



Debiasing Education Algorithms

Jamiu Adekunle Idowu¹ 

Accepted: 15 December 2023 / Published online: 4 January 2024
© The Author(s) 2023

Abstract

This systematic literature review investigates the fairness of machine learning algorithms in educational settings, focusing on recent studies and their proposed solutions to address biases. Applications analyzed include student dropout prediction, performance prediction, forum post classification, and recommender systems. We identify common strategies, such as adjusting sample weights, bias attenuation methods, fairness through un/awareness, and adversarial learning. Commonly used metrics for fairness assessment include ABROCA, group difference in performance, and disparity metrics. The review underscores the need for context-specific approaches to ensure equitable treatment and reveals that most studies found no strict tradeoff between fairness and accuracy. We recommend evaluating fairness of data and features before algorithmic fairness to prevent algorithms from receiving discriminatory inputs, expanding the scope of education fairness studies beyond gender and race to include other demographic attributes, and assessing the impact of fair algorithms on end users, as human perceptions may not align with algorithmic fairness measures.

Keywords Debiasing education algorithms · Fairness metrics · Algorithmic fairness · Systematic literature review

Introduction

Motivation

In 2020, the UK's decision to use an algorithm to predict A-level grades led to a public outcry. The algorithm disproportionately downgraded students from disadvantaged backgrounds while favoring those from affluent schools. This is just one out of many cases of how algorithms perpetuate existing biases. For instance, an

✉ Jamiu Adekunle Idowu
ucabaid@ucl.ac.uk

¹ Centre for Artificial Intelligence, Department of Computer Science, University College London, London, UK

algorithm developed to identify at-risk students performed worse for African-American students (Hu & Rangwala, 2020). Similarly, the predictive algorithms built by Yu et al. (2020) for undergraduate course grades and average GPA recorded better performance for male students than female students. In addition, a SpeechRater system for automated scoring of language assessment had to lower accuracy for students with speech impairment (Loukina & Buzick, 2017). Indeed, for AI applications to have a positive impact on education, it is crucial that their design considers fairness at every step (Suresh & Guttag, 2021; Gusenbauer & Haddaway, 2020).

In response to these challenges, the Learning Analytics and AIED research community has increasingly focused on fairness, equity, and responsibility. Recent studies have begun to address these issues more directly. For instance, Patterson et al. (2023) proposed the adoption of the anticipation, inclusion, responsiveness, and reflexivity (AIRR) framework (originally introduced by Stilgoe et al., 2013) in Learning Analytics. The study further applied the framework to creating an early-alert retention system at James Madison University. Also, Rets et al. (2023) provided six practical recommendations to enable the ethical use of predictive learning analytics in distance education. These include active involvement of end users in designing and implementing LA tools, inclusion through consideration of diverse student needs, stakeholder engagement, and a clear plan for student support interventions. Also, previous studies attempted to eliminate bias from education algorithms by removing sensitive features. However, this ‘fairness by unawareness’ approach is not a foolproof solution as bias can persist due to correlation with non-sensitive features (Jiang & Pardos, 2021) or the presence of proxy attributes (Pesach & Shmueli, 2022). For instance, brand associations like “Victoria Secret” may be highly correlated with gender, occupation/educational attainment may be correlated with socio-economic status, and geographic location may be highly correlated with race. In our review, we will examine various bias mitigation strategies implemented in the existing literature.

Meanwhile, a study by Gedrimiene et al. (2023) found accuracy to be a stronger predictor of students’ trustworthiness in an LA career recommender system compared to the students’ understanding of the origins of the recommendations (explainability). This aligns with Wang et al., 2022 (one of the papers reviewed in this SLR) which reported that in an online user study of more than 200 college students, participants preferred an original biased career recommender system over the debiased version of the system. Given these findings, this SLR will analyse the tradeoff between accuracy and fairness in education algorithms.

Moreover, to assess fairness of algorithms, it is crucial to choose an appropriate metric especially as studies have shown that it is impossible to satisfy multiple measures of fairness simultaneously. For instance, pursuing demographic parity will harm equalized odds (Baker & Hawn, 2021; Corbett-Davies et al., 2023). Therefore, it is important to analyse fairness measures being adopted in the existing literature on education algorithmic fairness.

This systematic literature review aims to address the following objectives:

- Research Objective 1: To identify the key metrics for assessing the fairness of education algorithms.

- Research Objective 2: To identify and analyze the bias mitigation strategies and sensitive features used in educational algorithms.
- Research Objective 3: To investigate the tradeoff between fairness and accuracy in educational algorithms.

Fairness in Machine Learning Algorithms

Fairness in machine learning has become an increasingly central topic, driven by the growing awareness of the societal implications of algorithmic decisions. While the design and training of algorithms can increase unfairness, biases often come from the data itself capturing historical prejudices, cultural stereotypes, or demographic disparities, which in turn influence model behaviour (Yu et al., 2020; Jiang & Pardo, 2021).

Although the definition of fairness is a subject of debate, at the broadest level, fairness of algorithms falls into two categories: individual fairness and group fairness.

Individual Fairness emphasizes that similar individuals should receive similar outcomes. Some common measures in this category include:

- **Fairness Through Awareness:** An algorithm is considered fair if it gives similar predictions to similar individuals (Dwork et al., 2012).
- **Fairness Through Unawareness:** Fairness is achieved if protected attributes like race, age, or gender are not explicitly incorporated in the decision-making process (Gajane & Pechenizkiy, 2017).
- **Counterfactual Fairness:** A decision is deemed fair for an individual if it remains the same in both the real and a counterfactual world where the individual belongs to a different demographic group (Kusner et al., 2017).

Group Fairness, on the other hand, focuses on ensuring equal statistical outcomes across distinct groups partitioned by protected attributes. Some widely used measures here are:

- **Equal Opportunity:** The probability of a person in a positive class being assigned to a positive outcome should be equal for both protected and unprotected group members (Hardt et al., 2016).
- **Demographic Parity:** The likelihood of a positive outcome should be the same regardless of whether the person is in the protected group (Dwork et al., 2012).
- **Equalized Odds:** The probabilities of true positives and false positives should be the same for both protected and unprotected groups (Hardt et al., 2016).

Meanwhile, despite the advancement in algorithmic fairness research, defining fairness remains challenging, as many existing definitions may not always align with real-world applications (Gardner et al., 2019). For instance, while demographic parity requires positive prediction probabilities to be consistent across groups, it may lead to imbalances when legitimate reasons exist for discrepancies among groups

(Gardner et al., 2019). Recognizing this, Yu et al. (2021) emphasized the importance of contextualizing fairness based on the application.

Method

This systematic literature review (SLR) collates, classifies and interprets the existing literature that fits the pre-defined eligibility criteria of the review. The 2020 updated PRISMA guideline is adopted for the review (Page et al., 2021). The stages involved in this SLR are discussed as follows:

Stage 1: Data Sources and Search Strategy

Scopus was selected as the principal search system for this systematic review because it is extensive (70,000,000+ publications) and can efficiently perform searches with high precision, recall, and reproducibility (Gusenbauer & Haddaway, 2020). In addition, ACM Digital Library and IEEE Xplore are used as supplementary search systems because they are specifically focused on the computer science subject (Gusenbauer & Haddaway, 2020). Table 1 provides more details into the data sources.

The keywords used for the search strategy are presented in Table 2.

