The Journal of Systems and Software 219 (2025) 112244

Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jss

Fairness for machine learning software in education: A systematic mapping study☆

Nga Pham [a,b], Pham Ngoc Hung [b], Anh Nguyen-Duc [c,*]

[a] Faculty of Information Technology, Dainam University, Hanoi, Viet Nam
[b] Faculty of Information Technology, VNU University of Engineering and Technology, Hanoi, Viet Nam
[c] Faculty of Information Technology, University of South Eastern Norway, Bø I Telemark, Norway

ARTICLE INFO

Keywords:
Fairness
Bias
AI
Definitions of fairness
AI software fairness

ABSTRACT

The integration of machine learning (ML) systems into various sectors, notably education, has great potential to transform business workflows and decision-making processes. However, this technological advancement brings forth critical ethical concerns, particularly concerning the fairness of decisions affecting diverse groups of people. Our objective was to systematically map out the landscape of ML fairness research in higher education by exploring seven key research questions. These questions span a range of topics from the types of ML algorithms used in education to the methods of fairness assessment and the results achieved in terms of equity. We included 63 primary studies published between 2002 and 2023. The most common setting for AI Fairness research are: traditional machine learning algorithms (Logistic Regression, Random Forest, Decision Tree), sensitive variables (gender, race, ethnicity), and various definitions of fairness (Group fairness, Demographic parity, Equalized odds). We also identify several future research directions, including fairness assurance for multiple sensitive variables, combining different fairness concepts and metrics, open-source benchmarking tools, and fairness testing for modern ML/AI models.

1. Introduction

Educational technologies increasingly use data and predictive models to provide support and analytical insights to students, instructors, and administrators (Chen et al., 2020a). Artificial Intelligence (AI)-enabled systems have the potential to transform educational activities for teachers, students, and administrators (Zhang and Aslan, 2021), providing personalized and efficient learning experiences for students. For teachers, AI can help automate the grading and assessment process, freeing them to focus on other aspects of teaching and learning. It can also enable the use of adaptive learning systems, which can adjust the difficulty and content of material based on a student's progress. Students can receive recommendations on lectures, classes, study programs, and even career development. This can help students learn more effectively and efficiently and can also help to identify and address any learning gaps or difficulties.

To realize the scenarios above in a real-world setting, several ethical issues with operating AI systems need to be addressed, including bias or unfair behaviors created by the systems (Du et al., 2021). The concept of fairness itself is subjective and context-dependent. From a requirement engineering perspective, different stakeholders may have different definitions of fairness, making it challenging to define, design, develop and test fairness attributes for AI systems (Mehrabi et al., 2019). From a technical perspective, AI systems are designed to learn and make decisions based on vast amounts of data, which can introduce biases and perpetuate inequalities if not carefully curated. The algorithms used in ML/AI software can inadvertently reflect the biases present in the data, leading to unfair or discriminatory outcomes. To understand and address these biases, a deep understanding of both the technical aspects of AI systems and the societal implications of their decisions is required.

Research interest in exploring, measuring, and ensuring AI fairness has grown rapidly in recent years. To name a few, Fair Machine Learning (Fair ML) is a dedicated research theme on the development of fairness-aware learning algorithms to avoid discrimination against minorities (Mehrabi et al., 2019). Since 2016, Explainable AI (XAI) has become popular as a research area, focusing on developing methods and techniques to make AI systems more transparent and understandable to humans (Došilović et al., 2018; Angelov et al., 2021). In 2018, AI Fairness 360 was introduced to the research community as an open-source toolkit offering a set of metrics for datasets, models and

☆ Editor: Nicole Novielli.

* Corresponding author.
E-mail address: angu@usn.no (A. Nguyen-Duc).

https://doi.org/10.1016/j.jss.2024.112244
Received 19 January 2024; Received in revised form 17 August 2024; Accepted 5 October 2024
Available online 11 October 2024

algorithms to mitigate bias (Bellamy et al., 2019). To advance research in a fast-moving area like AI fairness, it is important to understand the current landscape of research, both theoretically and empirically, to identify unanswered questions and potential areas of improvement. By staying informed about AI fairness research, we can actively contribute to the conversation and engage in meaningful discussions with other researchers, policymakers, and stakeholders.

Fairness is especially important for systems developed for education sectors. This phenomenon already has a long history in the field of education; scholars have studied inequalities and inequities in educational opportunities and outcomes, such as school segregation and achievement gaps (Engberg, 2004; Hughes, 2013; Huston, 2006; Mahmud, 2020; Minnaert and Janssen, 1997; Verdonk et al., 2009). The presence of achievement gaps can be understood as a shortcoming of fairness in educational outcomes, especially if it is the result of discriminatory behavior. It is unfair if students from low-income families score lower test scores due to the lack of access to study resources available to high-income families, but it is especially unfair if they score lower because their teacher or an algorithmic scoring system is biased against them (Chen et al., 2020a; Zhang and Aslan, 2021; Mahmud and Gagnon, 2023; Zhai et al., 2021; Pessach and Shmueli, 2022).

To the end of 2023, there have been several secondary studies and surveys on AI fairness (Du et al., 2021; Došilović et al., 2018; Angelov et al., 2021; Baker and Hawn, 2021; Hutchinson and Mitchell, 2019; Kleinberg et al., 2018; Hort et al., 2024; Chen et al., 2024; Soremekun et al., 2022; Wan et al., 2023; Caton and Haas, 2020; Tang et al., 2023), or secondary works reviewing research about AI in education (Chen et al., 2020a; Zhang and Aslan, 2021; Mahmud and Gagnon, 2023; Zhai et al., 2021; Pessach and Shmueli, 2022). However, there are a few attempts to review research on AI fairness in education (Baker and Hawn, 2021; Memarian and Doleck, 2023; Kizilcec and Lee, 2021; Casas-Roma and Conesa, 2021). In 2019, Elijah Mayfield et al. conducted research about this (Mayfield et al., 2019). This paper explores culturally relevant pedagogy and other teaching frameworks, identifying future equity work in NLP. They present case studies on intelligent tutoring systems, computer-assisted language learning, automated essay scoring, and classroom sentiment analysis, offering an actionable research agenda. In 2021, Casas-Roma et al. surveyed to explore the intersections between AI, online learning, and ethics to understand the ethical relationships surrounding the integration of AI in online learning environments. Casas-Roma and Conesa (2021). In 2023, Memarian et al. conducted a systematic review of 33 studies about fairness, accountability, transparency, and ethics in higher education (Memarian and Doleck, 2023). The review focuses on classifying studies into qualitative and quantitative research without systematic analysis of definitions, metrics, data, and algorithms of AI models.

The lack of an overview of the research landscape on AI fairness from both conceptual and operational levels in the education context, especially with an updated perspective in the last ten years, has motivated us to conduct a systematic mapping study (Petersen et al., 2015). As fairness research is diverse from both conceptual and operational levels, we want to classify research in many dimensions, from conceptualizing fairness to characteristics of AI models and details of fairness evaluations. This will provide actionable insights for future research to implement and evaluate approaches for AI fairness assurance in education. Furthermore, recent studies have highlighted a lack of empirical rigor in benchmarking bias mitigation approaches and underscored the crucial importance of addressing the accuracy-fairness trade-off (Chen et al., 2023; Hort et al., 2021). To build a foundation for future research, we aimed to explore the state-of-the-art algorithms that tackle the accuracy-fairness trade-off. Deriving from the research objective, we would like to address a general Research Question *What is the state-of-the-art research for conceptualizing, assuring, and evaluating AI fairness in the context of education?* This overall question is broken down into seven Research questions (RQs):

1. RQ1. Which AI algorithms/ approaches are reported in the primary studies in the education context?
2. RQ2. What are the problems investigated in the primary studies?
3. RQ3. What types of fairness are explored in the primary studies?
4. RQ4. What are the characteristics of the dataset used in the primary studies?
5. RQ5. What methods are employed for ensuring AI fairness in primary studies?
6. RQ6. What evaluation metrics are used to assess AI fairness in the primary studies?
7. RQ7. What methodologies are employed to assess fairness and performance in AI/ML models within the primary studies?

To summarize, the contribution of this study is threefold. Firstly, the review provides an in-depth overview of AI fairness and the characteristics reported in the context of education, where fairness is inherent. Secondly, the review presents various technical aspects of AI fairness assurance and their evaluation, which have direct implications for those who want to build a fair AI system for education. Thirdly, the review identifies a current research gap and room for future studies.

The remainder of this paper is structured as follows: Background on fairness in ML/AI is provided in Section 2. Section 3 presents the methodology. Section 4 presents the results. Section 5 discusses the results, and Section 6 concludes the paper.

2. Background

2.1. Definition of AI fairness

The concept of fairness has been debated in philosophy and psychology for a long time, but there is still no consensus on a universally accepted definition. Different perspectives, opinions, and cultural backgrounds can lead to different understandings of what constitutes justice. With the increasing prevalence of artificial intelligence (AI) and machine learning (ML) in various fields, many decisions that affect people are now being made by AI/ML systems. However, these systems are not immune to biases, as they are based on algorithms created by humans. Saxena et al. presented in their study that: “Fairness in the context of decision making can be understood as the absence of any biases or prejudices against an individual or group based on inherent characteristics” (Saxena et al., 2019). Despite this, there is still no clear agreement on what fairness means in the context of AI/ML, and there is no universally accepted concept of fairness in these fields. Some of the most commonly used definitions of fairness in AI/ML were presented by authors Hutchinson et al. and Mehrabi et al. in their studies (Mehrabi et al., 2019; Saxena et al., 2019), including the definitions in Table 1.

Furthermore, when exploring fairness, the issue of bias in AI/ML also needs to be considered. Fairness is often interpreted as the absence of bias, meaning that identifying bias is a key step in determining whether fairness is achieved. In their research, Baker et al. suggest that “By recognizing and rectifying specific biases, whether through theoretical approaches or trial-and-error methods, we move closer to achieving fairness” (Baker and Hawn, 2021). The notion of bias is a recurring theme in contemporary studies. Researchers such as Hutchinson et al. Mehrabi et al. and Baker et al. have offered various definitions of bias (Mehrabi et al., 2019; Baker and Hawn, 2021; Smith, 2020; Saleiro et al., 2019; Suresh and Guttag, 2021). Table 2 presents the common types of biases and their definitions, including historical bias, representation bias, measurement bias, evaluation bias, etc.

Table 1
Different types of fairness.

