JOURNAL OF LEARNING ANALYTICS

Volume 12(2), 125–139. https://doi.org/10.18608/jla.2025.8543

Unveiling Accuracy-Fairness Trade-Offs: Investigating Machine Learning Models in Student Performance Prediction

Raymond A. Opoku[1], Bo Pei[2] and Wanli Xing[3]

Abstract

While high-accuracy machine learning (ML) models for predicting student learning performance have been widely explored, their deployment in real educational settings can lead to unintended harm if the predictions are biased. This study systematically examines the trade-offs between prediction accuracy and fairness in ML models trained on the widely used Open University Learning Analytics Dataset (OULAD). We evaluated the relationship between model accuracy and fairness across various student demographic subgroups and investigated the extent to which fairness can be improved without significantly sacrificing accuracy. Our analysis revealed that standard ML models often exhibit bias; however, applying bias mitigation techniques can reduce these disparities while maintaining acceptable accuracy. Our findings emphasize the importance of auditing ML models for fairness to ensure that predictive insights are equitable across diverse student populations. We also discuss implications for best practices and challenges in achieving fair ML models for student performance prediction.

Notes for Practice

- Regular audits of machine learning models are crucial to identify biases and ensure equitable outcomes across diverse student populations.
- Practitioners should proactively apply bias mitigation techniques (e.g., reweighting, suppression, disparate impact remover) during model training to balance accuracy and fairness.
- Transparency in data preprocessing steps, such as aggregating student outcomes, is essential, since these decisions can significantly influence model bias and predictive accuracy.
- Integrating fairness-aware practices enhances trust and promotes the ethical use of predictive analytics in educational environments.

Keywords: Accuracy-fairness trade-offs, algorithmic bias, machine learning, virtual learning environments (VLEs), bias mitigation

Submitted: 30/06/2024 — **Accepted:** 21/05/2025 — **Published:** 31/07/2025

Corresponding author [1]Email: raymond.opoku@ufl.edu Address: School of Teaching & Learning, College of Education, University of Florida, 1221 SW 5th Ave, Gainesville, FL 32601, USA. ORCID iD: https://orcid.org/0000-0001-9892-0152
[2]Email: bpei@usf.edu Address: Department of Educational and Psychological Studies, College of Education, University of South Florida, 4110 USF Apple Dr., Tampa, FL 33620, USA. ORCID iD: https://orcid.org/0000-0002-6328-6929
[3]Email: wanli.xing@coe.ufl.edu Address: School of Teaching & Learning, College of Education, University of Florida, 1221 SW 5th Ave, Gainesville, FL 32601, USA. ORCID iD: https://orcid.org/0000-0002-1446-889X

1. Introduction

Virtual learning environments (VLEs) have transformed modern education by providing students with online access to course materials, assessment, communication tools and learning analytics (Romero & Ventura, 2020; Fazil et al., 2024; Khoudi et al., 2025; Pei & Xing, 2022). VLEs record student learning processes in detail, including learning logs, discussion posts, engagement patterns and performance outcomes, which machine learning (ML) models leverages to predict academic success and identify at-risk students for timely interventions (Martinez et al., 2025; Johnston et al., 2024; Pei & Xing, 2022; Xing & Du, 2018). However, deploying ML models in VLEs poses a critical challenge in ensuring that predictions are both accurate and fair across demographic groups (Baker & Hawn, 2022; Chinta et al., 2024; Idowu, 2024; Kizilcec & Lee, 2022; Shin et al., 2022). While recent research has provided extensive practical insights into addressing algorithmic bias (. i.e. a systemic and unfair disparities in ML model outcomes that disproportionately affect certain demographic groups) (Dwork et. al., 2012)

in online learning environments (Bayer et al., 2021; Mehrabi et al., 2022; Raftopoulos et al., 2025; Song et al., 2024; Shin et al., 2022; Liu et al., 2024; Lallé et al., 2024; Wongvorachan et al., 2024), there remains a critical gap in how bias mitigation techniques are applied and evaluated in this context. Our study aims to address this gap by not only identifying bias and comparing models but also by conducting a systematic analysis of various bias mitigation techniques and their impact on both fairness and accuracy in VLE-based predictive models.

1.1. Background

Algorithmic biases may arise in VLE data due to factors such as unequal access to technology, differences in prior knowledge, or cultural and socioeconomic variables that impact student engagement (Chinta et al., 2024; Kizilcec & Lee, 2022; Leite et al., 2021; Li et al., 2024). If not proactively addressed, these biases can lead to ML models making predictions that systematically disadvantage students from historically marginalized groups (Gardner et al., 2019; Li et al., 2024). Such algorithmic discrimination in educational analytics risks widening the existing achievement gaps (Zhang et al., 2023) and poses significant ethical challenges.

To tackle these issues, recent research has provided extensive practical insights into addressing the challenges relevant to the algorithmic bias in online learning environments (Deho et al., 2022; Lallé et al., 2024; Raftopoulos et al., 2025; Verger et al., 2024). For example, Shin et al. (2022) emphasized the importance of e-learning preparedness in ensuring fair learning analytics in higher education, underscoring that student readiness can influence how they interact with and are evaluated by digital learning systems. Similarly, Song et al. (2024) proposed a fair clustering approach to analyze self-regulated learning behaviours in VLEs, designed to account for diverse student backgrounds. In parallel, Liu et al. (2024) developed fair predictive models for an online math learning platform, emphasizing the balance between model accuracy with equitable outcomes across different student groups. Additionally, Lallé et al. (2024) investigated how demographic and contextual variables influence MOOC completion rates, underscoring fairness concerns in predictive modelling with large-scale online courses.

Despite these advances, the rapid evolution of ML models for student performance prediction has intensified concerns about the complex trade-offs between model accuracy and fairness (Deho et al., 2022; Dutta et al., 2020; Lallé et al., 2024; Liu & Vicente, 2022; Raftopoulos et al., 2025; Verger et al., 2024; Wang et al., 2021). While studies such as Deho et al. (2022) have emphasized the importance of fairness in educational ML applications, effectively managing these trade-offs in real-world settings remains a challenge. Fenu et al. (2022) further stressed the need for fairness in AI-driven education, yet the practical implications of navigating accuracy–fairness trade-offs are unexplored. Recent efforts by Peng and Jeang (2023) and Zhao et al. (2024) have advanced fairness-aware prediction models, but their focus on specific metrics, such as demographic parity and equal opportunity, leaves other fairness metrics like equalized odds and individual fairness ripe for further investigation.

Addressing these gaps requires robust bias mitigation techniques, several of which have been developed in recent years (Carey & Wu, 2023; Le Quy et al., 2023; Morik et al., 2020; Mehrabi et al., 2022; Verger et al., 2024). The reweighting (RW) technique, introduced by Kamiran and Calders (2012), ensures statistical independence between demographic attributes and outcomes by assigning weights to samples based on groups and label combinations. Specifically, the weight for each sample is proportional to the overall frequency of its label in the entire population and inversely proportional to the frequency of its label within its specific subgroup. The model is then trained on this adjusted dataset with reweighted samples. Particularly, for instance, this technique has been used to reduce bias in job interview video assessment (Köchling et al., 2021).

A preprocessing approach, Suppression (SUP) proposed by Kamiran and Calders (2009), eliminates bias by excluding key demographic attributes from the training dataset. This technique assumes that key demographic attributes are primary sources of bias. The process begins by removing the key demographic attribute from the dataset, followed by training a new ML model on this modified dataset. Researchers have also combined both SUP and RW methods to address bias in student dropout predictions (Kamiran & Calders, 2009; Kizilcec & Lee, 2022).

Similarly, the Disparate Impact Remover (DIR) method (Feldman et al., 2015), a preprocessing metric, adjusts feature distributions within the dataset to ensure equitable outcomes across groups. Here a fairness metric is applied to a dataset where S is the key demographic attribute, X represents the remaining attributes, and Y is the outcome variable. This technique adjusts the values of the features in X to ensure that all groups exhibit the same distribution for each variable, using percentile and quantile adjustments. The process begins with these adjustments, resulting in a restructured training dataset. This modified dataset is then used to train the ML model. Unlike some bias mitigation methods that directly alter the training data, DIR allows the model to be trained on the original data. After training, the model's outputs are modified to reduce disparities, and the effectiveness of the bias mitigation is assessed based on these adjusted outputs. The effectiveness of this approach in evaluating algorithmic bias has been demonstrated in various fields such as healthcare and education (Feldman et al., 2015).