Stage 2: Inclusion and Exclusion Criteria

This section outlines the inclusion and exclusion criteria for the SLR. We screened studies in three steps: first by titles and keywords, then abstracts, and finally through full-text reading. Studies that met the inclusion criteria were further evaluated in Stage 3 using the quality assessment criteria.

Inclusion Criteria:

- Papers that directly address the issues of algorithmic bias in educational settings or tools.
- Studies published from 2015 onwards.
- Research materials such as peer-reviewed journal articles, conference proceedings, white papers, etc. all with clearly defined research questions.

Exclusion Criteria:

- Non-research materials like newsletters, magazines, posters, surveys, invited talks, panels, keynotes, and tutorials.
- Duplicate research paper or article.

Implementation of Search Strategy:

We executed our search strategy across the three data sources using the queries specified in Table 2. Our decision to focus on studies from 2015 onwards was

Table 1 Data sources for identifying relevant studies

Data source	Type	Years covered	Justification
Scopus	Abstract and Citations database (journal articles, conference proceedings, etc.)	All years	Extensive publications, can be used as principal search system, full index, bulk download, repeatable, etc. (Gusenbauer & Haddaway, 2020).
ACM	Full text Digital Library	All years	Primary focus on Computer Science, can be used as a principal search system, full index, bulk download, repeatable, etc. (Gusenbauer & Haddaway, 2020).
IEEE Xplore	Digital Library database (journals, conference papers, etc.)	All years	Focus on Computer Science, full index, bulk download, repeatable, etc. (Gusenbauer & Haddaway, 2020).

Table 2 Search keywords

Term	Keywords
Artificial Intelligence	"machine learning" OR "artificial intelligence" OR AI OR algorithm* OR automate* OR recommend* AND
Fairness	bias* OR fair* OR ethic* OR trust* OR diverse* OR responsible AND
Education	education, teacher, student, pupil, classroom, grading, curricula*

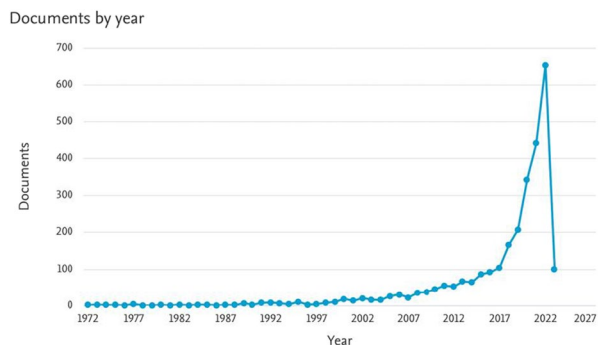
validated upon examining the distribution of publications, as the literature indicated an uptick in relevant studies within this timeframe as shown in Fig. 1 (for the principal search system, Scopus).

Stage 3: Quality Assessment

In this systematic literature review, we employed a rigorous quality assessment framework to ensure the validity and reliability of the studies included. This framework encompasses four quality assessment questions listed below. Only papers that satisfied at least 3 out of the 4 quality assessment questions were considered for further analysis.

- QA1. Did the study have clearly defined research questions and methodology?
- QA2. Did the study incorporate considerations of fairness during the algorithm development process, ensuring non-discrimination based on factors like gender, race, disability, socioeconomic status, etc.?
- QA3. Did the study report the performance of the algorithm using suitable metrics such as accuracy, F1 score, AUC-ROC, precision, recall, etc.?
- QA4. Was the study based on a real-world educational dataset?

Fig. 1 Distribution of targeted studies by year in Scopus



Results

A total of 3424 targeted studies were identified. 2747 of these are from the principal search system (Scopus), 357 from ACM, and 320 from IEEE Xplore. Using EndNote, a reference management package, 426 duplicates were detected and removed leaving 2998 studies whose titles and/or abstracts were screened.

By screening the studies' titles and keywords, 2865 papers were sieved out while the abstracts of the remaining 133 papers were fully read. At the end of the screening exercise, 86 papers were removed, leaving 47 papers as eligible for full text screening. After full text screening, 31 papers were dropped leaving 16 papers. A consultation with experts yielded one additional paper, Counterfactual Fairness, a study that developed a framework for modeling fairness using causal inference and demonstrated the use of the framework with fair prediction of students' success at Law School (Kusner et al., 2017). In the end, 12 papers made it through the 'thick and thin' of our quality assessment and became eligible papers. Figure 2 presents a PRISMA flow chart of the study selection process. The data extraction table containing the study design, dataset, methods, evaluation metrics, results, conclusion, and limitations of the eligible papers is presented in Appendix Table 6.

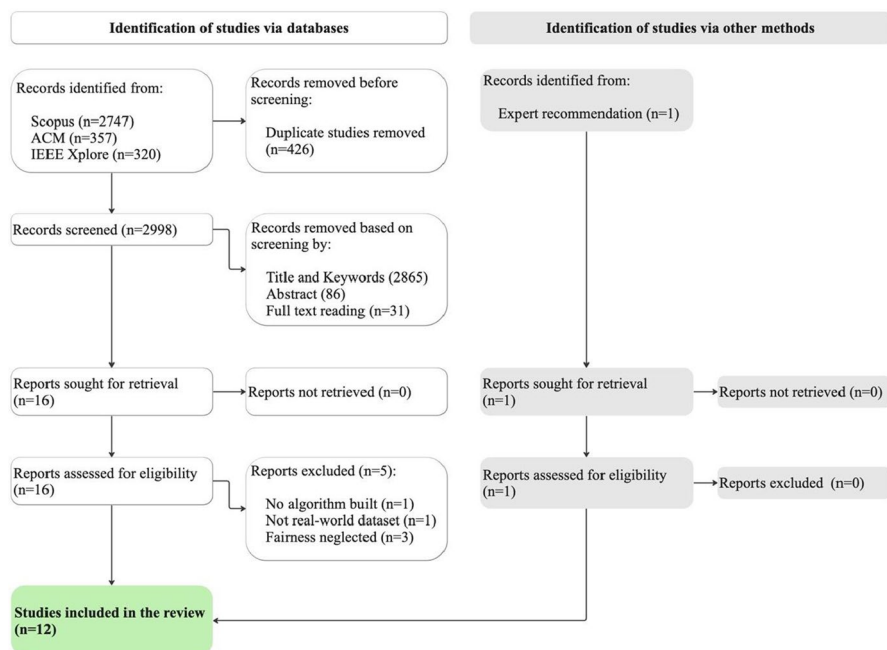


Fig. 2 Flow diagram of study selection using PRISMA 2020 guideline

Discussion

Research Objective 1: Measuring the Fairness of Education Algorithms

According to **S1** (Gardner et al., 2019), a necessary step towards correcting any unfairness in algorithms is to measure it first. However, there has been a challenge around finding the appropriate metric for measuring the fairness of education algorithms. Therefore, this section explores the metrics employed by selected studies to measure fairness in education algorithms as highlighted in Table 3.