Type of fairness	Definition	Explanation	Ref.
Equalized Odds	A predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y. $P(\hat{Y}=1\|A=0, Y=y) = P(\hat{Y}=1\|A=1, Y=y)$, $y \in \{0,1\}$	The protected and unprotected groups should have equal rates for true positives and false positives	Zhang and Aslan (2021), Hardt et al. (2016)
Equal Opportunity	A binary predictor $\hat{Y}$ satisfies equal opportunity with respect to A and Y if $P(\hat{Y}=1\|A=0, Y=1) = P(\hat{Y}=1\|A=1, Y=1)$	The protected and unprotected groups should have equal true positive rates	Zhang and Aslan (2021), Hardt et al. (2016), Verma and Rubin (2018)
Demographic Parity	A predictor $\hat{Y}$ satisfies demographic parity if $P(\hat{Y}\|A=0) = P(\hat{Y}\|A=1)$	The likelihood of a positive outcome should be the same regardless of whether the person is in the protected group (e.g., female)	Zhang and Aslan (2021), Verma and Rubin (2018)–Dwork et al. (2011)
Fairness Through Awareness	An algorithm is fair if it gives similar predictions to similar individuals, where	Any two individuals who are similar with respect to a similarity (inverse distance) metric defined for a particular task should receive a similar outcome	Zhang and Aslan (2021), Kusner et al. (2017), Dwork et al. (2011)
Fairness Through Unawareness	An algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process		Zhang and Aslan (2021), Kusner et al. (2017), Grgić-Hlača et al. (2016)
Treatment Equality	Treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories		Zhang and Aslan (2021), Berk et al. (2021)
Test Fairness	A score $S = S(x)$ is testfair (well-calibrated) if it reflects the same likelihood of recidivism irrespective of the individual's group membership, R. That is, if for all values of s, $P(Y=1\|S=s, R=b) = P(Y=1\|S=s, R=w)$	For any predicted probability score S, people in both protected and unprotected (female and male) groups must have an equal probability of correctly belonging to the positive class	Zhang and Aslan (2021), Verma and Rubin (2018), Chouldechova (2016)
Counterfactual Fairness	Predictor $\hat{Y}$ is counterfactually fair if under any context $X=x$ and $A=a$, $P(\hat{Y}_{(A\leftarrow a)}(U)=y\|X=x, A=a) = P(\hat{Y}_{(A\leftarrow a')}(U)=y\|X=x, A=a)$ (or all y and for any value a' attainable by A)	Intuition that a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group	Zhang and Aslan (2021), Kusner et al. (2017)
Fairness in Relational Domains	A notion of fairness that is able to capture the relational structure in a domain—not only by taking attributes of individuals into consideration but by taking into account the social, organizational, and other connections between individuals		Zhang and Aslan (2021), Farnadi et al. (2018)
Conditional Statistical Parity	For a set of legitimate factors L, predictor $\hat{Y}$ satisfies conditional statistical parity if $P(\hat{Y}\|L=1, A=0) = P(\hat{Y}\|L=1, A=1)$	People in both protected and unprotected (female and male) groups should have an equal probability of being assigned to a positive outcome given a set of legitimate factors L	Zhang and Aslan (2021), Verma and Rubin (2018), Corbett-Davies et al. (2017)

2.2. AI/ML in education

As advancements in AI and ML continue, these technologies are increasingly being used in the education field to make decisions that were once made by humans. Examples of AI/ML applications in education include automatic essay grading systems, systems that predict student graduation outcomes, enrollment systems, and student learning support systems (Chen et al., 2020a; Zhang and Aslan, 2021; Mahmud and Gagnon, 2023; Zhai et al., 2021). While these systems have helped streamline decision-making processes, it is important to consider fairness when using them, as important human decisions are involved (Baker and Hawn, 2021; Kizilcec and Lee, 2021). It has been observed that issues such as grading essays and speech can be influenced by the student's native language (Mayfield et al., 2019), and predicting enrollment and academic results can be affected by factors such as gender and race (Baker and Hawn, 2021; Hutchinson and Mitchell, 2019). Improving fairness in AI/ML systems in education and other fields is, therefore, crucial if these systems are to be trusted and fully replace human decision-making (Corbett-Davies et al., 2017; Smith, 2020).

In the domain of educational research, a significant increase in scientific publications related to AI in education (AIEd) has been observed (Chen et al., 2020). Xieling Chen et al. conducted a comprehensive analysis, identifying over 140 articles from six SSCI-recognized journals in educational technology, covering the period 1999–2019 (Chen et al., 2020). Similarly, Kai Siang Chan and colleagues explored AI applications in medical education, discussing the advantages and challenges in implementing intelligent teaching programs, automated grading, and accuracy assessment in virtual reality environments (Chan and Zary, 2019). Linja Chen et al. assessed the impact of AI across various educational domains, using over 40 selected articles, including journal publications and professional, governmental, and organizational reports. This study focused on AI's influence on teaching, learning, and educational management and administration (Chen et al., 2020a). Fati Tahiru et al. presented an overview of the opportunities, benefits, and challenges of applying Artificial Intelligence (AI)

Table 2
Different types of bias.

Type of bias	Definition	Ref.
Historical Bias	Historical bias is the already existing biases and socio-technical issues in the world and can seep into the data generation process even given a perfect sampling and feature selection	Baker and Hawn (2021), Suresh and Guttag (2021, 2019)
Representation Bias	Representation bias happens from the way we define and sample from a population	Mehrabi et al. (2019), Baker and Hawn (2021)
Measurement Bias	Measurement bias happens from the way we choose, utilize, and measure a particular feature	Mehrabi et al. (2019), Baker and Hawn (2021)
Evaluation Bias	Evaluation bias happens during model evaluation	Mehrabi et al. (2019), Baker and Hawn (2021)
Aggregation Bias	Aggregation bias happens when false conclusions are drawn for a subgroup based on observing other different subgroups or generally when false assumptions about a population affect the model's outcome and definition	Mehrabi et al. (2019), Baker and Hawn (2021)
Population Bias	Population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population represented in the dataset or platform from the original target population	Mehrabi et al. (2019), Olteanu et al. (2019)
Simpson's Paradox	The bias in the analysis of heterogeneous data. What can be observed in underlying subgroups may be quite different from what can be when these subgroups are aggregated	Mehrabi et al. (2019), Blyth (1972)
Longitudinal Data Fallacy	Observational studies often treat cross-sectional data as if it were longitudinal, which may create biases due to Simpson's paradox	Mehrabi et al. (2019)
Sampling Bias	Sampling bias arises due to non-random sampling of subgroups	Mehrabi et al. (2019)
Behavioral Bias	Behavioral bias arises from different user behavior across platforms, contexts, or different datasets	Mehrabi et al. (2019), Olteanu et al. (2019)
Content Production Bias	Content Production bias arises from structural, lexical, semantic, and syntactic differences in the contents generated by users	Mehrabi et al. (2019), Olteanu et al. (2019)
Linking Bias	Linking bias arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users	Mehrabi et al. (2019), Olteanu et al. (2019)
Temporal Bias	Temporal bias arises from differences in populations and behaviors over time	Mehrabi et al. (2019), Olteanu et al. (2019)
Popularity Bias	Items that are more popular tend to be exposed more. However, popularity metrics are subject to manipulation—for example, by fake reviews or social bots	Mehrabi et al. (2019), Nematzadeh et al. (2018)
Algorithmic Bias	Algorithmic bias is when the bias is not present in the input data and is added purely by the algorithm	Mehrabi et al. (2019), Baeza-Yates (2018)
User Interaction Bias	User Interaction bias is a type of bias that can not only be observed on the Web but also get triggered from two sources—the user interface and through the user itself by imposing his/her self-selected biased behavior and interaction	Mehrabi et al. (2019), Baeza-Yates (2018)
Presentation Bias	Presentation bias is a result of how information is presented	Mehrabi et al. (2019), Baeza-Yates (2018)
Ranking Bias	The idea that top-ranked results are the most relevant and important will result in the attraction of more clicks than others	Mehrabi et al. (2019)
Social Bias	Social bias happens when other people's actions or content coming from them affect our judgment	Mehrabi et al. (2019), Baeza-Yates (2018)
Emergent Bias	Emergent bias happens as a result of use and interaction with real users. This bias arises as a result of a change in population, cultural values, or societal knowledge usually sometime after the completion of design	Mehrabi et al. (2019), Friedman and Nissenbaum (1996)
Self-Selection Bias	Self-selection bias is a subtype of the selection or sampling bias in which subjects of the research select themselves	Mehrabi et al. (2019), Mester (2023)
Omitted Variable Bias	Omitted variable bias occurs when one or more important variables are left out of the model	Mehrabi et al. (2019), Mester (2023)
Cause-Effect Bias	Cause-effect bias can happen as a result of the fallacy that correlation implies causation	Mehrabi et al. (2019), Mester (2023)
Observer Bias	Observer bias happens when researchers subconsciously project their expectations onto the research	Mehrabi et al. (2019), Mester (2023)
Funding Bias	Funding bias arises when biased results are reported to support or satisfy the funding agency or financial supporter of the research study	Mehrabi et al. (2019), Mester (2023)

to the education sector and discussed the challenges in applying AI in Education through technological, organizational, and environmental aspects (Mahmud and Gagnon, 2023). Xuesong Zhai et al. presented a review of research on AI in education, focusing on articles published between 2010 and 2020 from Web of Science and examining three essential aspects of AI in knowledge processing outlined in the studies, including (a) knowledge representation (b) knowledge acquisition, and (c) knowledge extraction. This review will focus on AI techniques and tools integrated into education recently following the popularization of AI (Zhai et al., 2021). Ke Zhang et al. presented a research overview to help the AIEd community, including educators, education researchers, AI technology creators, and stakeholders (Zhang and Aslan, 2021), understand its current status, potential, challenges, and future directions.

2.3. Existing reviews on AI fairness in education

The exploration of fairness in AI systems, particularly within the educational context, has been a significant theme in recent scholarly literature. We identified some secondary work that is relevant to our study. In a 2021 study titled 'Algorithmic bias in education', Ryan S. Baker reviewed biased algorithms used in education (Baker and Hawn, 2021). The authors focused on discussing theoretical work regarding the root causes of algorithmic bias and reviewing the existing empirical literature on the specific ways that algorithmic bias has been observed in education. Their review aimed to understand who is affected and how the context surrounding the algorithms plays a role. They particularly emphasized biases that arise from how variables are operationalized and which datasets are utilized (Baker and Hawn, 2021). René F. Kizilcec, in a paper titled 'Algorithmic Fairness in Education', presented perspectives on fairness, prejudice, and discrimination while identifying the origins of bias and discrimination in education. From this, recommendations were made regarding the promotion of fairness in algorithms (Kizilcec and Lee, 2021). Most recently, in March 2023, Lin Li et al. published a paper with the title 'Moral Machines or Tyranny of the Majority? A Systematic Review on Predictive Bias in Education'. This systematic review aimed to analyze and summarize existing relevant studies from three different perspectives, namely, (i) protective attributes used to operationalize predictive bias; (ii) fairness measures used in various predictive tasks in education, and (iii) existing strategies used to mitigate predictive bias and enhance predictive fairness (Li et al., 2023).

While the above reviews provide us with a fairly comprehensive overview of research primary studies on fairness and bias in algorithms applied in education, our survey distinguishes itself by not only thoroughly cataloging the existing research in this field but also systematically organizing it. In addition to cataloging the AI/ML algorithms that have been utilized and the definitions of fairness and bias, we also compile the datasets commonly used by researchers when investigating fairness and bias in AI/ML systems in the field of education. Furthermore, our survey includes a comprehensive and clear presentation of methods, fairness metrics, and performance evaluations of these techniques. This systematic approach is the valuable distinction of our survey.