Finally, the Calibrated Equalized Odds Post-processing (CPP) method (Hardt et al., 2016; Pleiss et al., 2017) adjusts predicted probabilities to equalize the false-positive rate (FPR) and false-negative rate (FNR) across privileged and unprivileged groups, offering a post-processing solution to enhance fairness. By modifying the model's score outputs across different subgroups, this technique aims to satisfy the equalized odds criterion. In VLEs, where the objective is to identify

students at risk of poor performance or failure, achieving equitable outcomes is crucial. Therefore, we prioritize recall over precision, focusing on equalizing the false-negative rates (FNRs) among subgroups, ensuring that at-risk students from all demographics are identified and supported fairly.

1.2. Our Contribution

Through extensive experiments, we evaluated the effectiveness of these techniques in improving fairness while preserving model accuracy. These findings have significant implications for the field of learning analytics. As predictive models become increasingly integrated into educational decision-making processes, ensuring their fairness is crucial for promoting equitable learning outcomes. Our study contributes to this goal by providing insights for evaluating and mitigating bias in VLE-based predictive models. Our main contributions are as follows:

1. A systematic study of accuracy and fairness for student performance prediction across demographic subgroups in a large-scale data environment.
2. Empirical evaluation of the effectiveness of various bias mitigation techniques for improving fairness.
3. A systematic investigation of accuracy–fairness trade-offs and providing practical implications for equitable learning practices in VLEs.
4. Providing comprehensive guidance and recommendations for practitioners and educators on how to improve their considerations while engaging with ML and AI applications in educational settings.

2. Methods

In this section, we first introduce the datasets utilized in our study and offer an initial analysis of their key features. Subsequently, we describe the ML algorithms selected for prediction tasks and discuss four approaches implemented to mitigate algorithmic bias. Further, we specify the evaluation metrics employed to assess both the predictive accuracy and fairness aspects of the models in our experimental setup. For this study we focus on four demographic populations namely gender, disability status, age group, and Index of Multiple Deprivation.

2.1. Dataset

The OULAD (Kuzilek et al., 2017) is an open-source benchmark dataset containing demographics, virtual learning activity, and performance outcomes for over 35,000 adult learners enrolled in online undergraduate and postgraduate courses at The Open University, a distance learning institute based in the United Kingdom. The dataset includes students typically aged from their early 20s to over 60 and covers seven courses. It provides detailed information on student demographics (e.g., gender, age, disability status); VLE activity (e.g., clicks on course materials); and performance outcomes (e.g., final grades). For our preprocessing steps, we aggregated the "Fail" and "Withdrawn" classes to represent the "Fail" outcome, while combining "Distinction" and "Pass" classes to represent the "Pass" outcome. This dichotomization aligns with our objective of predicting whether a student will successfully complete a course or not. We acknowledge that this aggregation may influence the bias in our model and will discuss this further in our limitations section.

2.1.1. Participants and Features

The dataset comprises various features, including student demographic information, engagement metrics (such as the number of clicks on learning resources), and performance indicators. These features are used to train ML models for predicting whether a student will pass or fail a course. A total of nine features from 25,690 students were included in our study, which includes binary, ordinary, and continuous variables. Table 1 summarizes the key numeric features used in this study, and Figures 1-4 displays the categorical features.

Table 1. Summary of Numeric Features in the OULAD Dataset

Feature	Mean	Standard Deviation	Minimum	Median	Maximum
Number of Previous Attempts	0.161	0.474	0	0	6
Studied Credits	77.76	39.06	30	60	630
Number of Assessments	7.704	4.525	1	7	28
Average Score	72.66	15.59	0	75.58	100
Total Clicks	1,757.83	2,065.04	1	1,052	28,615

Table 1 summarizes the key numeric features from the OULAD, which includes various attributes related to student engagement and performance. The features summarized in this table provide insights into the student academic behaviours and outcomes, and are as follows:

Number of Previous Attempts indicates the number of previous attempts a student made to pass a course. The mean value of 0.16, with a standard deviation of 0.47, suggests that most students did not make multiple attempts.

Studied Credits represents the total number of credits a student studied, with an average of 77.76 and a standard deviation of 39.06. The data shows a wide range of studied credits, with a minimum of 30 and a maximum of 630, indicating varying levels of student engagement over multiple years.

Number of Assessments captured the number of assessments a student has completed, with a mean of 7.70 and a standard deviation of 4.52. The number of assessments ranged from 1 to 28 and reflects the diversity in course requirements and student participation.

Average Score reflects the average score a student achieved across assessments, with a mean of 72.66 and a standard deviation of 15.59. Scores ranged from 0 to 100, with a median score of 75.58, which indicates a generally high levels of performance among students.

Total Clicks measures the total number of clicks a student made in the VLE, which served as a proxy for student engagement. The mean number of clicks was 1,757.83, with a substantial standard deviation of 2,065.04, indicating significant variability in student engagement. The number of clicks ranged from 1 to 28,615.

2.1.2. Building the Ground Truth: Student Performance Outcome

The following figures summarize key demographic attributes from the OULAD used in our analysis. This includes gender, disability status, age group, and Index of Multiple Deprivation (IMD), which is a socioeconomic indicator.

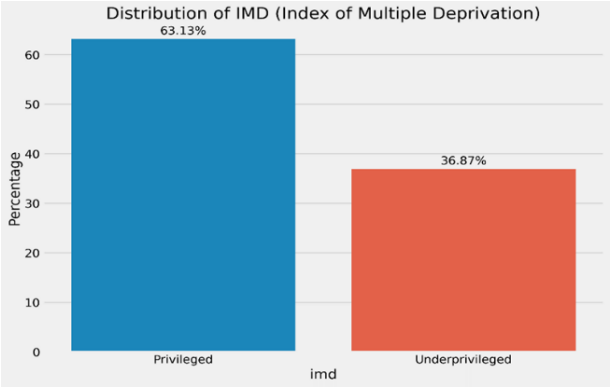

Figure 1. Distribution of IMD.

IMD: This feature categorizes students based on the IMD band. The dataset includes 10 unique IMD bands, with the 30–40% band being the most frequent, representing 2,889 students. These unique IMD bands were aggregated into two categories: 0%–59% indicating privileged (less deprived) and 60% or above indicating underprivileged (more deprived) students.

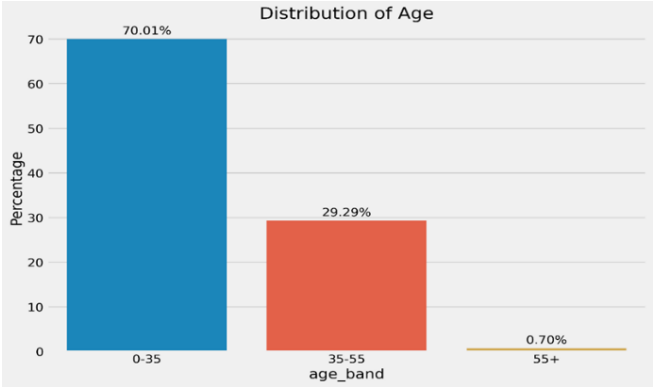

Figure 2. Distribution of Age.

Age_band: This feature classifies students into three age groups. The most common age group is 0–35, with 17,985 (70%) students falling into this category, indicating a younger demographic predominantly participating in the courses.

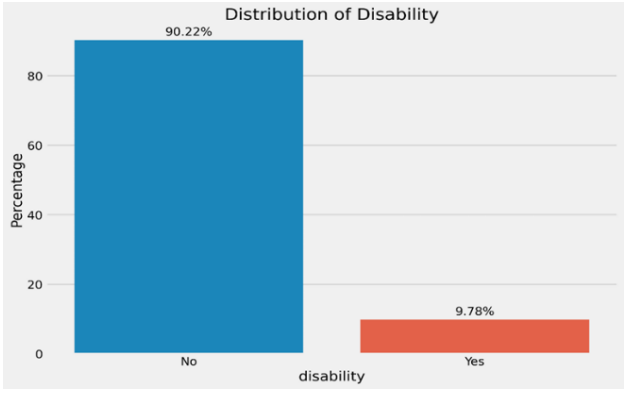

Figure 3. Distribution of Disability.