1. ABROCA (S1, S2 & S6)

As noted by Wang et al. (2022), most studies in learning analytics typically measure algorithmic fairness using statistical parity, equalised odds, and equalised opportunities. However, these methods are threshold dependent, for instance, equalised odds can only be achieved at points where the ROC curves cross thereby failing to evaluate model discrimination at any other threshold (Gard-

Table 3 Fairness metrics adopted by selected studies

Paper	Reference	ABROCA	Group Δ in performance	TPR TNR	FNR & FPR	Others	Performance metric
S1	Gardner et al. (2019)	✓					AUC-ROC
S2	Sha et al. (2022)	✓					AUC
S3	Wang et al. (2022)					U_{PAR}	*NDCG@K
S4	Yu et al. (2021)		✓				Accuracy & Recall
S5	Jiang and Pardos (2021)			✓			Accuracy
S6	Sha et al. (2021)	✓					Accuracy, AUC & F1
S7	Yu et al. (2020)				✓		Accuracy
S8	Hu and Rangwala (2020)		✓				Accuracy
S9	Lee and Kizilcec (2020)		✓	✓			Accuracy and F1
S10	Anderson et al. (2019)		✓		✓		AUC-ROC
S11	Loukina et al. (2019)		✓				Accuracy
S12	Kusner et al. (2017)					Counter-factual Fairness	RMSE

* NDCG@K - Normalized discounted cumulative gain at K

ner et al., 2019). Therefore, **S1** (Gardner et al., 2019) developed a methodology to measure unfairness in predictive models using slicing analysis. This method involves evaluating the model performance across multiple categories of the data. As a case study, **S1** applied the slicing approach to explore the gender-based differences in MOOC dropout prediction models. The study proposed a new fairness metric, ABROCA, the *Absolute Between-ROC Area*. This metric measures the absolute value of the area between the baseline group ROC curve ROC_b and the comparison group(s) ROC_c . The lower the ABROCA value, the less unfair the algorithm.

$$ABROCA = \int_0^1 |ROC_b(t) - ROC_c(t)| dt$$

Unlike fairness metrics like equalized odds, ABROCA is not threshold dependent. The metric reveals model discrimination between subgroups across all possible thresholds t . Also, as pointed out by **S1**, ABROCA evaluates model accuracy without focusing strictly on positive cases, making it relevant for learning analytics where a prediction, such as a student potentially dropping out, is not inherently good or bad, but rather used to inform future support actions. Another advantage of the metric is its minimal computational requirement, as it can be calculated directly from the results of a prediction modeling task (**S1**). Indeed, some of the relevant studies identified in this SLR adopted ABROCA; while **S6** (Sha et al., 2021) used ABROCA to assess fairness of students' forum post classification, **S2** (Sha et al., 2022) used the metric to assess fairness of student dropout prediction, student performance prediction, and forum post classification.

2. Group Difference in Performance (**S4, S8, S9, S10 & S11**)

Four of the selected studies measured fairness via group difference in performance. **S4** (Yu et al., 2021) used the difference between the Accuracy, Recall and TNR of its AWARE model (i.e., sensitive features present as predictors) and the BLIND model (sensitive features absent). Similarly, **S9** (Lee & Kizilcec, 2020) used demographic parity as one of its fairness metrics while **S10** (Anderson et al., 2019) measured the optimality of models via a comparison between the AUC ROC gained or lost when separate models were built for each sub-population. Meanwhile, **S11** (Loukina et al., 2019), the approach adopted involved three dimensions of fairness namely overall score accuracy; overall score difference; and conditional score difference.

3. True Positive Rate and True Negative Rate (**S5 & S9**)

S5 (Jiang & Pardos, 2021) examined race group fairness by reporting true positive rate (TPR), true negative rate (TNR), and accuracy according to equity of opportunity and equity of odds. Similarly, **S9** (Lee & Kizilcec, 2020) used equality of opportunity and positive predictive parity.

4. False Positive Rate and False Negative Rate (**S7 & S10**)

For **S7** (Yu et al., 2020), performance parity across student subpopulations was the approach to fairness measurement. The study computed disparity metrics and tested the significance of the disparities using one-sided two proportion z-test.

Higher ratios indicate a greater degree of unfairness. The disparity metrics are defined as follows:

$$\begin{aligned}\text{Accuracy disparity} &= acc_{ref}/acc_g \\ \text{FPR disparity} &= fpr_g/fpr_{ref} \\ \text{FNR disparity} &= fnr_g/fnr_{ref}\end{aligned}$$

where g is the disadvantaged student group and ref is the reference group.

Also, **S10** (Anderson et al., 2019) assessed the equity of graduation prediction models using False Positive Rates and False Negatives.

5. Others (**S3** & **S12**)

S3 (Wang et al., 2022) and **S12** (Kusner et al., 2017) took different unique approaches in assessing the fairness of students' career recommender system and prediction of success in law school. While **S1** used non-parity unfairness, UPAR, **S12** used counterfactual fairness. According **S3**, non-parity unfairness, UPAR, computes the absolute difference of the average ratings between two groups of users.

$$U_{PAR} = |E_g[y] - E_{\neg g}[y]|$$

where $E_g[y]$ is the average predicted score for one group (e.g. Male), and $E_{\neg g}[y]$ is the average predicted score for the other group (e.g. Female), the lower the U_{PAR} , the fairer the system.

Meanwhile, **S12**, focusing on predicting success in law school, proposed Counterfactual Fairness. This approach is rooted in the idea that a predictor is considered counterfactually fair if, for any given context, altering a protected attribute (like race or gender) doesn't change the prediction outcome when all other causally unrelated attributes remain constant. Formally, let A denote protected attributes and the remaining attributes as X , while Y denotes the desired output. Given a causal model represented by (U, V, F) where V is the union of A and X , **S12** postulated the following criterion for predictors of Y :

A predictor Y is said to be counterfactually fair if, for any specific context defined by $X=x$ and $A=a$, the following condition holds:

$$P(Y_{A \leftarrow a}(U) = y | X = x, A = a) = P(Y_{A \leftarrow a'}(U) = y | X = x, A = a)$$

for every possible outcome y and for any alternative value a' that the protected attribute A can assume.

Clearly, this SLR reveals the diverse nature of metrics for measuring education algorithmic fairness with common methods being ABROCA (**S1**, **S2** & **S6**), true positive rate and true negative rate (**S5** & **S9**), false positive rate, false negative rate, and disparity metrics (**S7** & **S10**) as well as group difference in performance (**S4**, **S8**, **S9**, **S10** & **S11**). Other methods include non-parity unfairness, UPAR (**S3**) and counterfactual fairness (**S12**).

Each method comes with its unique strengths and limitations, making it clear that there is no one-size-fits-all approach to fairness assessment. For instance, ABROCA is advantageous in scenarios where fairness needs to be evaluated independent of decision thresholds, offering a more holistic view of model discrimination. On the other hand, ABROCA might not be as effective in scenarios where specific decision thresholds are crucial, and this is exactly where measures like equalized odds do well. Also, counterfactual fairness, which ensures that predictions don't change when altering protected attributes like race or gender, is beneficial in contexts where direct discrimination needs to be explicitly addressed. However, this approach might be less suitable in situations where indirect bias or correlation with non-protected attributes is the primary concern.

In a nutshell, our review emphasizes the need for a nuanced understanding of research and application contexts highlighting that the choice of fairness metric should be aligned with the specific goals, characteristics, and ethical considerations of each educational algorithm.