3. Research methodology

The recent reviews have shown that AI/ML applications in education represent a broad and diverse field of research. Our interest is to classify and categorize characteristics of primary studies into fairness-relevant theoretical and empirical dimensions and, from that, identify research gaps that can contribute to developing practical fairness-aware AI/ML systems for educational sectors. As we do not focus on in-depth analysis of results from each primary study or forming theoretical frameworks, we select a systematic mapping study instead of a systematic literature review as our research approach (Berg et al., 2018). Petersen et al. suggest that, by categorizing the primary studies, a systematic mapping study provides a structure for the type of research reports and results that have been published (Petersen et al., 2015). The first step of the process involves posing RQs, which then help to generate a visual summary of the research results. Subsequent phases include screening studies based on their title, abstract, and keywords. The outcomes of this process aim to address the RQs, with the principal objective being to uncover research gaps in the examined area. Our systematic search comprises the following steps:

- Step 1: Development of the search protocol via a pilot search
- Step 2: Determination of inclusion and exclusion criteria for relevant publications.
- Step 3: Systematic search for relevant studies, followed by critical evaluation
- Step 4: Additional manual search. A manual search was performed to find more relevant papers.
- Step 5: Quality assessment. To identify the rigor of the remaining papers, a quality assessment was performed on the papers that provided empirical evidence. The complete assessment can be found in Appendix A
- Step 6: Data extraction and synthesis. From the primary papers, relevant data and information were extracted into a classification schema. A multi-step synthesis was performed to answer the research questions

Step 1 and Step 2 are initiated by the third author and conducted by the first and the third author. Step 3 to Step 5 was mainly done by the first author. The second and the third authors regularly follow the process and participate in resolving indecisive situations during the process. Step 6 was conducted separately by all authors, and then the results were discussed and merged. In the following sections, we describe the mentioned steps in detail.

3.1. Step 1: Development of search protocol

A pilot search was started with five seed papers. The search protocol, data extraction instruments, and analysis forms are reused from our previous secondary studies (Cico et al., 2021; Berg et al., 2018). Deriving from the main RQs, our search string has three parts: FAIRNESS (C1), ML/ AI (C2), and EDUCATION (C3). The synonyms of these terms were identified in the context of either Computer Science, Software Engineering, or Information Systems by interviewing field experts. Several trial searches were conducted to adjust the scope of the search string so that we do not include many irrelevant studies from different research fields. The most important information is the main focus of the articles on fairness in the educational context. Besides, we want to include as many studies as possible.

We also developed the inclusion/exclusion criteria on a small set of papers and adjusted the criteria based on this pilot before applying them to the full set of data.

After several trials, we ended up with the list of search words as shown in Table 3. The search string is formed as the formula: C1 AND C2 AND C3.

Our linguistic capability resulted in a limitation of the scope of publication to studies published in the English language. Several electronic databases were suggested by the second author, who has conducted several systematic literature reviews before. The list includes Google Scholar, Scopus, ISI Web of Science, IEEE Xplore, Current Contents, Kluwer Online, Computer Database, Science Direct, Springer Link, Inspec, ACM Digital Library, and ConnectedPaper.[1] Considering the availability of the databases for both the first and the last authors, the previous experiences of reviewers, flexible formulation of search strings with unlimited clauses, and easy exporting of paper lists in various formats, we decided to select the four databases: Google Scholar, IEEE Xplore, ACM Digital Library, and ConnectedPaper. Google Scholar is known for its vast, interdisciplinary repository of scholarly articles, providing broad coverage and availability in both authors' locations. IEEE Xplore and ACM Digital Library are specialized databases focusing on computer science, software engineering, and information systems, offering in-depth coverage of these areas. ConnectedPaper offers a unique feature of visualizing the connections between primary studies, which can be particularly helpful in understanding the landscape of

[1] https://www.connectedpapers.com

Table 3
Synonyms to key search words.

Search part	Main term	Synonyms
C1	Fairness	“fair” or “responsible” or “transparent” or “ethic” or “bias” or “discrimination”
C2	ML/AI	“Machine learning” or “Artificial Intelligence”
C3	Education	“school” or “university” or “college” or “education” or “teaching” or “pedagogy”

research and identifying key primary studies and trends. The search ranges from 1970 to 2023. We screened the sources based on the title, abstract, and keyword metadata to help us select studies relevant to our RQs.

3.2. Inclusion and exclusion criteria

To prepare for the next stage in the process, we developed the following inclusion criteria:

- IC1: the paper should investigate AI/ML fairness in Education as the main research topic
- IC2: the paper performs a type of empirical or theoretical evaluation
- IC3: it is possible to understand the details of AI/ML investigated in the paper
- IC4: it is possible to extract the details of evaluation datasets
- IC5: it is possible to extract the result of the evaluation

Both studies conducted in industry and in academic environments were considered. We also included both conceptual and empirical studies. Due to the nature of the field, we also consider papers published in open-access portals, including pre-print archived primary studies. We conducted a quality assessment round to ensure the quality of the primary studies that were included. The limitation was that we only searched for primary studies written in English.

The exclusion criteria (EX) are as follows:

- EX1: the paper does not investigate fairness or bias in the development and operations of AI/ML systems as the primary research topic.
- EX2: the paper investigates fairness or bias of AI/ML systems but not in the education context.
- EX3: the paper without full-text access.
- EX4: the paper with low quality (lower than the average score of 2 shown in Table 5).
- EX5: the paper is not presented in English.

3.3. Systematic search for relevant studies

Systematic search for primary studies: The search process took place from the beginning of January 2019 until the end of December 2023. By applying the search strings in the digital databases, we were able to acquire three sets of primary studies. From the ACM Digital Library, we retrieved 1869 primary studies. From IEEE Xplore, we retrieved 534 primary studies. From the first 20 pages of Google Scholar, we retrieved 1492 primary studies. Google Scholar’s algorithm prioritizes the relevance and impact of its search results; hence, the most pertinent and influential studies are likely to appear on the initial pages. The first 20 pages are likely to include the majority of significant and highly cited studies relevant to your topic. Beyond this point, the additional value gained from each new page of results is marginal and does not substantially contribute to the depth or quality of the review. We also searched from three portals independently to increase the confidence of the search results.

After searching the above three sources, based on the final selected articles, we performed another search through the Connected Papers website to expand the search for articles related to articles selected and published in 2023. Many related articles were found, but in this step, we only selected articles published in 2023 with titles and abstracts related to the content we are researching. That way, we retrieved 16 more articles in this step.

Select by title, keyword, and abstract: We removed articles that clearly have no connection to the topic of fairness in ML/ AI in the educational context. It is important that all these topics should be clearly displayed in the articles’ title or abstract. This process allowed us to eliminate a large number of irrelevant studies. The remaining numbers from each database are 52 (from ACM Digital), 35 (IEEE Xplore), and 424 (Google Scholar).

Select by full-text : The articles that cannot be determined with an abstract and titles will be further evaluated by reading full texts. The first author assessed the primary studies and sought to understand the empirical evaluation of ML/ AI fairness in an educational context. After this stage, the remaining numbers from each database are 41 (from ACM Digital), 15 (IEEE Xplore), and 99 (Google Scholar).

To manage references for removing duplicates and storing a large number of findings, we used Zotero[2] as a reference manager software. We also retrieve papers that are published in different venues from the same group of authors, examining the same research questions with a minor update (adding some new datasets or testing some new algorithms). In these cases, we decided to select only the most comprehensive version among these papers (see Fig. 1).

3.4. Additional manual search

A manual search was conducted with the participation of the first and the third authors using the forward snowballing technique [42], to identify additional papers not discovered by the search string. Google Scholar was used to examine the citations to the paper being examined. The publication lists of frequently appearing authors were also searched. We also used ConnectedPaper platform to find better links between the papers. This resulted in several papers as candidates for inclusion. After assessing the title, abstract, and finally, the full text, 16 more papers were found. Then, we removed four duplicate primary studies, and by reading the details of those primary studies, we retained eight primary studies for our research, which were published in 2023. We adopted a practice introduced by Hort et al. (2022) to solicit feedback on this mapping study by asking the authors of included papers to check for accurate reporting of their work as well as encourage them to point out other relevant work in the field. As a result, we received nine feedback emails from the authors of primary studies. While most of them confirm the comprehensiveness of our set of primary studies, they also suggested some more papers. These papers are not found in our systematic search due to either (1) uncommon self-archived portals that publish the papers, (2) papers published in 2024, (3) papers that use special terms that are not related to fairness and education. After applying our search protocol, inclusion criteria, and quality assessment, we were able to add four more studies to our paper set.

3.5. Demographics of primary studies

The final set of primary studies includes 63 articles. Fig. 2 presents the distribution of the selected primary studies across publication years with the growing interest in the topic over the years. The first paper that empirically evaluates fairness in ML/ AI systems is from 2002 (The

[2] https://www.zotero.org/

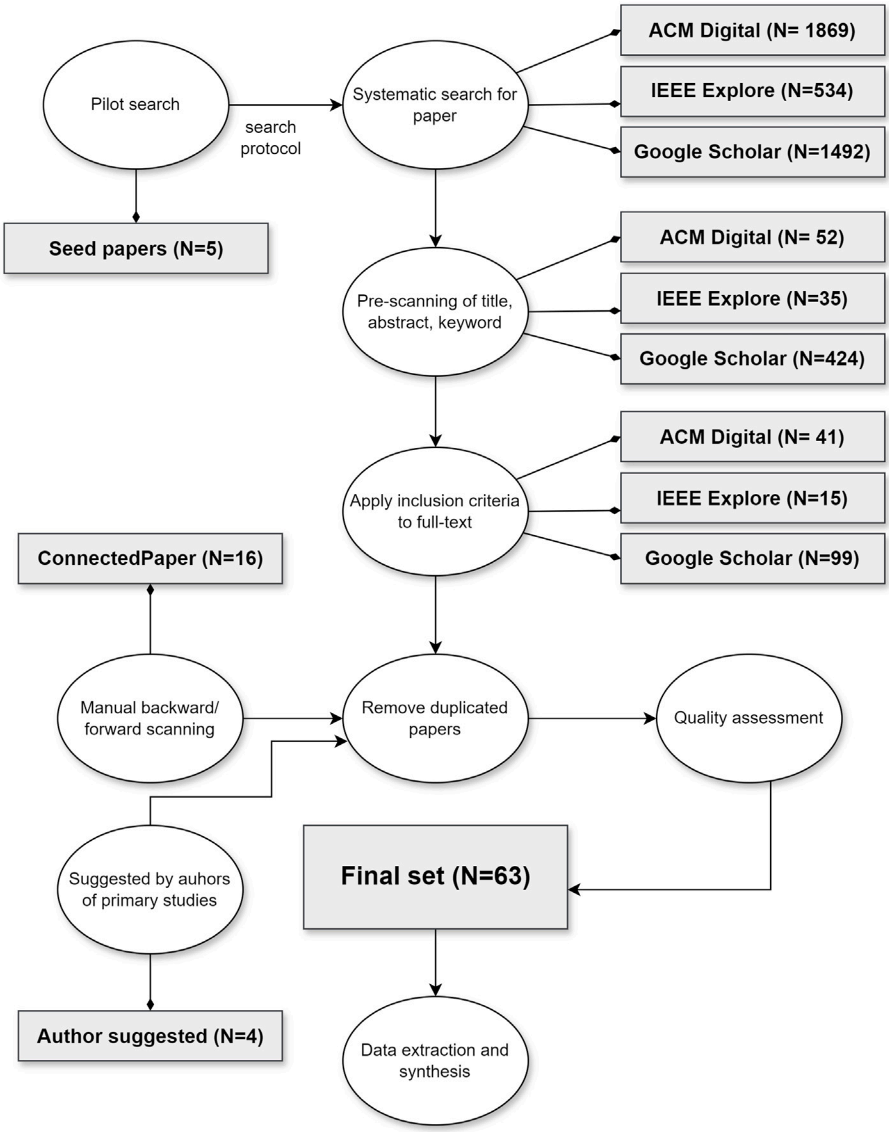

Fig. 1. The study selection process.