Disability: This binary feature indicates whether a student has a disability. Most students, 23,178 (90%), do not have a disability, highlighting the distribution of students with and without disabilities.

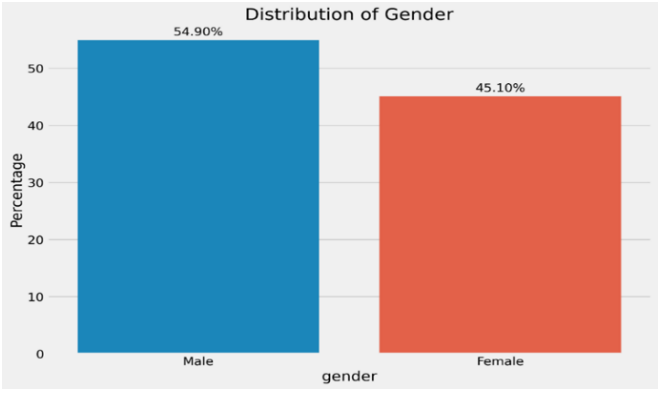

Figure 4. Distribution of Gender.

Gender: This feature records the gender of students. There are two unique values, with 11,604 (45%) students identified as female, providing a gender distribution in the dataset.

In this study, we define the ground truth for student performance based on the final course grades available in the OULAD dataset. The dataset provides information on student final results, which are categorized into four classes: Distinction, Pass, Fail, and Withdrawn. For our binary classification task, we combine the Distinction and Pass classes to represent the "Passed" outcome. The Fail and Withdrawn classes are combined to represent the "Failed" outcome. This dichotomization aligns with our objective of predicting whether a student will successfully complete a course or not.

It is important to acknowledge that using final course grades as the ground truth for student performance has its limitations. Grades may not fully capture the multidimensional aspects of learning, such as knowledge acquisition, skill development, and personal growth. Moreover, grading practices may vary across courses and instructors, introducing potential inconsistencies in the ground truth labels. However, given the available data in the OULAD dataset and the common use of grades as performance indicators in educational research, we consider final course grades as a reasonable proxy for student performance in this study. Future work could explore alternative or complementary measures of student success, such as self-reported learning outcomes or post-course assessments, to provide a more comprehensive evaluation of performance in VLEs.

2.2. Model Evaluation

In our study, we employed two commonly utilized ML algorithms to analyze the dataset: Logistic Regression (LR) and extreme gradient boosting (XGBoost). LR, a linear classification model, has been widely applied in various domains (Yu et al., 2011), while XGBoost, a powerful boosting technique, has gained popularity due to its exceptional performance (Chen & Guestrin, 2016). Rather than exploring the intricacies of these algorithms, our research focuses on their practical implementation and the assessment of their fairness metrics to provide insights on the applications ML models in educational settings.

To evaluate the performance of the LR and XGBoost models, we utilized k-fold cross-validation, a robust technique for model validation. The value of k was set to 10 for the datasets in our study. Additionally, we employed nested cross-validation, a method that allows for the optimization of hyperparameters while maintaining an unbiased assessment of model performance. This approach ensures that the models are fine-tuned to achieve the best possible results without overfitting to the specific characteristics of the dataset.

2.2.1. Performance Metrics

To assess the predictive capabilities of the machine learning models, we utilized two commonly employed evaluation metrics: the area under the receiver operating characteristic curve (AUC-ROC; Fawcett, 2006) and the balanced accuracy (BAcc; Wei & Dunbrack, 2013). The AUC-ROC serves as a measure of a model's discriminative power by quantifying the area beneath the curve generated by plotting true positive rates against false positive rates across a range of classification thresholds (Fawcett, 2006). This metric comprehensively assesses the model's ability to distinguish between different classes.

In contrast, BAcc is calculated by taking the average of sensitivity and specificity, making it particularly suitable for datasets with imbalanced class distributions, a common occurrence in educational data (Brodersen et al., 2010). By considering both the model's ability to accurately identify students who pass (sensitivity) and those who fail (specificity), BAcc offers a balanced evaluation of model performance, aligning with the objective of predicting student success or failure.

When investigating the trade-off between fairness and accuracy, we prioritize BAcc as it represents the model's performance at a specific decision threshold. This is particularly relevant in real-world scenarios where a fixed threshold is applied for classification purposes (Gardner et al., 2019). By focusing on BAcc, we can gain insights into the model's practical utility and its ability to make fair and accurate predictions at a given threshold (Fawcett, 2006).

We use both Balance Accuracy (BAcc) and AUC-ROC in our analysis for their complementary strengths. AUC-ROC provides an overall measure of model performance across all possible classification thresholds, BAcc focuses on model performance at a specific decision threshold, which is more relevant for assessing fairness in practical applications.

2.2.2. Fairness Metrics

In this study, we concentrated on the concept of group fairness (Bellamy et al., 2018), which necessitates that a model exhibits comparable performance across all subgroups defined by a particular attribute (Verma & Rubin, 2018). Specifically, we considered the equal opportunity criterion, which stipulates that a binary classifier can be deemed fair if the true-positive rates (TPR) are consistent across different groups (Dwork et al., 2011; Hardt et al., 2016).

We employed two fairness evaluation criteria to assess the fairness of the machine learning models in theoretical and contextual applications: the Equal Opportunity Difference (EOD; Hardt et al., 2016) and the Average Odds Difference (AOD; Bellamy et al., 2018), respectively. These metrics provide a comprehensive understanding of model fairness across different subgroups, namely gender, age, Index of Multiple Deprivation (IMD), and disability status.

(1) Equal Opportunity Difference (EOD)

The Equal Opportunity Difference (EOD) focuses on the true positive rates (TPR) across different subgroups defined by a key demographic attribute. A classifier is considered fair under the equal opportunity criterion if the TPRs are consistent across all subgroups (Hardt et al., 2016). EOD measures the maximum disparity in TPR between any two subgroups, with a value of 0 indicating perfect fairness.

Here, a value of 0 denotes perfect fairness (Hardt et al., 2016). This metric provides a practical and attainable measure of fairness by ensuring that the model's ability to correctly identify positive instances remains consistent across demographic subgroups. The mathematical definition of the metric is adapted from Hardt et al. (2016) as follows:

Consider a dataset $D = (X, Y, S)$, where X denotes the non-key demographic attributes, Y represents the target variable, and S is the key demographic attribute. Let $\hat{Y}$ be the predicted labels generated by a machine learning model. We define C as the set of possible class labels. In this study, we focused on binary classification tasks, i.e., $|C| = 2$, with $C = \{0, 1\}$ for a fail or pass. Let Ω_S denote the set of unique values that the key demographic attribute S can take. For instance, if the key demographic attribute is "race," then $\Omega_S = \{\text{Asian, Black, White, ...}\}$. A subgroup of the population, denoted as $D_{S=s}$, is defined as all the samples in the dataset D that have the same value s for the key demographic attribute S.

The True Positive Rate (TPR) for a specific subgroup $D_{S=s}$ is defined as:

$$\text{TPR}(D_{S=s}) = P(\hat{Y} = 1 \mid Y = 1, S = s)$$

This represents the probability of a positive prediction ($\hat{Y} = 1$) given that the true label is positive ($Y = 1$) and the key demographic attribute value is s.

The Equal Opportunity Difference (EOD) can then be defined as:

$$\text{EOD} = \max_{s \in \Omega_S} \text{TPR}(D_{S=s}) - \min_{s \in \Omega_S} \text{TPR}(D_{S=s})$$

Again, EOD measures the maximum disparity in TPR across all subgroups defined by the key demographic attribute S. A lower EOD value indicates better fairness, with $\text{EOD} = 0$ representing perfect equality of opportunity across all subgroups.

In our study, we prioritized the equal opportunity criterion, as our educational experts support the idea that fairness should ensure that students at risk of failing are equally identified across groups so that they are fairly provided with the necessary support and accommodations. This objective aligns with previous studies in ML for educational applications (Gardner et al., 2019; Hu & Rangwala, 2020).