Research Objective 2: Sensitive Features and Bias Mitigation Strategies

Sensitive Features: Beyond Gender and Race

This review reveals that most of the studies in the existing literature on education algorithmic fairness primarily concentrate on gender and race as sensitive features. As shown in Table 4, ten papers (**S1**, **S2**, **S3**, **S4**, **S6**, **S7**, **S8**, **S9**, **S10**, & **S12**) of the 12 papers reviewed used gender as sensitive feature while seven papers (**S4**, **S5**, **S7**, **S8**, **S9**, **S10**, & **S12**) used race as a sensitive feature. For instance, **S1** (Gardner et al., 2019) performed a slicing analysis along student gender alone, with the authors acknowledging the need for future work to consider multiple dimensions of student identity. Also, in its research on the effectiveness of class balancing techniques for fairness improvement, **S2** (Sha et al., 2022) focused only on student gender groups, with the study noting as part of its limitations, the need to consider other demographic attributes like first language and educational background. Similarly, the debiased college major recommender system by **S3** (Wang et al., 2022) mitigates only gender bias. Meanwhile, the study acknowledged that gender bias is only one of the biases that can harm career's choices, emphasizing the importance of addressing other types of biases in future studies.

Also, **S5** (Jiang & Pardos, 2021) primarily focused on race and did not thoroughly analyze potential biases that may have been introduced by other protected attributes such as parental income. In a similar vein, **S6** (Sha et al., 2021) concentrated only on gender and first language while **S8** (Hu & Rangwala, 2020) and **S9** (Lee & Kizilcec, 2020) focused only gender and race.

Notable exceptions are **S4** (Yu et al., 2021) and **S7** (Yu et al., 2021), which examined four sensitive features each: gender, race, first-generation student, and financial need. Meanwhile, only one paper **S11** (Loukina et al., 2019) considered disability. It is the same for native language, as only **S6** (Sha et al., 2021) considered it as a sensitive feature.

Table 4 Sensitive features used in selected studies

Paper	Dataset	Sensitive features				
		Sex	Race	Status*	Disability	Language
S1	Data from 44 MOOCs, 3,080,230 students in training & 1,111,591 in the test set.	✓				
S2	Moodle Data, <i>KDDCUP (2015)</i> and <i>OULA</i> Kuzilek et al. (2017)	✓				
S3	Dataset with information from 16,619 anonymised Facebook users	✓				
S4	81,858 online course-taking records for 24,198 students and 874 unique courses	✓	✓	✓		
S5	82,309 undergrads with 1.97 m enrollments and 10,430 courses from UC-Berkeley	✓	✓			
S6	3,703 randomly-selected discussion posts by Monash University students	✓				✓
S7	2,244 students enrolled in 10 online, STEM courses taught from 2016 to 2018	✓	✓	✓		
S8	A 10-year data at George Mason University from 2009 to 2019. Covering top five majors.	✓	✓			
S9	Admin data of 5,443 students from Fall 2014 to Spring 2019 from a US university	✓	✓			
S10	Data covering 14,706 first-time in US college undergrads, 2006–2012 (inclusive)	✓	✓			
S11	A corpus included 26,710 responses from 4,452 test-takers				✓	
S12	A survey data covering 163 law schools and 21,790 law students in the US	✓	✓			

Status* means social-economic status

Mitigating Bias in Education Algorithms

The papers included in this SLR leveraged various fairness strategies to debias education algorithms as shown in Table 5.

S2, S4, S6 and **S9** used **class balancing techniques** to improve fairness. In particular, **S2** (Sha et al., 2022) examined the effectiveness of 11 class balancing techniques in improving the fairness of three representative prediction tasks: forum post classification, student dropout prediction, and student performance prediction. The research considered four under-sampling techniques: Tomek's links, Near Miss, Edited Nearest Neighbour, and Condensed Nearest Neighbour; four over-sampling techniques: SMOTE, SMOTE (K-means), SMOTE (Borderline), and ADASYN; and four hybrid techniques (with different combinations of under-sampling and over-sampling techniques). Class balancing was performed on the training set to ensure gender parity. The study revealed that eight of the 11 class balancing techniques improved fairness in at least two of the three prediction tasks. Particularly, TomekLink and SMOTE-TomekLink improved fairness across all the three tasks. Meanwhile, compared to fairness improvement, accuracy was almost exclusively improved by over-sampling techniques.

Similarly, to address the issue of class imbalance, **S4** (Yu et al., 2021) with a focus on student drop-out prediction and **S9** (Lee & Kizilcec, 2020) with a focus on student performance prediction adjusted the sample weights to be inversely proportional to class frequencies during the training stage, ensuring that the model fairly considered both majority and minority classes during the learning process. Also, to address algorithmic bias in students' forum post classification, **S6** (Sha et al., 2021) explored the viability of equal sampling for observed demographic groups (gender and first-language background) in the model training process. The research trained six classifiers with two different training data samples, the original training sample

Table 5 Bias mitigation strategies adopted by reviewed papers

Paper	Bias mitigation strategy
S1	Slicing analysis; proposed ABROCA for fairness measurement
S2	Class balancing techniques
S3	Vector projection-based bias attenuation method
S4	Adjusted sample weights to address class imbalance.
S5	Fairness via unawareness, reweighing, adversarial learning, and inference strategy
S6	Equal sampling across demographic groups
S7	None
S8	Created a model, MCCM which is composed of two classifiers, each of which corresponds to a sensitive attribute
S9	Set different threshold values for each subgroup to achieve equality of opportunity in the testing set.
S10	None
S11	None
S12	Counterfactual Fairness

(all posts after removing the testing data) and the equal training sample (equal number of posts randomly selected for each demographic group). To ensure comparable results, the same testing data was used for evaluation. The study found that most classifiers became fairer for the demographic groups when using equal sampling.

Meanwhile, to mitigate gender bias in a career recommender system, **S3** (Wang et al., 2022) used vector projection-based bias attenuation method. As noted by the study, traditional word embeddings, trained on large data, often inherit racial and gender biases. For instance, vector arithmetic on embeddings can produce biased analogies, such as “*doctor – man + woman = nurse*”. The user embeddings in the career recommender system faced a similar issue. Therefore, the study introduced a debiasing step: given the embedding of a user, p_u , and a unit vector v_B representing the global gender bias in the same embedding space, the study debiased p_u by removing its projection on the gender bias vector v_B . According to **S3**, this bias attenuation method is beyond a simple “fairness through unawareness” technique, as it systematically removes bias related to both the sensitive feature (gender) and the proxy features.

Noting that the decision to include or exclude sensitive features as predictors in models is contentious in literature, **S4** (Yu et al., 2021) set out to address the challenge by exploring the implication of including or excluding protected attributes in student dropout prediction. The study compared BLIND (fairness through unawareness) and AWARE models (fairness through awareness). Fairness levels did not substantially differ between the models, regardless of whether the protected attributes were included or excluded (**S4**).

To minimize racial bias in student grade prediction, **S5** (Jiang & Pardos, 2021) trial several fairness strategies. The first fairness strategy used is “fairness by unawareness” (similar to **S4**) which was made the default (baseline). At the data construction stage, **S5** assigned weights to different race and grade labels to address data imbalance. At the model training stage, the study used adversarial learning strategy. Additional fairness strategies tried include the inference strategy and “alone strategy”. The inference strategy involved adding sensitive feature (race) to the input for training and removing it for prediction while the alone strategy involved training the model separately on each race group. According to **S5**, using weight loss function to balance race reduced the TPR, TNR, and accuracy for all race groups except Pacific Islanders. For all the three metrics used, no single strategy was always the best, however, the inference strategy recorded the most frequent best performance but it is also the worst in terms of fairness. Meanwhile, the adversarial learning strategy achieved the most fair results for all metrics.