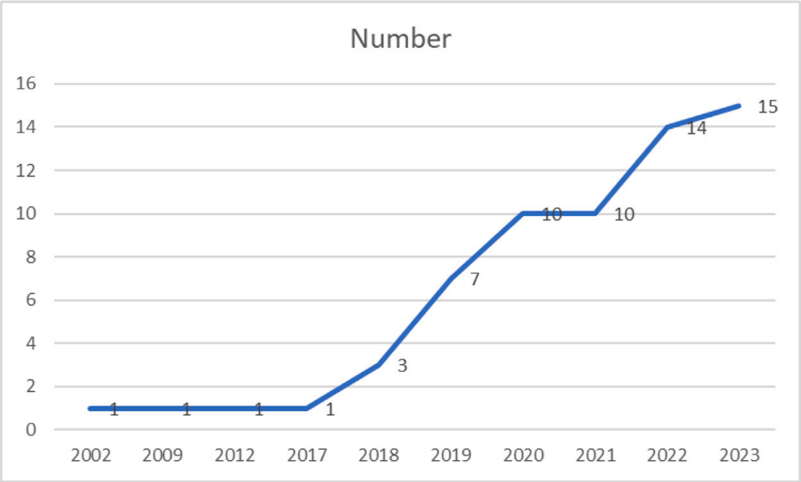

Fig. 2. Distribution of primaries across publication years.

Table 4
The distribution of primary studies across publication venue.

Venue	Ref.	No. of primary studies
arXiv.org	Saleiro et al. (2019), Kizilcec and Lee (2021), Lee and Kizilcec (2020), Manisha and Gujar (2018), Holstein and Doroudi (2021a), Di Carlo et al. (2023), Madaio et al. (2021), Han et al. (2024), Verger et al. (2023), Gándara et al. (2023)	10
International Conference on Educational Data Mining	Hutt et al. (2019), Yu et al. (2020), Paquette et al. (2020), Hu and Rangwala (2020), Anderson et al. (2019), Tschiatschek et al. (2022b), Arthurs and Alvero (2020),	7
Artificial Intelligence in Education	Baker and Hawn (2021), Bogina et al. (2022), Sha et al. (2021), Fenu et al. (2022)	4
Conference on Fairness, Accountability, and Transparency	Hutchinson and Mitchell (2019), Marcinkowski et al. (2020), Gardner et al. (2023), Friedler et al. (2019)	4
International Conference on Learning Analytics and Knowledge	Li et al. (2023), Gardner et al. (2019a), Verdugo et al. (2022), Li et al. (2021b)	4
Conference on Learning @ Scale	Kung and Yu (2020), Yu et al. (2021), Deho et al. (2023b)	3
Applied Measurement in Education	Clauser et al. (2002a), Bridgeman et al. (2012)	2
Conference on Human Factors in Computing	Holstein et al. (2019), Mashhadi et al. (2022)	2
IEEE Transactions on Learning Technologies	Sha et al. (2022), Deho et al. (2023a)	2
Neural Information Processing Systems	Alghamdi et al. (2022),Jeong et al. (2021)	2
International Conference on Computer Supported Education	Riazy et al. (2020), Rzepka et al. (2022)	2
AAAI/ACM Conference on AI, Ethics, and Society	Jiang and Pardos (2021)	1
IEEE Intelligent Systems	Du et al. (2021)	1
Data Science and Engineering	Grari et al. (2020)	1
ACM SIGCSE Bulletin	Elglaly and Liu (2023)	1
ACM Technical Symposium on Computer Science Education	Dobesh et al. (2023)	1
AI and Ethics	Akgun and Greenhow (2022)	1
Annual Meeting of the National Council on Measurement in Education (NCME), San Diego, CA, United States	Bridgeman et al. (2009)	1
Augmented Intelligence and Intelligent Tutoring Systems	Matias and Zipitria (2023)	1
Conference on Research in Equitable and Sustained Participation in Engineering, Computing, and Technology (RESPECT)	Karumbaiah and Brooks (2021)	1
Workshop on Ethics in NLP	Larson (2017)	1
International Conference on Multimedia Big Data (BigMM)	Kulkarni et al. (2021)	1
International Conference on Teaching, Assessment and Learning for Engineering (TALE)	Xiang et al. (2022)	1
International Learning Analytics & Knowledge Conference	Doroudi and Brunskill (2019)	1
Journal of Educational Measurement	Huggins-Manley et al. (2022)	1
Journal of Machine Learning Research	Wei et al. (2021)	1
Learning, Media and Technology	Sahlgren (2023)	1
Workshop on Innovative Use of NLP for Building Educational Applications	Loukina et al. (2019)	1
Expert Systems with Applications	Sha et al. (2023)	1
Frontiers in Education	Yee et al. (2023)	1
Education Sciences	Nezami et al. (2024)	1

article by Brian E. Clauser, Michael T. Kane, and David B. Swanson, “Validity Issues for Performance-Based Tests Scored With Computer-Automated Scoring Systems”,). Before that, studies investigated fairness in a theoretical way or in laboratory-type systems. Since 2019, there has been a significant growth in empirical studies of fairness in AI systems.

Table 4 revealed the top five most common sources of primary studies, including (1) Arxiv, (2) International Conference on Educational Data Mining, (3) Journal of Artificial Intelligence in Education, (4) Conference on Fairness, Accountability, and Transparency, and (5) Conference on Learning Analytics and Knowledge.

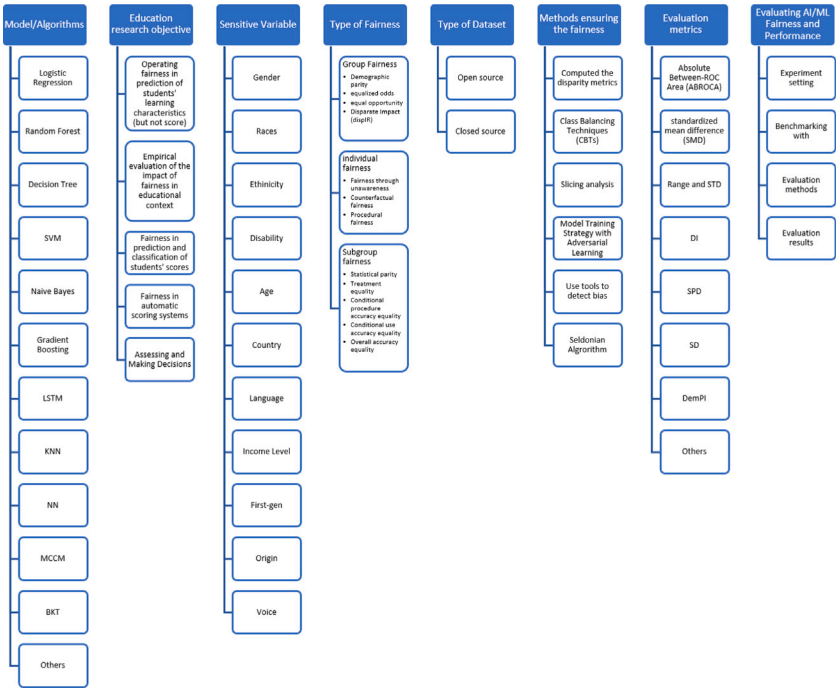
Model/Algorithms
Logistic Regression
Random Forest
Decision Tree
SVM
Naive Bayes
Gradient Boosting
LSTM
KNN
NN
MCCM
BKT
Others
Education research objective
Operating fairness in prediction of students' learning characteristics (but not score)
Empirical evaluation of the impact of fairness in educational context
Fairness in prediction and classification of students' scores
Fairness in automatic scoring systems
Assessing and Making Decisions
Sensitive Variable
Gender
Races
Ethinicity
Disability
Age
Country
Language
Income Level
First-gen
Origin
Voice
Type of Fairness
Group Fairness
• Demographic parity
• equalized odds
• equal opportunity
• Disparate Impact (dispIR)
individual fairness
• Fairness through unawareness
• Counterfactual fairness
• Procedural fairness
Subgroup fairness
• Statistical parity
• Treatment equality
• Conditional procedure accuracy equality
• Conditional use accuracy equality
• Overall accuracy equality
Type of Dataset
Open source
Closed source
Methods ensuring the fairness
Computed the disparity metrics
Class Balancing Techniques (CBTs)
Slicing analysis
Model Training Strategy with Adversarial Learning
Use tools to detect bias
Seldonian Algrorithm
Evaluation metrics
Absolute Between-ROC Area (ABROCA)
standardized mean difference (SMD)
Range and STD
DI
SPD
SD
DemPI
Others
Evaluating AI/ML Fairness and Performance
Experiment setting
Benchmarking with
Evaluation methods
Evaluation results

Fig. 3. Classification scheme derived from seed primary studies.

Table 5
Quality assessment checklist.

Category	Question
Problem Statement	Q1. Is the research objective sufficiently explained and well-motivated?
Research Design	Q2. Is the context of the study clearly stated? Q3. Is the research design prepared sufficiently?
Data Collection	Q4. Are the data collection & measures adequately described? Q5. Are the measures and constructs used in the study the most relevant for answering the research question?
Data Analysis	Q6. Is the data analysis used in the study adequately described? Q7a. Qualitative study: Are the interpretation of evidence clearly described? Q7b. Quantitative study: Are the effect sizes reported with assessed statistical significance? Q8. Are potential alternative explanations considered and discussed in the analysis?
Conclusion	Q9. Are the findings of the study clearly stated and supported by the results? Q10. Does the paper discuss limitations or validity?

3.6. Step 5: Quality assessment

A quality assessment of the paper was conducted to evaluate the relative strength of the reported evidence. We employed the quality assessment checklist from our previous work (Nguyen-Duc et al., 2015), with questions derived from quality criteria including rigorousness, credibility, and relevance.

In question Q1, we evaluated whether the research objective was clearly stated and supported by industry observation or existing theory. Questions Q2 and Q3 aimed to assess the study's preparation and the adequacy of context information provided. Questions Q4 and Q5 focused on evaluating the sufficiency of data collection methods, instruments, and measures, as well as examining if the constructs and measures aligned with the stated objective.

For questions Q6, Q7, and Q8, we examined whether the data analysis was adequately reported. In quantitative studies, we looked for the reporting of effect sizes and statistical significance, while in qualitative studies, we checked for the presentation of qualitative evidence interpretation, such as interview quotes or observation field notes. Additionally, we verified whether alternative explanations for the results were considered.

Questions Q9 and Q10 were designed to assess if the outcomes were clearly documented and if the study's threat to validity was adequately addressed. A four-point Likert scale was used to collect answers for quality assessment questions (Jamieson, 2004), where each question had four possible options: not mentioned at all (score 0), little mentioned (score 1), adequately addressed (score 2), and completely addressed (score 3). To ensure the reliability of selected primary studies, we considered only those with an average quality score equal to or greater than 2.

The quality assessment of the primary studies was independently conducted by all researchers. The results were compared and merged. Discussions will be conducted by the authors to resolve conflicts and agree on scoring the primary studies. The details of the final quality assessment are provided in Appendix A.

3.7. Step 6: Data extraction and synthesis

From the collected articles we extracted data from the articles to address the research questions. We employed a process called keywording (Cico et al., 2021; Berg et al., 2018), adopting a classification schema for extracting relevant information. To define the schema (Petersen et al., 2015), we conducted an analysis of existing systematic review primary studies on AI fairness to extract relevant terms. This process mirrors ground theory with open coding of keywords from papers extracted from the pilot search. We labeled categories in the scheme using the extracted terms and established clear boundaries among them to facilitate the classification of primary studies into one or multiple categories.