(2) Average Odds Difference (AOD)

The Average Odds Difference (AOD; Bellamy et al., 2018) is a fairness metric that assesses the equalized odds criterion, which requires a classifier to have both similar true positive rates (TPR) and false positive rates (FPR) across all subgroups (Hardt et al., 2016). AOD measures the average difference between the TPR and FPR across all subgroups, providing an aggregate assessment of the classifier's fairness (Bellamy et al., 2018).

While achieving perfect equality of these rates for all subgroups in real-world scenarios is challenging, AOD provides a comprehensive evaluation of model fairness by considering both TPR and FPR. This metric complements the EOD in our analysis, offering a more stringent fairness assessment. As such, we also used the Average Odds Difference (AOD) metric to quantify the *equalized odds* criterion.

For this study, if the false positive rate (FPR) for a specific subgroup s is defined as

$$\text{FPR}(D_{S=s}) = P(\hat{Y} = 1 \mid Y = 0, S = s)$$

then, AOD can be described as:

$$\text{AOD} = \frac{1}{2}\left(|\text{TPR}(D_{S=1}) - \text{TPR}(D_{S=0})| + |\text{FPR}(D_{S=1}) - \text{FPR}(D_{S=0})|\right)$$

Contextually, we focused on developing efficient predictive models for student performance in VLEs. At this point, overestimating student performance (i.e., students who are predicted to pass but fail) has more severe implications, as these students may miss crucial interventions and support. This can lead to a false sense of security and overlooked learning struggles, increasing the risk of failure. Educational experts support the idea that the focus on fairness should be on ensuring all at-risk students, including those who might not initially seem destined to fail, are equally identified across groups to receive necessary support and accommodation. This objective aligns with the use of the equal opportunity criterion, which is also considered by previous studies in ML for educational applications (Gardner et al., 2019; Hu & Rangwala, 2020).

When expert opinions are unavailable, an alternative fairness objective is the equalized odds criterion (Hardt et al., 2016), which imposes a more rigorous standard than the equal opportunity criterion (Hardt et al., 2016). According to the equalized odds criterion, a classifier is considered fair if it maintains consistent true positive rates (TPR) and false positive rates (FPR) across all subgroups. This means that the differences in both the TPR and FPR between different subgroups should be minimal or zero, as the following mathematical definition provided by Hardt et al. (2016)

$$P(\hat{Y} = 1 \mid Y = 1, S = 1) - P(\hat{Y} = 1 \mid Y = 1, S = 0) \leq \epsilon$$

and

$$P(\hat{Y} = 1 \mid Y = 0, S = 1) - P(\hat{Y} = 1 \mid Y = 0, S = 0) \leq \epsilon$$

where both the TPR and FPR are constrained by a small value ϵ, indicating that smaller differences across different groups denote better fairness of the models.

For this study, if the false positive rate (FPR) for a specific subgroup s is defined as:

$$\text{FPR}(D_{S=s}) = P(\hat{Y} = 1 \mid Y = 0, S = s)$$

then, AOD can be described as:

$$\text{AOD} = \frac{1}{2}\left(|\text{TPR}(D_{S=1}) - \text{TPR}(D_{S=0})| + |\text{FPR}(D_{S=1}) - \text{FPR}(D_{S=0})|\right)$$

The factor ½ in the AOD equation normalizes the sum of the absolute differences in the true positive rates and false positive rates, ensuring the AOD values falls between 0 and 1 for easier interpretation and comparison.

2.2.3. Bias Mitigation Techniques

Based on our review of existing bias mitigation techniques and their applicability with VLE data, we selected four approaches for our study: Reweighing (RW), Suppression (SUP), Disparate Impact Remover (DIR), and Calibrated Equalized Odds Post-processing (CPP). These techniques were chosen for their complementary strengths and their potential to address different aspects of bias in our predictive models.

RW and SUP are preprocessing techniques that aim to adjust the training data to reduce bias. RW assigns weights to samples based on their group and label combinations, while SUP removes potentially biasing attributes from the dataset. These methods have shown promise in addressing representation biases but may have limitations in capturing complex interactions between features.

DIR is a preprocessing technique that adjusts feature distributions to reduce disparate impact, offering a balance between bias mitigation and preservation for predictive power. CPP, on the other hand, is a post-processing technique that adjusts model output to satisfy fairness constraints. By including both pre- and post-processing methods, we aim to compare their effectiveness in the VLE context and identify potential synergies or trade-offs between these approaches.

3. Results and Discussion

In the following analysis, we investigate the performance of the LR and XGBoost models and the impact of bias mitigation techniques on their performance. To begin, we trained the ML models on the dataset and assessed their predictive capabilities alongside any presence of unfair bias. Subsequently, we applied a range of bias mitigation approaches to these models and evaluated their effectiveness by examining the trade-offs between predictive performance and fairness in the modified versions. This investigation was conducted using the OULAD dataset, which focuses on predicting student academic outcomes.

3.1. Initial Model Performance and Assessment

We evaluated the predictive performance and fairness of the LR and XGBoost models on the OULAD dataset, considering various key demographic attributes such as gender, age, disability, and socioeconomic status. Table 2 presents the AUC-ROC precision, recall, F1-score, accuracy, macro average, and weighted average metrics for the baseline models without any bias mitigation.

The LR model achieved an accuracy of 75%, with a macro average precision, recall, and F1-score all at 0.74. The AUC-ROC score for the logistic regression model is 0.8328, indicating a good ability to distinguish between the pass and fail classes.

Table 2. Logistic Regression Classification Report

Metric	Precision	Recall	F1-Score	Support
Class - Fail	0.71	0.70	0.71	11,039
Class - Pass	0.78	0.79	0.78	14,651
Accuracy			0.75	25,690
Macro Avg	0.74	0.74	0.74	25,690
Weighted Avg	0.75	0.75	0.75	25,690

The XGBoost model achieved an accuracy of 74%, with macro average precision, recall, and F1-score all at 0.74. The AUC-ROC score for the XGBoost model is 0.8433, which is slightly higher than that of the LR model, indicating marginally better performance in distinguishing between the pass and fail classes.

Table 3. XGBoost Classification Report

Metric	Precision	Recall	F1-Score	Support
Class - Fail	0.69	0.72	0.71	11,039
Class - Pass	0.78	0.76	0.77	14,651
Accuracy			0.74	25,690
Macro Avg	0.74	0.74	0.74	25,690
Weighted Avg	0.74	0.74	0.74	25,690

Both models demonstrate robust performance in predicting student success in a VLE, with XGBoost showing a slight edge in terms of AUC and ROC curve (Figure 5). These results from the baseline for evaluating the impact of bias mitigation techniques on model fairness and performance. The plot showcases the true positive rates (TPRs) for both the baseline outcomes.

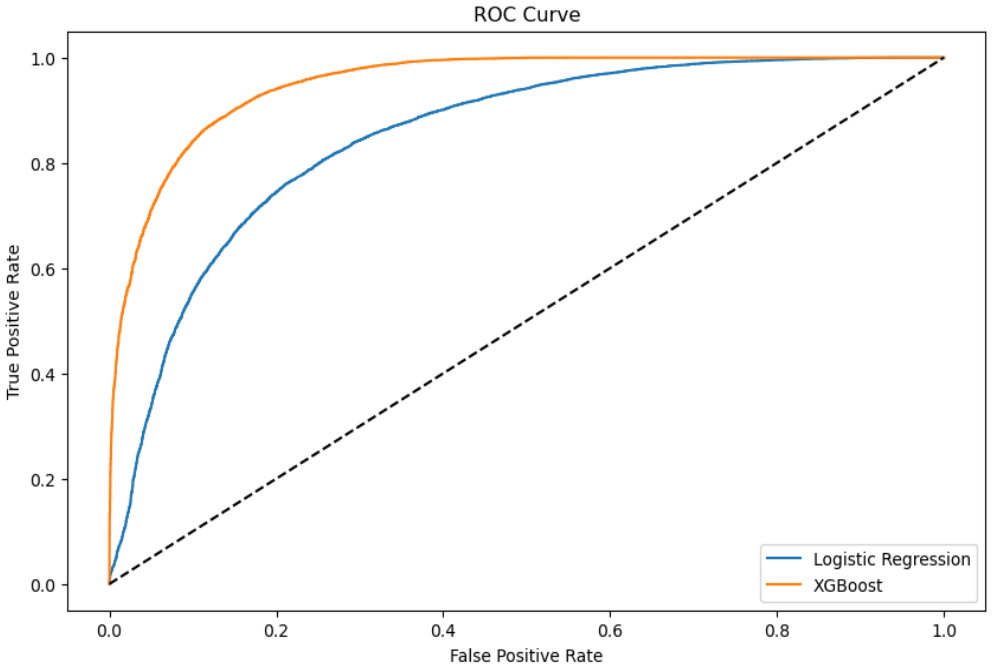

Figure 5. AUC-ROC curves comparing the performance of Logistic Regression and XGBoost classifiers on the OULAD.