S8 and **S12** took unique approaches as they proposed new framework for bias mitigation. Specifically, **S12** (Kusner et al., 2017) proposed counterfactual fairness as discussed in “[Research Objective 1: Measuring the Fairness of Education Algorithms](#)” section. On the other hand, **S8** (Hu & Rangwala, 2020), developed the multiple cooperative classifier model (MCCM) to improve individual fairness in student at-risk prediction. MCCM consists of two classifiers, each corresponding to a specific sensitive attribute (e.g., male or female). When given an individual’s feature vector x_i , both classifiers receive the input. The output of the classifier corresponding to the individual’s sensitive attribute s_i provides the probability of being positive,

while the output of the other classifier offers the probability of being positive if the individual's sensitive attribute were different (i.e. $1 - s_j$). The difference between the two outputs is measured using KL-divergence. Based on the assumption of metric free individual fairness, a prediction is fair when the difference between the two classifiers is negligible. To improve fairness, the research included a term that represents the KL-divergence in the model's objective function as fairness constraint. By minimizing this difference during the model training process and controlling the trade-off between accuracy and fairness with an hyperparameter λ , the model promotes fairness between different sensitive attributes. **S8** compared the proposed MCCM to baseline models such as individually fair algorithms like *Rawlsian Fairness*, *Learning Fair Representation*, and *Adversarial Learned Fair Representation* as well as an algorithm with no fairness constraint, Logistic Regression. The proposed MCCM is the best in mitigating gender bias in student at-risk prediction case study. Though the model was designed to improve individual fairness, it also achieved group fairness, underscoring the high correlation between both. Meanwhile, the LR model was highly biased as there was no fairness constraint imposed on it (**S8**).

In its own research, **S9** (Lee & Kizilcec, 2020) sought to mitigate racial and gender bias in Random Forest model predicting student performance by adjusting threshold values for each demographic sub-group. The study noted that optimizing the model to satisfy equality of opportunity perpetuates unfairness in terms of positive predictive parity and demographic parity emphasizing that it is not possible to satisfy all notions of fairness simultaneously.

The variety of fairness strategies discussed in the reviewed studies, such as class balancing techniques and adjusting sample weights (**S2**, **S4**, **S6** & **S9**), bias attenuation methods (**S3**), fairness through awareness/unawareness (**S4** & **S5**), adversarial learning (**S5**), multiple cooperative classifier models (**S8**) and counterfactual fairness (**S12**), emphasizes the need for a context-specific approach. By considering the unique characteristics and desired outcomes of education algorithms, researchers can ensure that they serve students in an equitable manner, applying the most suitable strategy for each specific scenario.

Fairness of Features and Data Before Fairness of Algorithms

S2, **S6**, **S7**, & **S8** revealed that it is critical to consider the fairness of data and features used in educational algorithms before delving into algorithmic fairness itself. Meanwhile, limited work has been done in evaluating bias in data and feature engineering (**S2**). As noted by **S7** (Yu et al., 2020) and **S8** (Hu & Rangwala, 2020), biased data and features can contribute to the unfairness of predictive algorithms, as they may reflect historical prejudices, demographic inequalities, and cultural stereotypes. According to **S2** (Sha et al., 2022), bias is inherent in most real-world education datasets. All the three datasets: Moodle dataset from Monash University, Xue-TangX *KDDCUP (2015) dataset*, and the Open University Learning Analytics (OULA) dataset exhibited distribution and hardness bias (Kuzilek et al., 2017). 60% of forum posts in the Moodle dataset were authored by female students while over 68% of the KDDCUP dataset were male student data. OULA had the lowest distribution bias with only 9% difference in male and female samples sizes. However, **S2**

noted that using class-balancing techniques reduced the hardness and distribution bias of the datasets.

Another eligible paper, **S6** (Sha et al., 2021), pointed out that extracting fairer features and evaluating feature fairness can prevent algorithms from receiving discriminatory inputs, especially as this information is difficult to detect later in model implementation.

To examine fairness as an attribute of data sources rather than algorithms, **S7** (Yu et al., 2020) evaluated the fairness impact of features from three data sources: institutional data (female, transfer, low income, first-gen, URM, SAT total score and High school GPA were extracted as features), LMS data (total clicks, total clicks by category, total time, and total time by category), and survey data (effort regulation, time management, environment management, and self-efficacy). According to the study, combining institutional and LMS data led to the most accurate prediction college success. On the other hand, features from the institutional data led to the most discriminatory behavior of the model. Meanwhile, features from survey dataset recorded the lowest accuracy while also showing high bias.

Research Objective 3: Fairness-Accuracy Tradeoffs

The trade-offs between fairness and accuracy in educational algorithms are a critical aspect of algorithm design (Pessach & Shmueli, 2022). These trade-offs have been investigated in various studies, however, papers (**S1**, **S2**, **S3**, **S4** & **S6**) in this SLR found no strict trade-off between model accuracy and fairness. Starting with **S1** (Gardner et al., 2019), the study did not notice a strict tradeoff between the performance and discrimination of its MOOC dropout models. Also, **S2** (Sha et al., 2022) found that fairness improvements were complementary with accuracy improvements when applying over-sampling class balancing techniques. Meanwhile, **S3** (Wang et al., 2022) presented an interesting case. A debiased career recommender system was found to be more accurate and fairer than its original biased version when evaluated using the measures of machine learning accuracy and fairness. However, when evaluated through an online user study of more than 200 college students, participants preferred the original biased system over the debiased system (**S3**). The findings of this study highlight that fair algorithms may not meet the expectations of end users, for refusing to confirm existing human bias. In its research, **S4** (Yu et al., 2021) found that including or excluding protected attributes in dropout prediction models did not significantly affect performance metrics like accuracy, recall, and TNR, neither did it lead to different levels of fairness. Similarly, in a study comparing machine learning and deep learning models in delivering fair predictions, **S6** (Sha et al., 2021) found that utilizing techniques such as equal sampling can help reduce model unfairness without sacrificing classification performance.

Meanwhile, a study by **S5** (Jiang & Pardos, 2021) which tried different fairness strategies revealed that adversarial learning achieved the best group fairness without sacrificing much with respect to performance metrics like TPR, TNR, and accuracy. However, another strategy that included sensitive attribute

(race) most frequently scored the highest for the performance metrics, but it was also the worst in terms of fairness, indicating the existence of fairness-accuracy tradeoff.

Conclusion and Future Research

In this systematic literature review, we analyzed twelve eligible studies on education algorithmic fairness. Our findings highlight the need for researchers to evaluate data and feature bias in the development of fair algorithms (**S2, S6, S7, & S8**). Despite its importance, limited studies currently address this aspect. Additionally, the current body of research on education algorithmic fairness predominantly focuses on gender and race. Meanwhile, in a review by Baker and Hawn (2021), there are available evidence of bias in other areas like military-connected status, disability, and socioeconomic status. Another takeaway from this SLR is that no one-size-fits-all solution exists for assessing the fairness of education algorithms, highlighting the importance of selecting metrics that effectively capture the nuances of one's context or application. Similarly, our analysis indicates that there is no strict trade-off between fairness and accuracy in education algorithms, as the relationship between the two can be complex and context dependent.