The extracted scheme, presented in Fig. 3, comprises eight categories: (1) ML/AI models or algorithms, (2) education mission or object, (3) sensitive variables, (4) type of fairness, (5) type of dataset, (6) method ensuring fairness, (7) evaluation metric, and (8) evaluation methodology. The scheme firstly lists the most popular AI/ML

Table 6
The list of AI algorithms.

Model /Algorithms	Definition	Paper	No. of primary studies
Logistic Regression	Statistical model analyzing binary outcomes by estimating probabilities using a logistic function	Di Carlo et al. (2023), Gándara et al. (2023), Yu et al. (2020), Hu and Rangwala (2020), Anderson et al. (2019), Sha et al. (2021), Gardner et al. (2023, 2019a), Verdugo et al. (2022), Kung and Yu (2020), Deho et al. (2023b), Clauser et al. (2002a), Bridgeman et al. (2012, 2009), Wei et al. (2021), Friedler et al. (2018), Belitz et al. (2021), Nezami et al. (2024), Alghamdi et al. (2022), Rzepka et al. (2022)	20
Random Forest (RF)	A supervised algorithm that uses an ensemble learning method consisting of a multitude of decision trees	Lee and Kizilcec (2020), Di Carlo et al. (2023), Gándara et al. (2023), Hutt et al. (2019), Yu et al. (2020), Anderson et al. (2019), Sha et al. (2021), Verdugo et al. (2022), Kung and Yu (2020), Deho et al. (2023b,a), Loukina et al. (2019), Nezami et al. (2024), Li et al. (2021a), Alghamdi et al. (2022)	14
Decision Tree (DT)	A model that makes decisions based on hierarchical, tree-like structures of rules and outcomes	Di Carlo et al. (2023), Anderson et al. (2019), Gardner et al. (2019a), Verdugo et al. (2022), Kung and Yu (2020), Grari et al. (2020), Riazy et al. (2020), Wei et al. (2021), Friedler et al. (2018), Belitz et al. (2021), Nezami et al. (2024), Rzepka et al. (2023, 2022)	14
Support Vector Machines (SVM)	Supervised learning models that classify data by finding the optimal separating hyperplane between classes	Di Carlo et al. (2023), Yu et al. (2020), Anderson et al. (2019), Gardner et al. (2019a), Verdugo et al. (2022), Kung and Yu (2020), Deho et al. (2023b), Friedler et al. (2018), Nezami et al. (2024)	11
Naive Bayes (NB)	A probabilistic classifier based on applying Bayes' theorem with strong, naive independence assumptions between features	Sha et al. (2021), Gardner et al. (2019a), Kung and Yu (2020), Riazy et al. (2020), Friedler et al. (2018), Belitz et al. (2021)	6
Gradient Boosting Machine Learning	An ensemble technique that optimizes predictive models by sequentially adding weak learners to minimize errors	Gardner et al. (2023), Deho et al. (2023b,a), Bayer et al. (2021), Alghamdi et al. (2022)	5
K nearest neighbor (KNN)	A non-parametric method used for classification and regression by analyzing the nearest data points	Verdugo et al. (2022), Nezami et al. (2024), Rzepka et al. (2023, 2022)	4
Long Short-Term Memory (LSTM)	Recurrent neural network capable of learning long-term dependencies using memory cells	Sha et al. (2021), Gardner et al. (2019a), Jiang and Pardos (2021)	3
Neural Network (NN)	a type of machine learning model that is designed to simulate the workings of the human brain	Manisha and Gujar (2018), Gardner et al. (2023)	2
Multiple cooperative classifier model (MCCM)	integrating multiple classifiers working together to improve accuracy and reliability in predictions	Kizilcec and Lee (2021), Hu and Rangwala (2020)	2
Bayesian Knowledge Tracing (BKT)	A probabilistic model tracking learners' knowledge over time using Bayesian inference and learner performance data	Doroudi and Brunskill (2019), Tschiatschek et al. (2022a)	2
Multi-Layer Perceptron (MLP)	Feedforward artificial neural network with multiple layers of nodes for complex learning	Riazy et al. (2020), Rzepka et al. (2023)	2
RANDOM, PERFECT, META	RANDOM model using a random sample drawn from a normal distribution. PERFECT model also contained a single feature. META model only relied on demographic information	Loukina et al. (2019)	1
Ordinary Least Squares (OLS) regression models	A statistical method estimating unknown parameters by minimizing the sum of squared differences from data.	Marcinkowski et al. (2020)	1
Seldonian Algorithm	An algorithmic framework ensuring predefined fairness and safety constraints in decision-making processes are met	Li et al. (2021b)	1
Five Collaborative Filtering algorithms (FCF)	UserKNN, ItemKNN, BPR, BiasedMF, SVD++	Gómez et al. (2021)	1
Additive Factor Model (AFM)	A statistical model used to understand the relationships between variables and outcomes...	Doroudi and Brunskill (2019)	1
Deep Knowledge Tracing (DKT)	Deep Knowledge Tracing (DKT) is a machine learning technique used for educational data analysis and modeling of students' knowledge	Tschiatschek et al. (2022a)	1
Bayesian–Bayesian Knowledge Tracing (B2KT)	An extension of the basic BKT model that uses a hierarchical Bayesian framework	Tschiatschek et al. (2022a)	1
Word2vec Skip-Gram (word embeddings)	The word2vec Skip-Gram model is a machine learning model used to create vector representations of words (word embeddings)	Arthurs and Alvero (2020)	1
Deep Learning (DL)	A neural network with multiple hidden layers that automatically learns features from complex data. It uses non-linear activation functions and backpropagation	Rzepka et al. (2022)	1

Table 7
The education research objective in primary studies.

The education misson	Description	Affected object	Reference	No. of papers
Operating fairness in prediction of students' learning characteristics (but not score)	A type of educational data mining that can be formulated in the form of a problem such as detecting factors that affect student learning outcomes; predicting students likely to drop out of school; predicting students' graduation results; detecting risks; etc...	College Students	Du et al. (2021), Kizilcec and Lee (2021), Lee and Kizilcec (2020), Madaio et al. (2021), Verger et al. (2023), Gándara et al. (2023), Hutt et al. (2019), Yu et al. (2020), Marcinkowski et al. (2020), Gardner et al. (2023), Verdugo et al. (2022), Li et al. (2021b), Kung and Yu (2020), Yu et al. (2021), Deho et al. (2023b,a), Riazy et al. (2020), Xiang et al. (2022), Huggins-Manley et al. (2022), Jeong et al. (2021), Sha et al. (2023), Gardner et al. (2019b), Rzepka et al. (2023, 2022)	23
Empirical evaluation of the impact of fairness in educational context	An issue mentioned by many studies when studying fairness in ML systems applied in education is promoting ML fairness in education. These studies provide recommendations to improve fairness, reduce bias for machine learning systems, or highlight the importance of studying fairness when designing applied machine learning products in education.	College student/ K-12	Arthurs and Alvero (2020), Paquette et al. (2020), Tschiatschek et al. (2022b), Bogina et al. (2022), Fenu et al. (2022), Verdugo et al. (2022), Holstein et al. (2019), Mashhadi et al. (2022), Sha et al. (2022), Elglaly and Liu (2023), Dobesh et al. (2023), Akgun and Greenhow (2022), Matias and Zipitria (2023), Karumbaiah and Brooks (2021), Doroudi and Brunskill (2019), Huggins-Manley et al. (2022), Wei et al. (2021), Sahlgren (2023), Holstein and Doroudi (2021b), Clauser et al. (2002b)	20
Fairness in prediction and classification of students' scores	Prediction/classification is a popular empirical research in educational data mining (EDM). Just like the student detection problem, the prediction problem focuses on tasks such as predicting/classifying graduation scores, predicting/classifying dropout status, predicting/classifying the possibility of late graduation, etc.	College Students/ High-school	Kizilcec and Lee (2021), Di Carlo et al. (2023), Anderson et al. (2019), Jiang and Pardos (2021), Sha et al. (2023), Nezami et al. (2024), Gardner et al. (2019b), Alghamdi et al. (2022)	9
Fairness in automatic scoring systems	Automatic scoring is one of the most mentioned applications in EDM. The system will rely on previously collected data used in the training process and thereby predict the results. A typical system can be mentioned as an automatic essay grading system.	College Student/ K-12 (kindergarten through grade 12)	Kizilcec and Lee (2021), Bridgeman et al. (2012, 2009), Loukina et al. (2019), Clauser et al. (2002b)	5
Assessing and Making Decisions	Evaluation and decision making is also an important task in EDM, which is based on the prediction/classification results, from which evaluation and decisions are made. These tasks may include assessing academic ability and recommending further courses; Evaluating the likelihood of success and making selection decisions; etc.	College Student	Baker and Hawn (2021), Kizilcec and Lee (2021), Sha et al. (2021), Yee et al. (2023)	4

algorithms extracted from the primary studies, including Naive Bayes, Random Forest, SVM, Decision Tree, and Logistic Regression. The education mission or object describes the research objective of the primary study that reflects its educational nature. Sensitive variables indicate variables considered sensitive in the context of fairness, such as gender, ethnicity, and country. The type of datasets is classified into open source or closed source. Type of fairness gives three main categories of fairness — group fairness, individual fairness, and subgroup fairness for common classification. Methods ensuring fairness classify techniques and strategies used to ensure fairness in models, such as class balancing techniques, slicing analysis, and model training strategies. Evaluation metrics count studies that use metrics like Absolute Between-ROC Area (ABROCA), standardized mean difference (SMD), range and standard deviation (STD), disparity index (DI), and statistical parity difference (SPD). Evaluating AI/ML Fairness and Performance presents data extracted about experiment setting, benchmarking, evaluation methods, and evaluation results.

After defining the classification schema resulting from the keywording process, we systematically extract data from the primary studies. We analyze the selected studies in the pool and identify a list of attributes connected to the previously constructed categories. Subsequently, we store the extracted studies in a systematic map, which we utilize to answer each of the RQs. We take great care in checking the following attributes from each paper source: Title, First author, Year of publication, Abstract, Keywords, Full text, and Publication source.

4. Result

This section presents our findings for the research questions: (1) AI algorithms and approaches reported in the primary studies in a higher education context (Section 4.1), (2) Investigated problems in the primary studies (Section 4.2), (3) Fairness and bias definitions used in the primary studies (Section 4.3), (4) Characteristics of the dataset used in the primary studies (Section 4.4), (5) Evaluation methods for AI fairness in the primary studies (Section 4.5), (6) Evaluation metric(s) are used to assess the AI fairness in the primary studies (Section 4.6) and (7) Techiques to asess the AI fairness and performance (Section 4.7).

4.1. Rq1. Which AI algorithms/approaches are reported from the primary studies in the higher education context?

From Table 6 it is evident that the most frequently utilized algorithm is logistic regression, with a total of 15 primary studies. Following this, other algorithms such as Decision Tree, Random Forest,

Table 8
Sensitive variables in studies.