These findings suggest that the predictive performance of ML models improves with larger training datasets, as illustrated by the OULAD dataset. Figure 6, a comparative analysis of group-specific true positive rates (TPRs) for LR classifiers, presents both baseline and post-mitigation outcomes. It showcases the true positive rates (TPR) for both the baseline and fair-aware LR models across the OULAD dataset and key demographic attributes. This provides a thorough comparison of fairness performance before and after bias mitigation. The plots reveal biases in the baseline classifier performance across different subgroups. As evident from the varying prevalence rates and unequal true positive rates (TPRs). These biases suggest disparate impact based on factors such as sex, age, index of multiple deprivation and disability. The top section is the base chart (before any bias mitigation is applied) and the bottom part displays the same model after mitigation has been applied. Additionally, we used Tukey's (1949) range test to conduct pairwise comparisons among groups, identifying statistically significant differences in TPR and FPR. We also mapped the 95% confidence intervals for these comparisons.

In Figure 6, each plot displays the results for a specific Dataset-Protected Attribute combination, with paired rows comparing the baseline classifiers to the bias-mitigated classifiers. The plots report the best outcomes from the tested bias mitigation algorithms. Points represent the mean TPR, with error bars indicating 95% confidence intervals derived from k-fold cross-validation. The baseline classifiers frequently exhibit biased behaviour due to factors such as inadequate representation, varying prevalence rates across groups, unequal feature distribution, or a combination of these factors within the four key demographic attributes. The comparative analysis across these demographic subgroups highlights the inherent biases present in the initial machine learning models and the effectiveness of different bias mitigation techniques. The DIR technique successfully aligns the TPRs across gender, age, and disability subgroups, although some residual bias remains. The SUP technique effectively mitigates biases related to socioeconomic status, as evidenced by the IMD analysis.

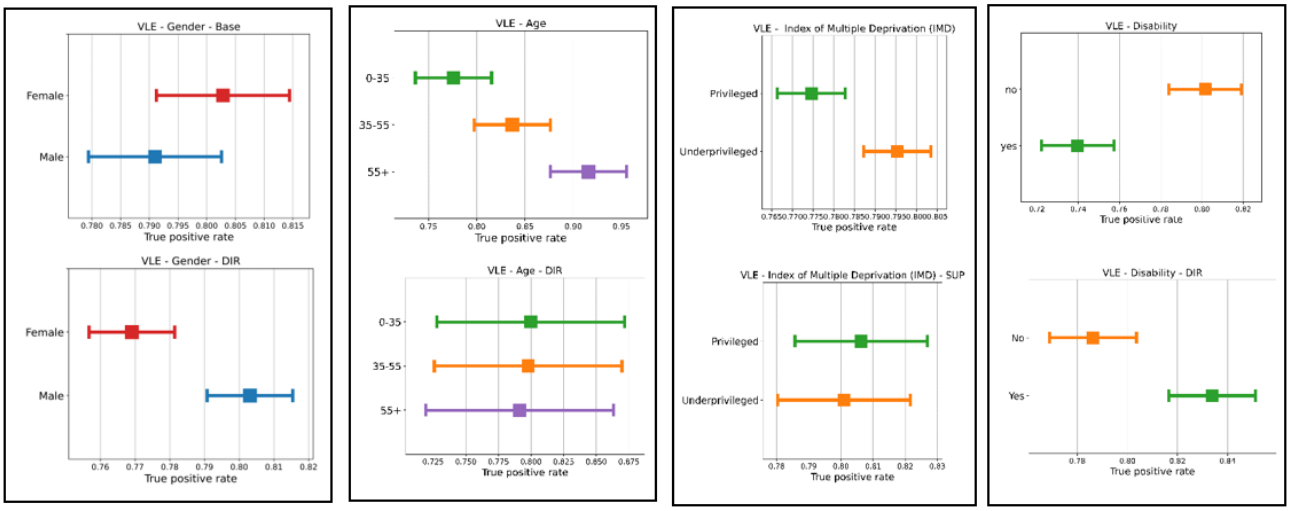

Figure 6. A comparative analysis of True Positive Rates (TPR) for various demographic subgroups (gender, age, disability, and IMD band) across different fairness conditions in VLEs using the OULAD. In each plot, colours are used to differentiate feature levels: red represents females and blue represents males in the gender subgroup analysis; green, orange, and purple represent different age bands (0–35, 35–55, 55+) in the age analysis; green and orange represent "Privileged" and "Underprivileged" categories in the IMD band analysis; and green and orange indicate "Yes" and "No" for disability status. Error bars indicate confidence intervals for the TPR estimates. This analysis aims to evaluate the effectiveness of different bias mitigation techniques on machine learning models, specifically focusing on LR classifiers.

Gender Analysis

Base Model (Gender): In Figure 6, the top plot shows the TPR for male and female students without any bias mitigation. The female subgroup (red) demonstrates a higher TPR compared to the male subgroup (blue), indicating a potential bias where the model predicts better for female students.

Disparate Impact Remover (Gender): In Figure 6 the bottom plot depicts the TPR after applying the Disparate Impact Remover (DIR) technique. Post-mitigation, the TPR for males and females is more balanced, though some disparity still exists, showing the effectiveness of DIR in reducing gender bias.

Age Analysis

Base Model (Age): In Figure 6, the top plot represents TPRs for different age groups (0–35, 35–55, 55+) in the base model. There is a noticeable disparity, with the youngest age group (0–35) having the highest TPR and the oldest age group (55+) having the lowest TPR, indicating an age bias in the predictions.

Disparate Impact Remover (Age): The bottom plot shows TPRs after applying DIR. The TPRs for all age groups are more closely aligned, indicating that DIR effectively mitigates age-related biases, although the oldest age group still lags slightly behind.

IMD Analysis

Base Model (IMD): The top plot illustrates the TPR for privileged and underprivileged groups based on the Index of Multiple Deprivation (IMD) in the base model (see Figure 6). The privileged group shows a slightly higher TPR, suggesting a bias favouring more privileged students.

Suppression (IMD): After applying the SUP technique, the bottom plot shows the TPR. The TPRs for both groups are now closer, indicating that SUP effectively reduces bias related to socioeconomic status (see Figure 6).

Disability Analysis

Base Model (Disability): The top plot shows the TPR for students with and without disabilities in the base model. There is a clear disparity, with non-disabled students having a higher TPR compared to disabled students, indicating a bias against disabled students (see Figure 6).

Disparate Impact Remover (Disability): The bottom plot shows the TPR after applying DIR. The TPRs for both groups are more aligned, demonstrating DIR's effectiveness in mitigating disability-related biases (see Figure 6).

3.2. Model Performance with Bias Mitigation

We evaluated the effectiveness of bias mitigation techniques by comparing fairness metrics between the baseline models and those obtained after applying these techniques. In VLEs, it is vital to ensure that the quest for fairness does not decrease the identification of at-risk students. The balance between fairness and accuracy, where improving fairness can sometimes reduce

predictive performance, is widely discussed in literature (Carey & Wu, 2023; Caton & Haas, 2023). In educational contexts, compromising accuracy for fairness can be problematic, as it may delay necessary support for struggling students. Thus, our goals are to improve accuracy and reduce discrimination. Figure 7 presents a comparative analysis of balanced accuracy versus equal opportunity difference (EOD) across four key demographic attributes - gender, age, IMD, and disability - in VLEs using LR models and various bias mitigation techniques.