Based on the gaps identified, we make the following recommendations for future research to advance the state of the art:

1. **Prioritize Fairness of Data and Features:** As indicated by **S2, S6, S7, & S8**, it is critical to assess the fairness of data and features used in educational algorithms before delving into algorithmic fairness itself. This is essential to prevent algorithms from receiving discriminatory inputs.
2. **Broaden the scope of education algorithmic fairness studies** by incorporating demographic attributes such as socioeconomic status and disability when analyzing sensitive features. This SLR revealed that most papers in the existing literature focus primarily on gender and race as sensitive features. Specifically, ten papers (**S1, S2, S3, S4, S6, S7, S8, S9, S10, & S12**) out of the 12 papers in this review used gender as a sensitive feature while seven papers (**S4, S5, S7, S8, S9, S10, & S12**) used race as a sensitive feature. In contrast, only one paper each examined fairness related to native language (**S6**) and disability (**S11**). Notably, none of the eligible studies in this SLR examined military-connected status or age.
3. **Consider End User's Perspective:** As reported by **S3**, the expectations of end users may not align with algorithm's fairness measures. Therefore, investigating user preferences can help bridge the gap between machine learning fairness and human perceptions.
4. **Harmonize Fairness with Accuracy:** When exploring novel debiasing procedures for algorithms, focus on taking advantage of the complementary nature of fairness and accuracy, rather than compromising one for another as papers (**S1, S2, S3, S4 & S6**) in our review indicate there is no strict tradeoff between the two.

Appendix

Table 6 Data extracted from eligible papers

Ref	Study design	Dataset	Methods	Evaluation metrics	Results	Conclusions	Limitations
S1 Gardner et al. (2019)	Proposed “ABROCA” a new method for unfairness measurement in predictive student models. ABROCA meas- ures the absolute value of the area between the baseline group ROC curve ROCb (e.g. Male) and the comparison groups ROCc (e.g. Female). Replication study	- Raw data from 44 MOOCs. - 3,080,230 students in the training dataset - 1,111,591 students in the test set.	- Slicing Analysis - Feature extrac- tion - Hyperparameter optimization - 5 ML models for experiment	- ABROCA for fairness measure- ment = $\int_0^1 ROC_b - ROC_c dt $ - AUC-ROC for accuracy	Mean for 44 MOOCs: ABROCA LSTM = 0.0351 LR = 0.0353 SVM = 0.0311 Naïve- Bayes = 0.0177 CART = 0.0280 AUC LSTM = 0.653 LR = 0.669 SVM = 0.693 Naïve- Bayes = 0.558 CART = 0.715	- ABROCA offers more insights into model discrimination than equalized odds. - Model discrimi- nation is related to gender and course cur- ricular. - No strict tradeoff between model performance and discrimina- tion	- Slicing was done only along gender as sensitive feature. - Focused only on measuring fair- ness, nothing on mitigation.

Table 6 (continued)

Ref	Study design	Dataset	Methods	Evaluation metrics	Results	Conclusions	Limitations
S2	Sha et al. (2022)	Investigated the impact of 11 class balancing techniques on fairness of three representative prediction tasks: Forum post classification, student dropout prediction and student performance prediction.	Three datasets used: - Moodle dataset: 3,703 forum discussion posts by Monash University students - <i>KDDCUP (2015) dataset</i> with 39 courses and 180,785 students from online learning platform Xue-Tang X. - Open University Learning Analytics dataset with 24,806 students' interactions in the VLE on 22 courses. 80/20% Train/Test split All 3 datasets are publicly available.	- 22 features extracted for KDDCUP dropout prediction; CFIN model (i.e. DNN + CNN). - 112 features extracted for OULA performance prediction; GridSearch and Random Forest model implemented. - Forum post classification task involves BERT embedding of posts; CLSTM (CNN + LSTM) model implemented using Tensorflow. - Class balancing techniques used include four under-sampling, four over-sampling and three combined sampling techniques.	Distribution and hardness bias to measure data bias. AUC for predictive accuracy ABROCA for predictive fairness Best results are stated: Hardness bias: 0.001 for Forum post task, 0.014 for KDDCUP, and 0.01 for OULA. AUC: 0.844, 0.89, and 0.802 respectively for Forum post, KDDCUP, and OULA tasks. ABROCA: 0.055, 0.019, and 0.018 respectively for Forum post, KDDCUP, and OULA tasks.	8 out of the 11 CBTs improved fairness in at least two out of the three tasks Over-sampling techniques methods improved both fairness and accuracy. No tradeoff between fairness and accuracy.	Addressed only the data aspect of algorithmic bias but bias may also occur during annotation of a training sample, feature engineering, and model training. Focused only on student gender groups. Need to explore other demographic attributes.

Table 6 (continued)

Ref	Study design	Dataset	Methods	Evaluation metrics	Results	Conclusions	Limitations
S3	Studied human-fair-AI interaction: 1. Case study – developed a debiased recommender system that mitigates gender bias in college major recommendations 2. Conducted an online user study with over 200 student participants.	<ul style="list-style-type: none"> - Dataset with information from 16,619 anonymised Facebook users; 60% female and 40% male, 1,380 unique college majors, 143,303 unique items that a user can like on Facebook and 3.5 million + user-item pairs. - Online user study with 202 participants: 48% females and 49.5% males. Only 18.3% open to career suggestions 	<ul style="list-style-type: none"> - Constructed user embedding via Neural collaborative filtering (NCF) model for predicting the items a user “likes”. - Then, trained a Logistic Regression classifier to predict top college majors using the embedding as input features. - Vector projection-based bias attenuation method to gender-debias the recommendations. - Train-test split: 70/30% 	<ul style="list-style-type: none"> - Normalized discounted cumulative gain at K Higher NDCG@K means better accuracy - Non-parity unfairness. Lower UPAR means fairer system. - T-test for user acceptance study 	<ul style="list-style-type: none"> GenderDebiased: NDCCG@3: 0.005 NDCCG@10: 0.01 NDCCG@20: 0.013 $U_{PAR} = 1.1188$ Mean t-test acceptance score: 0.279 GenderAware: NDCCG@3: 0.0009 NDCCG@10: 0.005 NDCCG@20: 0.007 $U_{PAR} = 1.1445$ Mean t-test acceptance score: 0.372 	<ul style="list-style-type: none"> - ML evaluation showed that the gender-debiased recommender is fairer and more accurate than gender-aware recommender but online user study showed that participants prefer the biased system. - For ML evaluation, Gender-debiased recommender is fairer without any loss of prediction accuracy. 	<ul style="list-style-type: none"> - 80% of the participants are not open to new career suggestions as most are university students. - Focused only on applicants but career-related bias exist also in college admission decision makers.