Sensitive variable	Description	Reference	No. of primary studies
Gender	A person's biological sex, which can be male, female, or non-binary.	Baker and Hawn (2021), Hutchinson and Mitchell (2019), Kizilcec and Lee (2021), Lee and Kizilcec (2020), Manisha and Gujar (2018), Han et al. (2024), Verger et al. (2023), Yu et al. (2020), Paquette et al. (2020), Hu and Rangwala (2020), Sha et al. (2021), Gardner et al. (2019a), Verdugo et al. (2022), Li et al. (2021b), Kung and Yu (2020), Yu et al. (2021), Deho et al. (2023b), Bridgeman et al. (2012), Deho et al. (2023a), Grari et al. (2020), Dobesh et al. (2023), Akgun and Greenhow (2022), Bridgeman et al. (2009), Larson (2017), Riazy et al. (2020), Kulkarni et al. (2021), Wei et al. (2021), Sha et al. (2023), Yee et al. (2023), Friedler et al. (2018), Belitz et al. (2021), Nezami et al. (2024), Bayer et al. (2021), Islam et al. (2021), Rzepka et al. (2023, 2022)	36
Race	Physical characteristics, such as skin color, hair texture, and facial features, which can be used to categorize people into different racial groups.	Hutchinson and Mitchell (2019), Han et al. (2024), Di Carlo et al. (2023), Paquette et al. (2020), Hu and Rangwala (2020), Gardner et al. (2023), Li et al. (2021b), Grari et al. (2020), Elglaly and Liu (2023), Wei et al. (2021), Jeong et al. (2021), Friedler et al. (2018), Belitz et al. (2021), Nezami et al. (2024), Bayer et al. (2021), Alghamdi et al. (2022)	17
Ethnicity /Disability	A person's cultural and racial identity can be influenced by factors such as ancestry, language, and shared cultural practices.	Baker and Hawn (2021), Verger et al. (2023), Yu et al. (2020), Paquette et al. (2020), Bridgeman et al. (2012), Deho et al. (2023a), Karumbaiah and Brooks (2021), Dobesh et al. (2023), Doroudi and Brunskill (2019), Bayer et al. (2021), Tschiatschek et al. (2022a)	11
Age	Person's chronological age.	Hutchinson and Mitchell (2019), Han et al. (2024), Verger et al. (2023), Kung and Yu (2020), Deho et al. (2023a), Grari et al. (2020), Riazy et al. (2020), Wei et al. (2021), Yee et al. (2023)	9
Country	The nation or sovereign state in which a person lives or was born.	Baker and Hawn (2021), Bridgeman et al. (2012, 2009), Loukina et al. (2019), Gómez et al. (2021), Islam et al. (2021)	6
Language	Person's native language or the language they speak most fluently.	Sha et al. (2021), Deho et al. (2023a), Sha et al. (2023), Rzepka et al. (2023, 2022)	5
Income Level	A person who has a low income or not.	Yu et al. (2020), Kung and Yu (2020), Arthurs and Alvero (2020)	3
Year of study (First-gen)	In education, it is understood as first-year students — subjects who are confused with information about schools and majors.	Kung and Yu (2020), Yu et al. (2021)	2
Origin	Place or country of a person's birth or ancestry.	Deho et al. (2023b), Riazy et al. (2020)	2
Parental background	Parental Education Background refers to the level of formal education that a child's parents or guardians have achieved	Rzepka et al. (2023, 2022)	2
Home literacy environment	Home Literacy Environment refers to the availability and quality of reading materials, as well as literacy-related activities and interactions within a child's home	Rzepka et al. (2023, 2022)	2

and Support Vector Machine have also been commonly employed. However, there is a comparatively lower number of studies that delve into advanced AI approaches like Long Short-Term Memory or Multi-Layer Perceptron. This trend may be indicative of the nature of the problems under investigation, specifically in the realm of classifying and predicting students' performance based on textual data. It appears that these problems may not derive significant benefits from employing advanced machine learning models. Notably, the application of Neural Network models and ensembling methods has not demonstrated substantial improvements in the outcomes, as reported by previous research (Li et al., 2022)

4.2. RQ2 — What are the investigated problems in the primary studies?

To answer this question, we have extracted information from the primary studies related to issues such as study objectives, research subjects, sensitive variables, and concern for fairness/bias. This information is presented in Table 7, and Table 8 below.

Table 7 lists the study objectives of primary studies that mainly analyze different types of educational data. We can group the study objectives into five main types: (1) prediction of student learning outcomes, (2) promoting systematic education, (3) classifying students based on performance, (4) automatic assessment, and (5) assessing and making decisions. The primary studies mostly propose approaches to ensure the fairness of existing ML/AI systems for the above educational objectives. Typically, the studies conduct experiments on the trade-off between fairness and model performance, evaluating how changes to enhance fairness might affect overall performance. The outcomes can be packaged as methods to reduce or eliminate bias or fairness assurance tools.

Table 8 presents the sensitive variables addressed in primary studies. Sensitive variables refer to attributes of individuals or groups that could potentially lead to bias or unfair treatment when used in ML or AI systems. These variables are considered "sensitive" because they are often linked to social, cultural, or legal factors that can lead to discrimination or unequal treatment. Among the primary studies, gender emerges as the most frequently examined sensitive variable, highlighting ongoing concerns about gender bias in the education sector. Following gender, race and ethnicity/disability are also major focal points, reflecting their critical roles in discussions around societal disparities. Age and country of origin are other significant variables

Table 9
Primary studies with the investigated types of fairness.

Type of fairness		Reference	No. of primary studies
Group fairness	Group fairness	Du et al. (2021), Hutchinson and Mitchell (2019), Saleiro et al. (2019), Di Carlo et al. (2023), Verger et al. (2023), Gándara et al. (2023), Yu et al. (2020), Anderson et al. (2019), Sha et al. (2021), Kung and Yu (2020), Deho et al. (2023b), Mashhadi et al. (2022), Deho et al. (2023a), Elglaly and Liu (2023), Bridgeman et al. (2009), Larson (2017), Riazy et al. (2020), Kulkarni et al. (2021), Gómez et al. (2021), Islam et al. (2021), Alghamdi et al. (2022), Rzepka et al. (2023, 2022)	24
	Equalized odds	Baker and Hawn (2021), Di Carlo et al. (2023), Anderson et al. (2019), Gardner et al. (2019a), Verdugo et al. (2022), Li et al. (2021b), Mashhadi et al. (2022), Jiang and Pardos (2021), Elglaly and Liu (2023), Riazy et al. (2020), Wei et al. (2021), Jeong et al. (2021), Nezami et al. (2024), Alghamdi et al. (2022)	15
	Demographic parity	Du et al. (2021), Baker and Hawn (2021), Lee and Kizilcec (2020), Manisha and Gujar (2018), Han et al. (2024), Sha et al. (2021), Gardner et al. (2019a), Verdugo et al. (2022), Deho et al. (2023b), Mashhadi et al. (2022), Jiang and Pardos (2021), Grari et al. (2020), Sha et al. (2023), Bayer et al. (2021)	14
	Equal opportunity	Baker and Hawn (2021), Lee and Kizilcec (2020), Han et al. (2024), Gándara et al. (2023), Anderson et al. (2019), Verdugo et al. (2022), Jiang and Pardos (2021), Nezami et al. (2024)	8
	Disparate Impact (dispIR)	Verdugo et al. (2022)	1
Individual fairness	Individual fairness	Hutchinson and Mitchell (2019), Saleiro et al. (2019), Kizilcec and Lee (2021), Madaio et al. (2021), Hutt et al. (2019), Hu and Rangwala (2020), Gardner et al. (2019a), Deho et al. (2023b), Mashhadi et al. (2022), Tschiatschek et al. (2022a), Islam et al. (2021)	11
	Fairness through unawareness	Kizilcec and Lee (2021), Hutt et al. (2019), Yu et al. (2021), Jiang and Pardos (2021), Sahlgren (2023)	5
	Counterfactual fairness	Kizilcec and Lee (2021), Hutt et al. (2019)	2
	Procedural fairness	Doroudi and Brunskill (2019), Belitz et al. (2021)	2
Subgroup fairness	Statistical parity	Baker and Hawn (2021), Di Carlo et al. (2023), Verdugo et al. (2022), Wei et al. (2021), Sahlgren (2023), Loukina et al. (2019), Nezami et al. (2024)	7
	Subgroup fairness	Hutchinson and Mitchell (2019), Gardner et al. (2023), Bayer et al. (2021)	3
	Conditional procedure accuracy equality	Mashhadi et al. (2022), Loukina et al. (2019)	2
	Treatment equality	Loukina et al. (2019)	1
	Conditional use accuracy equality	Loukina et al. (2019)	1
	Overall accuracy equality	Loukina et al. (2019)	1

that are scrutinized for their potential to influence AI behavior subtly, impacting how individuals are evaluated or treated by automated systems.

4.3. RQ3 — What fairness and bias definitions are used in the primary studies?

Many fairness definitions are used in the primary studies. Specifically, the fairness definitions mentioned include Demographic parity; Individual fairness; Subgroup fairness; Group fairness; Treatment equality; Counterfactual fairness; Equalized odds; Equal opportunity; Overall accuracy equality; Statistical parity; Conditional procedure accuracy equality; Conditional use accuracy equality; Fairness through unawareness; procedural fairness; and Disparate Impact (dispIR). The number of primary studies mentioning these definitions is detailed in Table 9. Among them, group fairness, demographic parity, and equalized odds are the most common types of fairness to be examined.

Group-based fairness essentially compares the outcome of the classification algorithm for two or more groups defined through the sensitive variables. Individual-based fairness considers the outcome for each participating individual. There is no consensus in literature whether individual or group fairness should be prioritized.

The results of Table 9 show that the definitions used in primary studies mostly focus on the definition of group equity, which can be explained based on the unique characteristics of these studies in the field of education, specifically as follows:"

- Dependency on Distance Metrics: A significant challenge for individual fairness is its heavy reliance on the choice of distance metrics. These metrics themselves can cause fairness issues because they may not accurately reflect the true nature of differences between individuals. For example, two students might have similar scores but come from vastly different social and economic backgrounds, leading to different needs and challenges in their learning.
- The conflicting Treating similar individuals similarly may not yield satisfactory fairness results at the group level. In education, if we focus solely on individual fairness without considering

Table 10
Characteristics of datasets.

Dataset	Ref.	Type	Sources	No. of studies
Students' scores from university courses	Lee and Kizilcec (2020), Yu et al. (2020), Hu and Rangwala (2020), Li et al. (2021b), Yu et al. (2021), Deho et al. (2023b,a), Riazy et al. (2020), Sha et al. (2023)	Closed sources	Not Available	9
MOOC STEM courses	Gardner et al. (2019a), Kung and Yu (2020), Sha et al. (2023), Yee et al. (2023), Gómez et al. (2021), Rzepka et al. (2023, 2022)	Closed source	Not Available	7
USA High School Longitudinal K12 pupils' scores	Gándara et al. (2023), Akgun and Greenhow (2022), Jeong et al. (2021), Sha et al. (2023)	Open source	https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018140	4
Student GPA scores from Chile, USA, etc	Anderson et al. (2019), Verdugo et al. (2022), Jiang and Pardos (2021), Okur et al. (2018)	Closed source	Not Available	4
Foreign language scores	Loukina et al. (2019), Wang et al. (2018)	Closed source	Not available	2
Face attributes dataset	Han et al. (2024), Kulkarni et al. (2021)	Open source	CelebA Dataset: https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html	2
ELS (Education Longitudinal Study) and IPEDS data set	Nezami et al. (2024), Di Carlo et al. (2023)	Open source	(https://www.icpsr.umich.edu/web/ICPSR/studies/4275);(https://surveys.nces.ed.gov/Ipeds/	2
ETS (Educational Testing Service)	Bridgeman et al. (2009)	Closed source	Not available	1
The CAE submitted by applicants to a multi-campus, US public university system	Arthurs and Alvero (2020)	Closed source	Not Available	1
National Student Clearinghouse (NSC)	Hutt et al. (2019)	Closed source	Not Available	1
PISA (the Programme for International Student Assessment) in 65 countries	Li et al. (2021a)	Closed source	Not available	1
Flickr images of people	Kulkarni et al. (2021)	Open source	https://www.flickr.com/photos/tags/images/	1
Open University Learning Analytics Dataset dataset	Verger et al. (2023)	Open source	https://analyse.kmi.open.ac.uk/open_dataset	1
HSLS (High School Longitudinal Study)	Alghamdi et al. (2022)	Closed source	Not available	1
ENEM Dataset	Alghamdi et al. (2022)	Open source	https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem	1

group-level factors such as race, gender, and economic status, we may overlook systemic injustices that minority groups face.