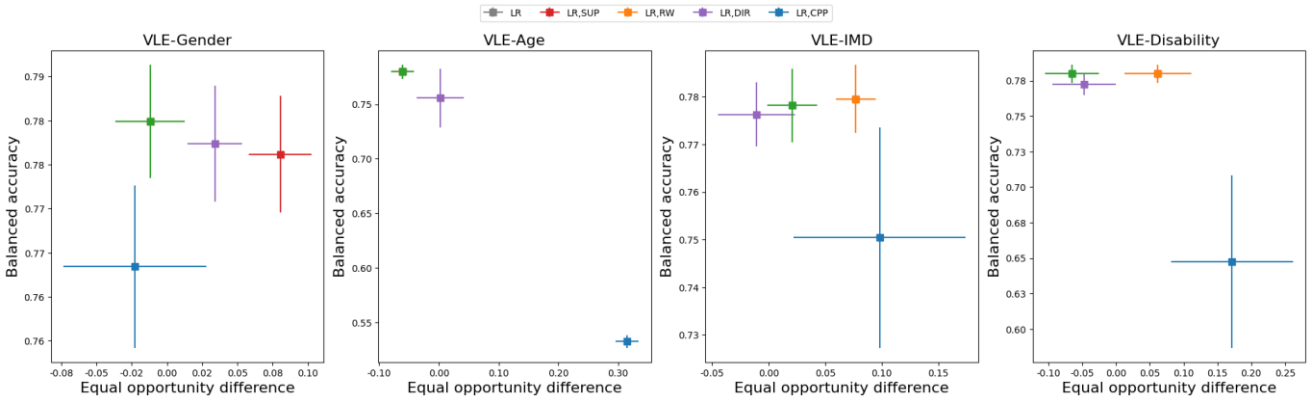

Figure 7. The accuracy–fairness trade-off, as measured by the Equal Opportunity Difference (EOD) and Balanced Accuracy (BAcc), is illustrated for the original model and the classifiers obtained after applying four bias mitigation techniques to the LR models. Each data point represents the average BAcc for TRP, while the error bars depict the standard deviation across the k-fold cross-validation procedure. The plots are organized to display the results for each subgroup, considering specific combinations of datasets and key demographic attributes.

Gender

In the context of gender in Figure 7, the LR model's balanced accuracy hovers around 0.78 for all bias mitigation techniques. The EOD remains relatively low, indicating minimal disparity in true positive rates between male and female students. This suggests that the mitigation techniques effectively reduce bias without significantly compromising the model's accuracy. The effectiveness of such bias mitigation is crucial as highlighted by Fenu et al. (2022), who stress the importance of gender fairness in educational AI systems.

Age

For age, the chart in Figure 7 shows a substantial spread in both balanced accuracy and EOD. The Calibrated Equalized Odds Post-processing (CPP) method exhibits a lower accuracy and a high EOD, suggesting it struggles to balance fairness and performance for this attribute. In contrast, the other techniques manage to keep the EOD minimal while maintaining higher accuracy. This finding aligns with the challenges discussed by Caton and Haas (2023) and Corbett-Davies et. al., (2023), who emphasize the trade-offs between fairness and accuracy, especially in educational settings where different age groups may have varying learning capabilities and access to resources.

IMD

The IMD analysis in Figure 7 indicates that many bias mitigation techniques achieve balanced accuracy close to 0.78 with minimal EOD. However, the CPP method again shows a significant drop in accuracy coupled with a higher EOD, reflecting a struggle to achieve equity without sacrificing performance. This resonates with the findings of Zhang et al. (2023), who discuss the inherent difficulties in addressing socioeconomic disparities through algorithmic interventions.

Disability

For disability in Figure 7, the models demonstrate a wide variation in balanced accuracy and EOD, with the CPP method showing the largest disparity. This suggests that while some techniques like DIR and RW manage to keep both metrics at reasonable levels, the CPP method struggles significantly. This outcome supports the literature on educational fairness, such as the work by Kizilcec and Lee (2022), which highlights the complexities of mitigating bias for students with disabilities, given their diverse needs and interactions with the learning environment.

3.3. Summary of Analysis

The results underscore the complex relationship between fairness and accuracy in VLEs, echoing the ongoing discussions in the literature about algorithmic fairness. As Kizilcec and Lee (2022) note, applying fairness techniques in educational settings requires careful consideration of the specific context and the key demographic attributes involved. The varied performance of bias mitigation techniques across different attributes in this study highlights the need for tailored approaches to algorithmic fairness, rather than one-size-fits-all solutions.

Moreover, the significant challenges faced by the CPP method in this analysis align with the broader critique of post-processing techniques, as discussed by Pleiss et al. (2017), who argue that while post-processing can address disparities in outcomes, it often does so at the cost of predictive performance.

The comparative analysis presented in this chart emphasizes the importance of selecting appropriate bias mitigation techniques for different key demographic attributes in virtual learning environments. The findings suggest that while some debiasing methods like DIR and RW can balance fairness and accuracy effectively, others like CPP and SUP may introduce significant trade-offs. These insights are crucial for developing fair and effective educational AI systems, contributing to the broader discourse on algorithmic fairness in education.

4. Discussion

The increasing adoption of ML in VLEs has brought to light the critical issue of algorithmic bias in education. While predictive models have the potential to provide insights regarding personalizing learning experiences and identifying at-risk students for timely interventions, they also risk perpetuating or amplifying existing biases and inequities in education (Kizilcec & Lee, 2022; Xu et al., 2022). For example, models trained on historical data may reflect societal biases related to gender, race, or socioeconomic status, leading to disparate impacts on certain student populations. Additionally, biased predictive models may reinforce stereotypes, limit opportunities, or misallocate resources, exacerbating achievement gaps and hindering educational equity.

This study aimed to investigate the presence of bias in standard ML approaches applied to VLE data and evaluate the effectiveness of various bias mitigation techniques in promoting fairness across demographic subgroups. The findings contribute to the growing body of research on algorithmic fairness in higher education and highlight the importance of developing context-specific and nuanced approaches to ensure equitable outcomes for all students. As VLEs continue to play a significant role in shaping the educational experiences of college and university students, addressing algorithmic bias becomes crucial for promoting inclusive and equitable learning environments.

4.1. Uncovering Bias in Conventional ML Approaches

The results of this study demonstrate that standard machine learning approaches, such as LR and XGBoost, can exhibit biased performance. The initial model assessment revealed disparities in true positive rates (TPRs) across various demographic subgroups, including gender, age, socioeconomic status (IMD), and disability. These findings align with previous research highlighting the presence of bias in educational AI systems (Kizilcec & Lee, 2022; Xu et al., 2022).

The observed biases can be attributed to several factors, such as inadequate representation of certain subgroups in the training data, varying prevalence rates across groups, and unequal feature distributions (Li et al., 2024). These factors can lead to models that systematically disadvantage students from historically marginalized groups, perpetuating educational inequities (Gardner et al., 2019; Li et al., 2024). The findings underscore the importance of auditing predictive models for fairness and applying debiasing interventions to ensure equitable student support. As Fenu et al. (2022) emphasize, addressing algorithmic bias is crucial for developing fair and inclusive educational AI systems that cater to the diverse needs of all students.

4.2. Effectiveness of Bias Mitigation Techniques

The comparative analysis of bias mitigation techniques demonstrates that algorithmic fairness can be improved in predictive models for student learning performance. The application of methods such as DIR, RW, and SUP effectively reduced disparities in TPRs across demographic subgroups, as evidenced by the decreased EOD values. The effectiveness of these techniques varied across key demographic attributes, highlighting the need for context-specific approaches to algorithmic fairness. While DIR showed considerable success in mitigating biases related to gender, age, and disability, SUP was particularly effective in addressing socioeconomic disparities captured by the IMD.

These findings contribute to the growing body of literature on algorithmic fairness in education (Kizilcec & Lee, 2022) by demonstrating the potential of bias mitigation techniques to promote equity in VLE-based predictive models. However, the residual biases observed in some cases underscore the challenges of achieving perfect fairness and the need for ongoing research and refinement of debiasing methods.