Table 6 (continued)

Ref	Study design	Dataset	Methods	Evaluation metrics	Results	Conclusions	Limitations
S4	Yu et al. (2021) Investigated whether sensitive features should be included in model Case study: College drop-out prediction with or without sensitive features (gender, URM, first-generation student and high-financial need).	Sample contains 81,858 online course-taking records for 24,198 unique students and 874 unique courses; and 564,104 residential course-taking records for 93,457 unique students and 2,877 unique courses.	Extracted 58 features. AWARE experiment included all features while BLIND experiment excluded four protected attributes. LR and Gradient Boosting were implemented. Train-Test split done by cohort: Training set has 17,259 online and 79,182 residential students while test set has 6,939 online and 14,275 residential students. 5-fold CV Grid-Search, indicator variables for missing values, feature scaling, and adjusted sample weights to address class imbalance.	- Accuracy - Recall - True Negative Rate (TNR) Measure fairness via Group difference in performance metric	GBT Accuracy, Recall and TNR slightly higher by 0.2, 0.1, and 0.1% in AWARE model compared to BLIND for online students. No difference for residential. LR Accuracy, Recall and TNR slightly lesser by 0.2, 0.1, and 0.1% in AWARE compared to BLIND for online students. Difference not statistically significant ($p < 0.1$)	Including or excluding protected attributes does not generally affect fairness in terms of any metric in any enrollment format. Performance do not differ significantly between the BLIND and AWARE models. GBT slightly performed better than LR.	Did not investigate the impact of counterfactual protected attributes on fairness.

Table 6 (continued)

Ref	Study design	Dataset	Methods	Evaluation metrics	Results	Conclusions	Limitations
S5 Jiang and Pardos (2021)	Trial different fairness strategies for a grade prediction algorithm with respect to race. Fairness strategies were applied at three stages of the model pipeline: data construction, model training, and inference (prediction).	Anonymised students' enrolment data from UC Berkeley. 82,309 undergraduates with a total of 1.97 million enrollments and 10,430 unique courses.	<ul style="list-style-type: none"> - LSTM model implemented. - Fairness Strategies: <ol style="list-style-type: none"> 1) Fairness through unawareness as baseline. 2) Data construction stage: Added race and other features; assigned weights to balance race and grade labels. 3) Model training stage: Adversarial learning 4) Prediction stage: inference strategy, i.e. remove sensitive feature (race) for prediction. - Training, validation, and test data are in the ratio 13:1:1. 	Accuracy, True Positive Rate, and True Negative Rate reported according to equity of opportunity and equity of odds.	<p>Adversarial learning strategy has best group fairness as follows:</p> <p>TPR: Range = 9.48 & STD = 3.8</p> <p>TNR: Range = 8.76 & STD = 3.45</p> <p>Accuracy: Range = 1.5 & STD = 0.62</p> <p>NB: If group fairness is fully attained, Range and STD will be 0.</p>	<p>Using weight loss function to balance race reduced the TPR, TNR, and accuracy for all race groups except Pacific Islanders.</p> <p>Inference strategy recorded the best performance, but it is the worst in terms of fairness.</p> <p>Adversarial learning strategy achieved the fairest results for all metrics.</p>	Focused on race; did not analyze bias that may be introduced by other sensitive features like gender and parental income.

Table 6 (continued)

Ref	Study design	Dataset	Methods	Evaluation metrics	Results	Conclusions	Limitations
S6	Assessed algorithmic fairness in classification of forum posts. Fairness assessed along demographic groups (sex and English as first or second language).	3,703 randomly-selected discussion posts made by Monash University students including 2,339 (63%) content-relevant posts and 1,364 (37%) content-irrelevant posts. Dataset is publicly available.	<ul style="list-style-type: none"> - Evaluated four traditional ML models: SVM, Random Forests, Logistic regression, and Naïve Bayes and two DL models: Bi-LSTM and CNN-LSTM. - Text embeddings generated using the tool Bert-as-service. - Trained classifiers with two samples: original sample and balanced sample (along gender and first language). - Training, validation, and test set in ratio 70:10:20. 	Accuracy, Cohen's κ , AUC, and F1 score for performance evaluation ABROCA for fairness evaluation.	<ul style="list-style-type: none"> - CNN-LSTM has highest performance: 0.795 AUC for original sample; 0.792 and 0.802 AUC for balanced sample along gender and first language respectively. - Best fairness in original sample: ABROCA of 0.007 for gender, and 0.032 for language group. - In balanced sample: ABROCA of 0.003 for sex group and 0.012 for language. 	<ul style="list-style-type: none"> - Traditional ML models provided fairer results while DL models yielded better performance. - Strict performance-for-fairness trade-off is not evident. - Equal sampling across demographic groups reduced model unfairness. 	Focused only on two demographic groups: gender and first language.

Table 6 (continued)

Ref	Study design	Dataset	Methods	Evaluation metrics	Results	Conclusions	Limitations
S7	Yu et al. (2020) Fairness and performance evaluation of the utility of different data sources in predicting short and long-term student success. Data sources examined include LMS data, institutional data and survey data.	Sample students enrolled in 10 online, introductory STEM courses taught from 2016 to 2018 at a large, public research US university. Original dataset has 2,244 students. After data cleaning, the final sample size was 2,093.	Institutional features included student demographics and academic achievement prior to college. Click features were derived from the LMS data. Survey features included four constructs of self-regulated learning skills and self-efficacy from pre-course surveys. SVM, LR, and RF models were implemented.	Accuracy, False positive rate (FPR), False negative rate (FNR). Fairness measured via performance parity.	The combination of institutional and LMS data gave best accuracy ($\Delta = 0.052$, $p < 0.001$ for short term and $\Delta = 0.037$, $p = 0.014$ for long term). Institutional data consistently underestimates historically disadvantaged student groups while LMS data overestimate some of these groups more often.	Survey data is neither fair nor accurate. Inclusion of institutional features seemed to produce the most unfair outcome. All data sources exhibit bias.	Scope of the feature sets was limited and not representative of the full potential of different data sources.

Table 6 (continued)

Ref	Study design	Dataset	Methods	Evaluation metrics	Results	Conclusions	Limitations
S8 Hu and Rangwala (2020)	Proposed Multiple Cooperative Classifier Model to improve individual fairness in at-risk student prediction.	A 10-year data at George Mason University from Fall 2009 to Fall 2019. Covering top five majors. A course is chosen only if at least 300 students have taken it.	<ul style="list-style-type: none"> - Created a model, MCCM which is composed of two classifiers, each of which corresponds to a sensitive attribute, e.g., male or female. - Used Rawlsian Fairness, LR, Learning Fair Representation, and Adversarial Learned Fair Representation as baselines. 	Accuracy, Discrimination and Consistency.	The proposed model MCCM recorded the best performance in mitigating bias in terms of discrimination.	Though designed for individual fairness, the proposed MCCM achieved both group fairness and individual fairness, because group and individual fairness are highly correlated.	Focused only on gender and race groups.

Table 6 (continued)

Ref	Study design	Dataset	Methods	Evaluation metrics	Results	Conclusions	Limitations
S9 Lee and Kizilcece (2020)	<ul style="list-style-type: none"> - Predicted student success based on university administrative records. - Applied post-hoc adjustments to improve model fairness. 	<ul style="list-style-type: none"> - Student-level administrative data spanning Fall 2014 to Spring 2019 was obtained from a US Research University. - The final processed data has 5,443 instances and 56 features. - Duplicates, missing course grades, courses taken multiple times, and grades other than letter (A-F) or pass/fail grades were removed. 	<ul style="list-style-type: none"> - Random Forest model was implemented. - Weighted each instance with the inverse of its label proportion to achieve label balance in the training set. - Indicator value of .999 imputed for missing standardized test scores. - To improve fairness, different threshold values for each subgroup was set such that equality of opportunity is achieved in the testing set. 	<ul style="list-style-type: none"> - Equality of opportunity, Demographic parity and Positive predictive parity 	<ul style="list-style-type: none"> - Overall accuracy of the model on the test data is 0.73 with an F-score of 0.80. - After correction for equality of opportunity via threshold adjustment for each group, the subgroup accuracy remains similar for both groups. 	<ul style="list-style-type: none"> - Optimizing the model to satisfy equality of opportunity perpetuates unfairness in terms of demographic parity and positive predictive parity for both gender and racial-ethnic groups. - Model exhibits gender and racial bias in two out of three fairness measures considered. 	<ul style="list-style-type: none"> - Analysis focused on two binary protected attributes by racial-ethnicity and gender alone.