- Imbalanced Data Characteristics: An important reason for using group fairness definitions is that educational datasets often exhibit imbalances regarding sensitive attributes. When different groups have uneven amounts of data, focusing only on individual fairness can exacerbate the injustices faced by underrepresented groups. Thus, using group fairness definitions ensures that all groups are treated equitably and are not disadvantaged by the initial data imbalance.

However, ensuring both individual fairness and group fairness is crucial. This presents an opportunity for researchers to evaluate these aspects further, a promising area for future research to ensure both individual and group fairness, as well as subgroup fairness.

4.4. RQ4. What are the characteristics of the dataset used in the primary studies?

By extracting data from articles, we compare the usage of different types of datasets (25 papers using closed source datasets, while 8 papers using open source datasets). Detailed information about the datasets is presented in Table 10. This table represents the most common variables but is not exhaustive, and some of the papers cited used additional variables. The common type of data used for educational AI/ML systems includes demographic characteristics of students, information about courses, and student's performance in the courses. The dominant use of closed-source datasets can be explained for several reasons:

- Privacy and Sensitivity Concerns: Educational data often contains sensitive information about students, including demographics, academic performance, and behavioral traits. With the recent enforcement of General Data Protection Regulation (GDPR), sharing these types of data is even more restricted.
- Tailor-made data: educational institutions often have specific data requirements tailored to their unique educational settings and objectives. Closed-source datasets can be curated directly by educational organizations' needs.

4.5. RQ5 — What methods are employed for ensuring AI fairness in primary studies?

With the statistical documents, the authors have also used and proposed solutions to detect and ensure fairness. Methods covered

Table 11
Methods ensuring the fairness.

Method	Description	Ref.	No. of primary studies
Computed disparity variables	Learning an unbiased classifier by making the prediction independent of the sensitive attribute. The goal is to maximize the predictor's ability to predict the label, while simultaneously training the adversary to minimize the ability to predict the sensitive attribute from the output predictions (Zhang et al., 2018)	Manisha and Gujar (2018), Verger et al. (2023), Yu et al. (2020), Kung and Yu (2020), Yu et al. (2021), Deho et al. (2023a), Riazy et al. (2020), Kulkarni et al. (2021), Loukina et al. (2019), Gómez et al. (2021)	10
Class Balancing Techniques (CBTs)	To over-sample the class labels with fewer data samples or under-sample the class labels with more data samples to help reach parity in the training data	Han et al. (2024), Sha et al. (2021), Fenu et al. (2022), Gardner et al. (2023), Sha et al. (2023), Nezami et al. (2024)	6
Use tools to detect bias	Use available tools for discrimination discovery and bias mitigation such as AI Fairness 360; Aequitas; Google Analogy Test Set (GATS); Synthetic minority oversampling technique (SMOTE); and Equitable measures	Di Carlo et al. (2023), Mashhadi et al. (2022), Arthurs and Alvero (2020), Rzepka et al. (2023)	4
Slicing analysis	Breaking down performance measures by different dimensions or categories of the data (Sculley et al., 2018)	Hutt et al. (2019), Li et al. (2021b)	2
Model Training Strategy with Adversarial Learning	Adversarial learning is a technique that has been used to attempt to learn bias-free deep representations from biased data (Beutel et al., 2017). Its mission is to enforce the deep representations to be maximally informative for predicting the labels of the main task while minimally discriminative for predicting sensitive attributes (Du et al., 2021)	Jiang and Pardos (2021), Grari et al. (2020)	2
Seldonian Algorithm	Seldonian Algorithm relies on reasonable but false assumptions, such as appeals to the central limit theorem (Thomas et al., 2019)	Li et al. (2021b)	1
The FairProjection	FairProjection is a parallelizable algorithm used to adjust the outputs of machine learning models to ensure group fairness. It employs the ADMM method for optimization and comes with guarantees on sample complexity and convergence rate, ensuring that the final results are fair and efficient	Alghamdi et al. (2022)	1

include Slicing analysis; Seldonian algorithm; Model training strategy with adversarial learning method; Multiple Collaborative Classification Model (MCCM); and Computed disparity metrics, detailed information about the methods is presented in Table 11.

The authors also emphasize interventions that can occur in all stages of the learning model. This includes actions such as auditing the data during the pre-processing stage, scrutinizing the model during the in-process step, and inspecting the prediction output during any post-processing step.

4.6. RQ6- What evaluation metrics are used to assess AI fairness in the primary studies?

To assess fairness in machine learning systems, various measurement parameters are provided, such as the Absolute Area between ROCs (ABROCA), Standardized Mean Difference (SMD), Range and Standard Deviation (STD), Disparate Impact (DI), Statistical Parity Difference (SPD), and other parameters. This parameter is utilized to measure a biased estimate of the probability of correctly classifying a randomly selected data point from the majority group to the non-majority group. Detailed information about the measurement parameters and fairness assessment tools is presented in Table 12.

Among primary studies aimed at improving fairness, the findings also indicate encouraging outcomes. The primary studies underscore the significance of fair evaluation for machine learning systems in the educational domain. These articles highlight the crucial nature of ensuring fairness in machine learning systems implemented in education, proposing methods to assess and evaluate fairness while offering solutions to enhance system fairness. With notably positive results, such as the fairness assessment index exhibiting minimal error (0.003 (Gender) (Hu and Rangwala, 2020); 0.008 (Hutt et al., 2019)), without substantial reductions in model accuracy, and overall model accuracies ranging from 0.6 to 0.84, there are optimistic indications of progress in achieving fairness in machine learning systems applied in education. Lee and Kizilcec (2020), Hutt et al. (2019), Yu et al. (2020), Hu and Rangwala (2020), Gardner et al. (2019a), Li et al. (2021b), Yu et al. (2021), Jiang and Pardos (2021), Loukina et al. (2019), Li et al. (2021a), Okur et al. (2018)

4.7. RQ7- What methodologies are employed to assess fairness and performance in AI/ML models within the primary studies?

The primary studies reported various settings and evaluation approaches for their proposed fairness assurance methods. We can summarize commonalities among the empirical studies into four categories: experiment settings, bench-marking methods, evaluation methods, and evaluation results.

Regarding experiment settings, studies predominantly use hyperparameters by manually setting values of variables like learning rate, number of epochs, batch size, etc, before training models. Only seven studies use default settings for their experiments. We note that such settings are crucial as they determine the model's initial conditions and its adaptability through parameter tuning.

Benchmarking provides a comparative framework for model assessment. The most common evaluation approach here is to compare the fairness-fixed models with their original versions (33 studies). We also found studies that compare the fairness achieved with results from manual judgment by humans (5 studies).

Table 12
Metrics reported for fairness evaluation.

Metrics	Description	Type of fairness	Ref.	No. of studies
Absolute Between-ROC Area (ABROCA)	Measures the absolute value of the area between the baseline group ROC curve ROCb and those of one or more comparison groups ROCc.	Group fairness	Verger et al. (2023), Hutt et al. (2019), Sha et al. (2021), Gardner et al. (2019a), Deho et al. (2023b), Sha et al. (2022), Riazy et al. (2020), Jeong et al. (2021), Sha et al. (2023), Rzepka et al. (2023, 2022)	11
Disparate Impact (DI)	Let us A represent the protected attribute, with 1 denoting the privileged group and 0 denoting the unprivileged group. Let y denote the actual label and $\hat{y}$ denote the predicted label, where 1 is the favorable class and 0 is the unfavorable class) DI= $\frac{P(\hat{y}=1\mid A=0)}{P(\hat{y}=1\mid A=1)}$	Group fairness	Hutchinson and Mitchell (2019), Friedler et al. (2019), Manisha and Gujar (2018), Grari et al. (2020), Saleiro et al. (2019), Anderson et al. (2019), Verdugo et al. (2022), Du et al. (2021), Riazy et al. (2020), Deho et al. (2023b)	10
DemPI, Disparate Impact, Equalized Opportunity, Equalized Odd, Sufficiency	– DemPI: $\lvert \Pr(C(X)=1\mid A=1)-\Pr(C(X)=1\mid A=0)\rvert \le \alpha$ – Disparate Impact: $\frac{\Pr(C(X)=1\mid A=1)}{\Pr(C(X)=1\mid A=0)} > \alpha$ – Equalized Opportunity: $\lvert \Pr(C(X)=1\mid A=1, Y=1)\ -\Pr(C(X)=1\mid A=0, Y=1)\rvert \le \alpha$ – Equalized Odd: $\lvert \Pr(C(X)=1\mid A=1, Y=y)\ -\Pr(C(X)=1\mid A=0, Y=y)\rvert \le \alpha$ – Sufficiency: $\lvert \Pr(y=1\mid C(X)=1, A=1)\ -\Pr(Y=1\mid C=1, A=0)\rvert \le \alpha$	Most quantitative definitions of fairness	Han et al. (2024), Gándara et al. (2023), Verdugo et al. (2022), Rzepka et al. (2022, 2023)	5
Range and Standard Deviation (SD)	Range (i.e., max value–min value) and SD of each metric over all the groups could be deemed as a group fairness measure, with lower values corresponding to less disparity between race groups and therefore greater fairness	Group fairness	Marcinkowski et al. (2020), Deho et al. (2023a), Jiang and Pardos (2021)	3
Statistical Parity Difference (SPD)	Lets A represent the protected attribute, with 1 denoting the privileged group and 0 denoting the unprivileged group. Let y denote the actual label and $\hat{y}$ denote the predicted label, where 1 is the favorable class and 0 is the unfavorable class) $SPD = P(\hat{y}=1\mid A=0) - P(\hat{y}=1\mid A=1)$	Group fairness	Kulkarni et al. (2021), Du et al. (2021), Alghamdi et al. (2022)	2
Mean Equalized Odds (MEO)	Evaluate fairness using three fairness metrics across demographic groups: measure multi-class performance, we extend the definition of MEO as: MEO= $\max_{i\in Y} \max_{s_1,s_2\in S}(\lvert \mathrm{TPR}_i(s_1)-\mathrm{TPR}_i(s_2)\rvert + \lvert \mathrm{FPR}_i(s_1)-\mathrm{FPR}_i(s_2)\rvert)/2.$ Where $\mathrm{TPR}_i(s) = P(Y_b = i \mid Y = i, S = s)$ $\mathrm{FPR}_i(s) = P(Y_b = i \mid Y \neq i, S = s)$	Group fairness	Alghamdi et al. (2022), Di Carlo et al. (2023)	2
AUC Gap (Area Under the Receiver Operating Characteristic Curve)	$AUC(f(\theta)) = \int_0^1 TPR(FPR(f_t(\theta)))dt$ where t indicates a prediction threshold applied to the predictions of the model (i.e. using the decision rule $f_t(\theta, x) = 1(f(\theta, x) \ge t)$ and TPR, FPR are the true positive rate and false positive rates, respectively. AUC Gap: To measure fairness across subgroups, a metric that accounts for the disparities in predictive performance across a set of arbitrarily many (possibly overlapping) subgroups G. The AUC Gap as: $max_{(g,g'\in G)}\lvert E_{D_k}[f(\theta(D_{k,g}))] - E_{D_k}[f(\theta(D_{k,g'}))]\rvert$ where $D_{k,g}$ and $D_{k,g'}$ indicate the subset of the data in group g and g', respectively.	Subgroup fairness	Han et al. (2024), Gardner et al. (2023)	2
Predictive parity (PP)	which is satisfied if both subgroups have equal predictive positive value $P(Y=1\mid R=1, S=s1) = P(Y=1\mid R=1, S=s2)$	Group fairness	Rzepka et al. (2022, 2023)	2
Predictive Equality (PE)	A classifier satisfied if both subgroups have equal false positive rates $P(R=1\mid Y=0, S=s1) = P(R=1\mid Y=0, S=s2)$	Group fairness	Rzepka et al. (2022, 2023)	2
MADD (Model Absolute Density Distance)	Measures the difference between the probability distributions of the model's outcomes $D^a_{G_0}$ and $D^a_{G_1}$: the density vectors of the respective groups G_0 and G_1 of the sensitive feature a. $MADD(D^a_{G_0}, D^a_{G_1}) = \sum_{(k=0)}^{m} \lvert d^a_{G_0,k} - d^a_{G_1,k}\rvert$ (Verger et al., 2023)	Group fairness	Verger et al. (2023)	1
Standardized Mean Difference (SMD)	Standardized mean difference (SMD) is a standard measure used to evaluate the fairness of automated scoring engines. SMD for each group is the average difference between such standardized human and system scores within this group (System score–human score).	Group fairness	Loukina et al. (2019)	1

