Education".

References

[1] A. Abdollahimohammad and R. Ja'afar. Learning Style Scales: A valid and reliable questionnaire. *Journal of educational evaluation for health professions*, 11, 2014.

[2] Y. Ahn and Y.-R. Lin. FairSight: Visual analytics for fairness in decision making. *IEEE Transactions on Visualization and Computer Graphics*, 26:1086–1095, 1 2019.

[3] I. Ajunwa and D. Greene. Platforms at Work: Automated hiring platforms and other new intermediaries in the organization of work. In *Work and labor in the digital age*, volume 33, pages 61–91. Emerald Publishing Limited, 2019.

[4] G. Alicioglu and B. Sun. A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics*, 102:502–520, 2022.

[5] R. Bellamy and others. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63:4–1, 2019.

[6] Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 46–56, 2019.

[7] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

[8] D. Dua and C. Graff. UCI machine learning repository, 2017.

[9] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[10] E. E. Firat and R. S. Laramee. Towards a survey of interactive visualization for education. *EG UK Computer Graphics & Visual Computing, Eurographics Proceedings*, pages 91–101, 2018.

[11] E. Fouh, M. Akbar, and C. A. Shaffer. The role of visualization in computer science education. *Computers in the Schools*, 29(1-2):95–117, 2012.

[12] E. Fouh, V. Karavirta, D. A. Breakiron, S. Hamouda, S. Hall, T. L. Naps, and C. A. Shaffer. Design and architecture of an interactive etextbook–The OpenDSA system. *Science of computer programming*, 88:22–40, 2014.

[13] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143, 2021.

[14] P. Garg, J. Villasenor, and V. Foggo. Fairness Metrics: A comparative analysis. In *IEEE International Conference on Big Data (Big Data)*, pages 3662–3666, 2020.

[15] B. Ghai, M. N. Hoque, and K. Mueller. WordBias: An interactive visual tool for discovering intersectional biases encoded in word embeddings. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[16] B. Ghai and K. Mueller. D-BIAS: A causality-based human-in-the-loop system for tackling algorithmic bias. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):473–482, 2022.

[17] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[18] S. Grissom, M. F. McNally, and T. Naps. Algorithm visualization in CS education: comparing levels of student engagement. In *Proceedings of the ACM Symposium on Software Visualization*, pages 87–94, 2003.

[19] B. J. Grosz, D. G. Grant, K. Vredenburgh, J. Behrends, L. Hu, A. Simmons, and J. Waldo. Embedded EthiCS: integrating ethics across cs education. *Communications of the ACM*, 62(8):54–61, 2019.

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

[21] S. Haroz. StatCheck simple edition web application. `http://statcheck.steveharoz.com`, 2021.

[22] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 203–212, 2010.

[23] C. D. Hundhausen, S. A. Douglas, and J. T. Stasko. A meta-study of algorithm visualization effectiveness. *Journal of Visual Languages and Computing*, 13(3):259–290, 2002.

[24] Z. Jin, X. Wang, F. Cheng, C. Sun, Q. Liu, and H. Qu. Shortcutlens: A visual analytics approach for exploring shortcuts in natural language understanding dataset. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[25] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

[26] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929, 2012.

[27] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 35–50. Springer, 2012.

[28] D. R. Krathwohl. A revision of Bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

[29] B. C. Kwon, U. Kartoun, S. Khurshid, M. Yurochkin, S. Maity, D. G. Brockman, A. V. Khera, P. T. Ellinor, S. A. Lubitz, and K. Ng. RMExplorer: A visual analytics approach to explore the performance and the fairness of disease risk models on population subgroups. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pages 50–54, 2022.

[30] B. C. Kwon, J. Lee, C. Chung, N. Lee, H.-J. Choi, and J. Choo. DASH: Visual analytics for debiasing image classification via user-driven synthetic data augmentation. In M. Agus, W. Aigner, and T. Hoellt, editors, *EuroVis 2022 - Short Papers*. The Eurographics Association, 2022.

[31] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, R. Capobianco, E. Bertini, R. Giot, and M. Angelini. State of the art of visual analytics for explainable deep learning. *Computer Graphics Forum*, 42(1):319–355, 2023.

[32] M. S. A. Lee and J. Singh. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[33] S. Liu and L. N. Vicente. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science*, 19(3):513–537, 2022.

[34] A. Mashhadi, A. Zolyomi, and J. Quedado. A case study of integrating fairness visualization tools in machine learning education. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.

[35] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[36] S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[37] A. Mosca, A. Ottley, and R. Chang. Does interaction improve Bayesian reasoning with visualization? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[38] T. L. Naps, G. Rößling, V. Almstrum, W. Dann, R. Fleischer, C. Hundhausen, A. Korhonen, L. Malmi, M. McNally, S. Rodger, et al. Exploring the role of visualization and engagement in computer science education. *ACM SIGCSE Bulletin*, 35(2):131–152, 2002.

[39] P. R. Osztian, Z. Kátai, and E. Osztian. Algorithm visualization environments: Degree of interactivity as an influence on student-learning. In *2020 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE, 2020.

[40] S. Passi and M. Vorvoreanu. Overreliance on AI literature review. Microsoft Research, 2022.

[41] J. M. Pérez, J. C. Giudici, and F. Luque. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks, 2021.

[42] D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[43] A. Rathore, S. Dev, J. M. Phillips, V. Srikumar, Y. Zheng, C.-C. M. Yeh, J. Wang, W. Zhang, and B. Wang. VERB: Visualizing and interpreting bias mitigation techniques geometrically for word representations. *ACM Transactions on Interactive Intelligent Systems*, 2022.

[44] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

[45] P. Saraiya, C. A. Shaffer, D. S. McCrickard, and C. North. Effective features of algorithm visualizations. In *Proceedings of the 35th SIGCSE technical symposium on Computer Science Education*, pages 382–386, 2004.

[46] D. Schweitzer and W. Brown. Interactive visualization for the active learning classroom. In *Proceedings of the 38th SIGCSE technical symposium on Computer science education*, pages 208–212, 2007.

[47] S. Šimoňák. Increasing the engagement level in algorithms and data structures course by driving algorithm visualizations. *Informatica*, 44(3), 2020.

[48] J. Urquiza-Fuentes and J. Á. Velázquez-Iturbide. A survey of successful evaluations of program visualization and algorithm animation systems. *ACM Transactions on Computing Education (TOCE)*, 9(2):1–21, 2009.

[49] Q. Wang, Z. Xu, Z. Chen, Y. Wang, S. Liu, and H. Qu. Visual analysis of discrimination in machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1470–1480, 2020.

[50] M. Q. Wang Baldonado, A. Woodruff, and A. Kuchinsky. Guidelines for using multiple views in information visualization. In *Proceedings of the working conference on Advanced visual interfaces*, pages 110–119, 2000.

[51] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The What-If Tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2019.

[52] B. Wilson, J. Hoffman, and J. Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019.

[53] T. Xie, Y. Ma, J. Kang, H. Tong, and R. Maciejewski. FairRankVis: A visual analytics framework for exploring algorithmic fairness in graph mining models. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):368–377, 2021.

[54] J. N. Yan, Z. Gu, H. Lin, and J. M. Rzeszotarski. Silva: Interactively assessing machine learning fairness using causality. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[55] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333, 2013.

[56] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[57] X. Zhang, J. P. Ono, H. Song, L. Gou, K.-L. Ma, and L. Ren. SliceTeller: A data slice-driven approach for machine learning model validation. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):842–852, 2022.