Table 6 (continued)

Ref	Study design	Dataset	Methods	Evaluation metrics	Results	Conclusions	Limitations
S10 Anderson et al. (2019)	Assessed the equity and optimality of graduation prediction models.	Data from public, US R1 research. Includes 14,706 first-time in college undergrads, admitted in Fall semesters between 2006–2012 (inclusive).	<ul style="list-style-type: none"> - Feature set covers academic performance, financial information, pre-admission information, and extra-curricular activities. - Linear Kernel SVM, Decision Trees, Random Forests, Logistic Regressions, and Stochastic Gradient Descent classifier were implemented. - Model parameters selected via 5-fold CV Grid Search. Train-test split: 80:20%	Equity of models measured via false positive and false negative predictions. AUC-ROC for performance	Decision Tree = 0.798 SVM = 0.805 Logistic Regression = 0.807 Random Forest = 0.800 SGD = 0.814	<ul style="list-style-type: none"> - No model, or population, saw a meaningful change in performance when trained only on one population. - Models are not perfectly equitable. 	Need to investigate if the model implementation affects student outcomes, and whether the slight unfairness observed translate to real-world differences.

Table 6 (continued)

Ref	Study design	Dataset	Methods	Evaluation metrics	Results	Conclusions	Limitations
S11 Loukina et al. (2019)	Explored multiple dimensions of fairness. Case study: Simulated and real data to assess impact of English language proficiency test-takers' on automated native language scores.	<ul style="list-style-type: none"> - Created a corpus with uniform distribution of native languages by randomly sampling a similar number of test-takers for each version of the test. - The final corpus included 26,710 responses from 4,452 test-takers (742 for each language). 	<p>Four different simulated models implemented: random model, perfect model, an almost perfect model, and metadata-based model.</p> <p>Train-test split: 75:25%</p>	<p>Three dimensions of fairness:</p> <p>Overall score accuracy: Overall score difference; and Conditional score difference</p>	<ul style="list-style-type: none"> - The speakers of all languages were disadvantaged by the META model. - GER speakers were underscored while JPN speakers were overscored by the perfect model. - On actual models, the model trained separately for each native language is most fair in terms of overall score differences. 	<ul style="list-style-type: none"> - Total fairness may not be achievable and different definitions of fairness may require different solutions. - None of three definitions of fairness is, in principle, more important than another. 	<p>Considered human scores to be the true 'gold standard' measure of language proficiency, however, human scores are likely to contain a certain amount of error and bias.</p>

Table 6 (continued)

Ref	Study design	Dataset	Methods	Evaluation metrics	Results	Conclusions	Limitations
S12 Kusner et al. (2017)	Developed a framework for modeling fairness using tools from causal inference.	A survey data covering 163 law schools and 21,790 law students in the United States with information on LSAT score, GPA, etc., Provided by Law School Admission Council	Counterfactual fairness defined as: a decision to an individual is fair if it is the same in the actual world and a counterfactual world where the individual belonged to a different demographic group.	RMSE for accuracy Counterfactual fairness for fairness	Full and unaware models are fair w.r.t. to sex as they have very weak causal link between sex and GPA/LSAT.	The proposed counterfactual fairness framework takes protected attributes into account in modelling.	Implemented only for one use case

Funding The author gratefully acknowledges support from Chevening Scholarship, the UK government's global scholarship programme, funded by the Foreign, Commonwealth, and Development Office (FCDO).

Declarations

Competing Interests The author has no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Anderson, H., Boodhwani, A., & Baker, R. S. (2019). Assessing the Fairness of Graduation Predictions. In *EDM*.
- Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32, <https://doi.org/10.1007/s40593-021-00285-9>
- Corbett-Davies, S., Gaebler, J., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312), 1–117.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226).
- Gajane, P., & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*. Retrieved from <https://arxiv.org/abs/1710.03184>
- Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the Fairness of Predictive Student models through Slicing Analysis. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 225–234. <https://doi.org/10.1145/3303772.3303791>.
- Gedrimiene, E., Celik, I., Mäkitalo, K., & Muukkonen, H. (2023). Transparency and trustworthiness in user intentions to follow career recommendations from a learning analytics tool. *Journal of Learning Analytics*, 10(1), 54–70.
- Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, 11(2), 181–217.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)* (pp. 3323–3331). Curran Associates Inc.
- Hu, Q., & Rangwala, H. (2020). Towards fair educational data mining: A case study on detecting at-risk students. International Educational Data Mining Society.
- Jiang, W., & Pardos, Z. A. (2021). Towards equity and algorithmic fairness in student grade prediction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 608–617).
- Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)* (pp. 4069–4079). Curran Associates Inc.
- Kuzilek, J., Hlostá, M., & Zdrahal, Z. (2017). Open University Learning Analytics dataset Sci. Data, 4, 170171. <https://doi.org/10.1038/sdata.2017.171>
- Lee, H., & Kizilcec, R. F. (2020). Evaluation of fairness trade-offs in predicting student success. *arXiv preprint arXiv:2007.00088*. Retrieved from <https://arxiv.org/abs/2007.00088>
- Loukina, A., & Buzick, H. (2017). Use of automated scoring in spoken language assessments for test takers with speech impairments. *ETS Research Report Series*, 2017(1), 1–10.

- Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 1–10).
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, 88.
- Patterson, C., York, E., Maxham, D., Molina, R., & Mabrey, P. (2023). Applying a responsible innovation framework in developing an equitable early alert system:: A case study. *Journal of Learning Analytics*, 10(1), 24–36.
- Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3), 1–44.
- Rets, I., Herodotou, C., & Gillespie, A. (2023). Six practical recommendations enabling ethical use of predictive learning analytics in distance education. *Journal of Learning Analytics*, 10(1), 149–167. <https://doi.org/10.18608/jla.2023.7743>
- Sha, L., Raković, M., Das, A., Gašević, D., & Chen, G. (2022). Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Transactions on Learning Technologies*, 15(4), 481–492.
- Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, V. M., Gasevic, D., & Chen, G. (2021). Assessing algorithmic fairness in automatic classifiers of educational forum posts. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part I 22* (pp. 381–394). Springer International Publishing.
- Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, 42(9), 1568–1580. <https://doi.org/10.1016/j.respol.2013.05.008>
- Suresh, H., & Gutttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1–9).
- Wang, C., Wang, K., Bian, A., Islam, R., Keya, K. N., Foulds, J., & Pan, S. (2022). Do Humans Prefer Debaised AI Algorithms? A Case Study in Career Recommendation. In *27th International Conference on Intelligent User Interfaces* (pp. 134–147).
- Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes? In *Proceedings of the Eighth ACM Conference on Learning@ Scale* (pp. 91–100).
- Yu, R., Li, Q., Fischer, C., Doroudi, S., & Xu, D. (2020). Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data. *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, 292–301.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.