4.3. Navigating the Accuracy–Fairness Dilemma

The analysis of the accuracy–fairness trade-off reveals the complex relationship between these two objectives in VLE-based predictive models. While some bias mitigation techniques, such as DIR and RW, improved fairness metrics (EOD) while maintaining relatively high balanced accuracy (BAcc), others, like CPP, exhibited significant drops in accuracy coupled with higher EOD values. These findings align with the broader discourse on the inherent tensions between fairness and accuracy in algorithmic decision-making (Caton & Haas, 2023). In educational contexts, compromising accuracy for fairness can be particularly problematic, as it may lead to delayed or inadequate support for struggling students. However, the results also suggest that the accuracy–fairness trade-off is not always inevitable. The success of techniques like DIR and RW in balancing

these objectives indicates that carefully designed and context-specific interventions can help mitigate biases while preserving predictive performance (Fenu et al., 2022).

The varied performance of bias mitigation techniques across different key demographic attributes highlights the need for nuanced approaches to algorithmic fairness in education. As Kizilcec and Lee (2022) argue, developing fair and effective educational AI systems requires a deep understanding of the specific contexts, the key demographic attributes involved, and the potential impact of algorithmic decisions on student outcomes. Moreover, the challenges faced by post-processing techniques like CPP in this study echo the broader critiques of these methods, which often address disparities in outcomes at the cost of predictive performance (Pleiss et al., 2017). This underscores the importance of considering alternative approaches, such as preprocessing and in-processing techniques, to pursue algorithmic fairness.

5. Limitations and Future Research

Despite the practical implications our study has offered, there are still some limitations necessitating further investigations. One limitation of our study is the aggregation of "Fail" and "Withdrawn" categories into a single "Failed" outcome. Although this simplification supports our binary classification goal, it may obscure subtle nuances inherent in multi-class settings, potentially masking important distinctions between the two groups and introducing or amplifying biases in our model.

Future research could investigate more nuanced outcome categories or assess the impact of aggregation on model fairness and accuracy. Our analysis primarily relied on the OULAD dataset, which, while comprehensive, may not fully represent diverse virtual learning environments (VLEs) or student populations. Although the study yielded valuable insights, it did not account for various influential factors in real-world teaching and learning contexts, potentially limiting its contributions. Additionally, our focus on four common bias mitigation techniques leaves room for exploring alternative approaches that may yield different results. Lastly, the performance metrics used (AUC-ROC and BAcc) may not fully capture all aspects of model performance and fairness.

Lastly, we did not evaluate the long-term effects of these mitigation strategies in real-world educational settings. To enhance algorithmic fairness in VLE predictive analytics, we recommend best practices such as collecting and preprocessing VLE data with fairness in mind, assessing model fairness across multiple key demographic attributes and metrics, applying and comparing various bias mitigation techniques, and engaging in transparent communication and collaboration with educational stakeholders.

As the field continues to evolve, ongoing research and collaboration between educators, researchers, and policymakers will be crucial in advancing algorithmic fairness and ensuring that AI-driven interventions in education serve to narrow, rather than widen, achievement gaps.

6. Conclusion

The findings of this study contribute to the growing body of research on algorithmic fairness in education by demonstrating the presence of bias in AI predictive models and the potential of bias mitigation techniques to promote equity. The results highlight the complexity between fairness and accuracy, emphasizing the need for context-specific and nuanced approaches to developing fair and effective educational AI systems. The overall analytical pipeline outlined in our study provides a roadmap for further exploration and innovation in the field of algorithmic fairness in education. By expanding the scope of analysis, investigating long-term impacts, developing tailored mitigation techniques, exploring hybrid approaches, conducting qualitative research, examining ethical intersections, and fostering interdisciplinary collaboration, researchers can continue to push the boundaries of knowledge and practice in this critical area.

Declaration of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors declared no financial support for the research, authorship, and/or publication of this article.

References

Baker, R. S. & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Bayer, V., Hlosta, M., & Fernandez, M. (2021). Learning analytics and fairness: Do existing algorithms serve everyone equally? In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Artificial intelligence in education*: *22nd international conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, proceedings, part II* (pp. 71–75). Springer. https://doi.org/10.1007/978-3-030-78270-2_12

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). *AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias*. arXiv. https://doi.org/10.48550/arxiv.1810.01943

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. *2010 20th international conference on pattern recognition* (pp. 3121–3124). IEEE. https://doi.org/10.1109/ICPR.2010.764

Carey, A. N., & Wu, X. (2023). The statistical fairness field guide: Perspectives from social and formal sciences. *AI and Ethics, 3*(1), 1–23. https://doi.org/10.1007/s43681-022-00183-3

Caton, S., & Haas, C. (2023). Fairness in machine learning: A survey. *ACM Computing Surveys, 56*(7), Article 166. https://doi.org/10.1145/3616865

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *KDD '16: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chinta, S. V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Le Quy, T., & Zhang, W. (2024). *FairAIED: Navigating fairness, bias, and ethics in educational AI applications*. ArXiv. https://doi.org/10.48550/arxiv.2407.18745

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The Measure and Mismeasure of Fairness. *Journal of Machine Learning Research, 24*(312), 1–117. https://www.jmlr.org/papers/volume24/22-1511/22-1511.pdf

Deho, O. B., Zhan, C., Li, J., Liu, J., Liu, L., & Duy Le, T. (2022). How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? *British Journal of Educational Technology, 53*(4), 822–843. https://doi.org/10.1111/bjet.13217

Dutta, S., Wei, D., Yueksel, H., Chen, P.-Y., Liu, S., & Varshney, K. R. (2020). *Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing.* ArXiv. doi.org/10.48550/arxiv.1910.07870

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). *Fairness through awareness.* ArXiv. doi.org/10.48550/arxiv.1104.3913

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. doi.org/10.1016/j.patrec.2005.10.010

Fazil, M., Rísquez, A., & Halpin, C. (2024). A novel deep learning model for student performance prediction using engagement data. *Journal of Learning Analytics, 11*(2), 23–41. doi.org/10.18608/jla.2024.7985

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). *Certifying and removing disparate impact.* ArXiv. doi.org/10.48550/arxiv.1412.3756

Fenu, G., Galici, R., & Marras, M. (2022). Experts' view on challenges and needs for fairness in artificial intelligence for education. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial intelligence in education: 23rd international conference, AIED 2022, Durham, UK, July 27–31, 2022, proceedings, part I* (pp. 243–255). Springer. doi.org/10.1007/978-3-031-11644-5_20

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In S. Hsiao, J. Cunningham, K. McCarthy, G. Lynch, C. Brooks, R. Ferguson, & U. Hoppe (Eds.), *LAK19: Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 225–234). ACM Press. doi.org/10.1145/3303772.3303791

Hardt, M., Price, E., & Srebro, N. (2016). *Equality of opportunity in supervised learning.* ArXiv. doi.org/10.48550/arxiv.1610.02413

Hu, Q., & Rangwala, H. (2020). Towards fair educational data mining: A case study on detecting at-risk students. In A. N. Rafferty, J. Whitehill, C. Romero, & V. Cavalli-Sforza (Eds.), *Proceedings of the 13th international conference on educational data mining (EDM 2020)* (pp. 431–437). International Educational Data Mining Society. https://files.eric.ed.gov/fulltext/ED608050.pdf

Idowu, J. A. (2024). Debiasing education algorithms. *International Journal of Artificial Intelligence in Education, 34*(4), 1510–1540. doi.org/10.1007/s40593-023-00389-4

Johnston, L. J., Griffin, J. E., Manolopoulou, I., & Jendoubi, T. (2024). *Uncovering student engagement patterns in Moodle with interpretable machine learning.* ArXiv. doi.org/10.48550/arxiv.2412.11826

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1–33. doi.org/10.1007/s10115-011-0463-8

Kamiran, F., & Calders, T. (2009). Classifying without discriminating. In *2009 2nd international conference on computer, control and communication* (pp. 1–6). IEEE. doi.org/10.1109/IC4.2009.4909197

Khoudi, Z., Hafidi, N., Nachaoui, M., & Lyaqini, S. (2025). New approach to enhancing student performance prediction using machine learning techniques and clickstream data in virtual learning environments. *SN Computer Science, 6*, Article 139. https://doi.org/10.1007/s42979-024-03622-6

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In W. Holmes & K. Porayska-Pomsta (Eds.), *The ethics of artificial intelligence in education: Practices, challenges, and debates* (pp. 174–202). Routledge. doi.org/10.4324/9780429329067-10

Köchling, A., Riazy, S., Wehner, M. C., & Simbeck, K. (2021). Highly accurate, but still discriminatory. *Business & Information Systems Engineering, 63*(1), 39–54. doi.org/10.1007/s12599-020-00673-w

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University learning analytics dataset. *Scientific Data, 4*, Article 170171. doi.org/10.1038/sdata.2017.171

Lallé, S., Bouchet, F., Verger, M., & Luengo, V. (2024). Fairness of MOOC completion predictions across demographics and contextual variables. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, I. I. Bittencourt (Eds.), *Artificial intelligence in education: 25th international conference, AIED 2024, Recife, Brazil, July 8–12, 2024, proceedings, part I* (pp. 379–393). Springer. doi.org/10.1007/978-3-031-64302-6_27

Leite, W. L., Jing, Z., Kuang, H., Kim, D., & Huggins-Manley, A. C. (2021). Multilevel mixture modeling with propensity score weights for quasi-experimental evaluation of virtual learning environments. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(6), 964–982. doi.org/10.1080/10705511.2021.1919895

Le Quy, T., Nguyen, T. H., Friege, G., & Ntoutsi, E. (2023). Evaluation of group fairness measures in student performance prediction problems. In I. Koprinska, P. Mignone, R. Guidotti, S. Jaroszewicz, H. Fröning, F. Gullo, P. M. Ferreira, D. Roqueiro, G. Ceddia, S. Nowaczyk, J. Gama, R. Ribeiro, R. Gavaldà, E. Masciari, Z. Ras, E. Ritacco, F. Naretto, A. Theissler, P. Biecek, … S. Pashami (Eds.), *Machine learning and principles and practice of knowledge discovery in databases: International workshops of ECML PKDD 2022, Grenoble, France, September 19–23, 2022, proceedings, part I* (pp. 119–136). Springer. doi.org/10.1007/978-3-031-23618-1_8

Liu, S., & Vicente, L. N. (2022). *Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach.* arXiv. doi.org/10.48550/arxiv.2008.01132

Liu, Z., Jiao, X., Li, C., & Xing, W. (2024). Fair prediction of students' summative performance changes using online learning behavior data. *Proceedings of the 17th international conference on educational data mining* (pp. 686–691). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.12729918

Li, C., Xing, W., & Leite, W. (2024). Using fair AI to predict students' math learning outcomes in an online platform. *Interactive Learning Environments*, *32*(3), 1117–1136. doi.org/10.1080/10494820.2022.2115076

Martinez, A. L. J., Sood, K., & Mahto, R. (2025). Early detection of at-risk students using machine learning. In H. R. Arabnia, L. Deligiannidis, S. Amirian, F. Ghareh Mohammadi, & F. Shenavarmasouleh (Eds.), *Foundations of computer science and frontiers in education: Computer science and computer engineering: 20th international conference, FCS 2024, and 20th international conference, FECS 2024, held as part of the world congress in computer science, computer engineering and applied computing, CSCE 2024, Las Vegas, NV, USA, July 22–25, 2024, revised selected papers* (pp. 396–406). Springer. https://doi.org/10.1007/978-3-031-85930-4_36

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR), 54*(6), Article 115. https://doi.org/10.1145/3457607

Morik, M., Singh, A., Hong, J., & Joachims, T. (2020). Controlling fairness and bias in dynamic learning-to-rank. In J. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J.-R. Wen, & Y. Liu (Eds.), *SIGIR '20: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 429–438). ACM Press. doi.org/10.1145/3397271.3401100

Pei, B., & Xing, W. (2022). An interpretable pipeline for identifying at-risk students. *Journal of Educational Computing Research*, 60(2), 380–405. doi.org/10.1177/07356331211038168

Peng, L., & Jeang, B. (2023). Prediction model of students' learning behavior on learning effect in online live class based on machine learning algorithm. In Y. Zhong (Ed.), *Fifth international conference on computer information science and artificial intelligence (CISAI 2022)* (Article 1256650). SPIE. doi.org/10.1117/12.2669155

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). *On fairness and calibration.* ArXiv. doi.org/10.48550/arxiv.1709.02012

Raftopoulos, G., Davrazos, G., & Kotsiantis, S. (2025). Evaluating fairness strategies in educational data mining: A comparative study of bias mitigation techniques. *Electronics, 14*(9), Article 1856. https://doi.org/10.3390/electronics14091856

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs: Data Mining and Knowledge Discovery, 10*(3), Article e1355. doi.org/10.1002/widm.1355

Shin, J., Bulut, O., & Pinto, W. N., Jr. (2022). E-learning preparedness: A key consideration to promote fair learning analytics development in higher education. *Proceedings of the 15th international conference on educational data mining* (pp. 673–678). International Educational Data Mining Society. doi.org/10.5281/zenodo.6853111

Song, Y., Li, C., Xing, W., Li, S., & Hannah, H. (2024). A fair clustering approach to self-regulated learning behaviors in a virtual learning environment. In B. Flanagan, B. Wasson, & D. Gašević (Eds.), *LAK '24: Proceedings of the 14th learning analytics and knowledge conference* (pp. 771–778). ACM Press. https://doi.org/10.1145/3636555.3636863

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics, 5*(2), 99–114. doi.org/10.2307/3001913

Verger, M., Fan, C., Lallé, S., Bouchet, F., & Luengo, V. (2024). A comprehensive study on evaluating and mitigating algorithmic unfairness with the MADD metric. *Journal of Educational Data Mining, 16*(1), 365–409. doi.org/10.5281/zenodo.12180668

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In Y. Brun, B. Johnson, & A. Meliou (Eds.), *FairWare '18: Proceedings of the international workshop on software fairness*, 1–7. doi.org/10.1145/3194770.3194776

Wang, Y., Wang, X., Beutel, A., Prost, F., Chen, J., & Chi, E. H. (2021). Understanding and improving fairness–accuracy trade-offs in multi-task learning. In F. Zhu, B. C. Ooi, C. Miao, H. Wang, I. Skrypnyk, W. Hsu, & S. Chawla (Eds.), *KDD '21: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining* (pp. 1748–1757). ACM Press. doi.org/10.1145/3447548.3467326

Wei, Q., & Dunbrack, R. L., Jr. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLOS ONE, 8*(7), Article e67863. doi.org/10.1371/journal.pone.0067863

Wongvorachan, T., Bulut, O., Liu, J. X., & Mazzullo, E. (2024). A comparison of bias mitigation techniques for educational classification tasks using supervised machine learning. *Information*, *15*(6), Article 326. doi.org/10.3390/info15060326

Xing, W., & Du, D. (2018). Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research, 57*(3), 547–570. doi.org/10.1177/0735633118757015

Xu, J., Xiao, Y., Wang, W. H., Ning, Y., Shenkman, E. A., Bian, J., & Wang, F. (2022). Algorithmic fairness in computational medicine. *eBioMedicine, 84*, Article 104250. doi.org/10.1016/j.ebiom.2022.104250

Yu, H.-F., Huang, F.-L., & Lin, C.-J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning, 85*(1–2), 41–75. doi.org/10.1007/s10994-010-5221-8

Zhang, F., Xing, W., & Li, C. (2023). Predicting students' Algebra I performance using reinforcement learning with multi-group fairness. In I. Hilliger, H. Khosravi, B. Rienties, & S. Dawson (Eds.), *LAK23: 13th international learning analytics and knowledge conference* (pp. 657–662). ACM Press. doi.org/10.1145/3576050.3576104

Zhao, C., Mi, F., Wu, X., Jiang, K., Khan, L., & Chen, F. (2024). Dynamic environment responsive online meta-learning with fairness awareness. *ACM Transactions on Knowledge Discovery from Data, 18*(6), Article 153. doi.org/10.1145/3648684