Regarding the evaluation methods, we documented various techniques used for evaluating the fairness and performance of AI models. Cross-validation was a predominant method (22 studies), facilitating the evaluation of the model by training on multiple data subsets. Comparisons between human and system performance (5 studies) were critical for contextualizing AI/ML model outcomes within human performance benchmarks. Slicing analysis (3 studies) specifically examined model performance across different demographic or data segments, addressing potential biases in model predictions.

Table 13
Techniques for evaluating AI/ML fairness and performance.

Type of evaluation		References	No. of primary studies
Experiment setting	Hyperparameter	Saleiro et al. (2019), Di Carlo et al. (2023), Kulkarni et al. (2021), Marcinkowski et al. (2020), Sha et al. (2022), Yu et al. (2021), Loukina et al. (2019), Verger et al. (2023), Sha et al. (2023), Grari et al. (2020), Xiang et al. (2022), Gardner et al. (2023), Hutt et al. (2019), Hu and Rangwala (2020), Jeong et al. (2021), Li et al. (2021b), Deho et al. (2023a), Gándara et al. (2023), Yee et al. (2023), Verdugo et al. (2022), Riazy et al. (2020), Yu et al. (2020), Manisha and Gujar (2018), Nezami et al. (2024), Doroudi and Brunskill (2019), Arthurs and Alvero (2020), Alghamdi et al. (2022), Anderson et al. (2019), Rzepka et al. (2023)	29
	Default setting	Tschiatschek et al. (2022b), Paquette et al. (2020), Lee and Kizilcec (2020), Deho et al. (2023b), Kung and Yu (2020), Friedler et al. (2019), Rzepka et al. (2022)	7
Benchmarking with	Empirical comparing with the original model	Saleiro et al. (2019), Di Carlo et al. (2023), Sha et al. (2021, 2022), Wei et al. (2021), Yu et al. (2021), Loukina et al. (2019), Han et al. (2024), Verger et al. (2023), Sha et al. (2023), Friedler et al. (2019), Grari et al. (2020), Anderson et al. (2019), Xiang et al. (2022), Gardner et al. (2023, 2019a), Kung and Yu (2020), Hu and Rangwala (2020), Li et al. (2021b), Deho et al. (2023a,b), Gándara et al. (2023), Yee et al. (2023), Lee and Kizilcec (2020), Yu et al. (2020), Jiang and Pardos (2021), Manisha and Gujar (2018), Nezami et al. (2024), Doroudi and Brunskill (2019), Alghamdi et al. (2022), Arthurs and Alvero (2020), Rzepka et al. (2023, 2022)	33
	Comparing to ground truth	Marcinkowski et al. (2020), Bridgeman et al. (2012), Clauser et al. (2002b), Bridgeman et al. (2009), Tschiatschek et al. (2022b), Arthurs and Alvero (2020)	6
Evaluation methods	Cross-validation	Saleiro et al. (2019), Di Carlo et al. (2023), Wei et al. (2021), Yu et al. (2021), Sha et al. (2023), Friedler et al. (2019), Grari et al. (2020), Gardner et al. (2023), Kung and Yu (2020), Jeong et al. (2021), Li et al. (2021b), Deho et al. (2023a), Gándara et al. (2023), Yee et al. (2023), Verdugo et al. (2022), Riazy et al. (2020), Yu et al. (2020), Manisha and Gujar (2018), Nezami et al. (2024), Alghamdi et al. (2022), Arthurs and Alvero (2020), Rzepka et al. (2023)	22
	Other	Kulkarni et al. (2021), Deho et al. (2023b), Hu and Rangwala (2020), Jiang and Pardos (2021), Tschiatschek et al. (2022b), Rzepka et al. (2022)	6
	Compare between Human and Machine	Bridgeman et al. (2009), Clauser et al. (2002b), Bridgeman et al. (2012), Loukina et al. (2019), Marcinkowski et al. (2020)	5
	Slicing analysis	Hutt et al. (2019), Gardner et al. (2019a), Anderson et al. (2019)	3
Evaluation results	Improve Fairness	Sha et al. (2021), Wei et al. (2021), Sha et al. (2023), Marcinkowski et al. (2020), Holstein et al. (2019), Sha et al. (2022), Loukina et al. (2019), Verger et al. (2023), Li et al. (2021b), Jeong et al. (2021), Kung and Yu (2020), Bridgeman et al. (2012), Gardner et al. (2019a), Lee and Kizilcec (2020), Riazy et al. (2020), Jiang and Pardos (2021), Manisha and Gujar (2018), Nezami et al. (2024), Bridgeman et al. (2009), Doroudi and Brunskill (2019), Tschiatschek et al. (2022b), Arthurs and Alvero (2020), Rzepka et al. (2023)	23
	Improve both Fairness and Performance	Saleiro et al. (2019), Di Carlo et al. (2023), Han et al. (2024), Deho et al. (2023a), Hutt et al. (2019), Grari et al. (2020), Xiang et al. (2022), Gardner et al. (2023), Paquette et al. (2020), Alghamdi et al. (2022)	10
	Other	Anderson et al. (2019), Yu et al. (2021), Hu and Rangwala (2020), Deho et al. (2023b), Gándara et al. (2023), Yu et al. (2020), Clauser et al. (2002b), Madaio et al. (2021), Rzepka et al. (2022)	9

Regarding the evaluation results, the effectiveness of these methodologies in enhancing fairness and performance was directly reflected in the study outcomes: 23 studies specifically noted enhancements in fairness, underscoring the targeted efforts to mitigate biases within AI/ML models, 10 studies reported simultaneous improvements in fairness and performance, indicating successful optimization across both metrics.

Table 14
Summary of findings.

Research Questions	Findings
RQ1	The primary studies utilized a variety of AI models, including Naive Bayes, Random Forest, Support Vector Machines, Decision Trees, Logistic Regression, K-Nearest Neighbor, Long Short-Term Memory, Natural Language Processing, Neural Networks, Random Models, Ordinary Least Squares Regression, Seldonian Algorithm, among others.
RQ2	Major problems investigated encompass predicting/detecting student performance, classifying GPA scores, evaluation and decision-making processes, automatic grading, and ML equity in education.
RQ3	Diverse definitions of fairness and bias were employed, including Demographic Parity, Individual Fairness, Subgroup Fairness, Group Fairness, Treatment Equality, Counterfactual Fairness, Equalized Odds, Equal Opportunity, Overall Accuracy Equality, Statistical Parity, Conditional Procedure Accuracy Equality, Conditional Use Accuracy Equality, Fairness through Unawareness, Procedural Fairness, and Disparate Impact.
RQ4	The datasets are varied, comprising both open datasets and proprietary datasets from universities. Notable datasets include GPA, Course Data, Foreign Language Scores, University Data Scores, College Applications, and Nationally Representative Student Data.
RQ5	A variety of methods to detect and ensure fairness in AI systems are implemented. They focused on evaluating disparities in sensitive attributes to identify and mitigate bias, and they are applied at different stages of the learning model, including pre-processing, in-process, and post-processing steps.
RQ6	Various metrics were used to evaluate fairness in AI, such as the Absolute Area between ROCs, Standardized Mean Difference, Range and SD, and others like DemPI, dispIR, eqOPP, eqODD, Suffic, and AUC Gap. These metrics typically measure the probability of correctly classifying a data point from the majority to the non-majority group.
RQ7	The primary studies explored various techniques for evaluating AI/ML fairness and performance. These techniques included model settings (default and hyperparameters), benchmarking (comparing to ground truth and original models), evaluation methods (cross-validation, slicing analysis, and human vs. system comparisons), and evaluation results (improvements in fairness, performance, or both).

Besides, various other results highlighted the nuanced outcomes dependent on specific model applications and settings. Detailed information is presented in Table 13

5. Discussion

This section summarizes our findings (Section 5.1), discusses threats to validities (Section 5.2), and implications for future research (Section 5.3)

5.1. Summary of findings

Through this comprehensive survey, we have not only thoroughly cataloged the existing research on the fairness of machine learning systems in education but also organized it systematically. In a departure from previous studies, we have cataloged the AI/ML algorithms that have been used, highlighting the definitions of fairness and bias in question and specifically pointing to datasets commonly used by researchers when investigating fairness in machine learning systems in education. Furthermore, our survey includes a comprehensive and clear presentation of the methods and fairness evaluation metrics of these techniques. The answers to seven RQs are summarized in Table 14.

Interesting observations are noted:

- The top 3 machine learning algorithms were widely used in education including Logistic Regression (20); Random Forest (15); Decision Tree (14). It seems that studies about fairness education sectors somehow lag behind software engineering research, where deep learning is the contemporary trend.
- The top 3 of the sensitives variables were widely used in education including gender (36); race (17); ethnicity/Disability (11). Domain-specific sensitive variables include year of study and accent voice.
- The top 3 definitions of fairness used in ML for education are Group fairness (24); Equalized odds (14); Demographic parity (14).
- The datasets used are mostly closed datasets collected from universities or courses managed by responsible institutions.
- The top 3 of Fairness measures are ABROCA (11); DI (10); and DemPI/EO (5).
- For model evaluation techniques, the researchers mainly use hyper-parameter techniques (29) with cross-validation methods (22).

Mehrabi et al. in their highly cited survey on ML bias in 2019, classify fair AI systems based on their approaches in either pre-processing, in-processing, or post-processing phases. Fairness definitions are grouped into individual, subgroup, or group levels. We adopted many classifications from this study in our scheme Fig. 3. Besides sharing similar observations on the current trends and challenges of studying fairness, we also classify empirical evaluations of fairness assurance approaches. Future work can refer directly to our paper for algorithms, datasets, metrics, experiment settings and benchmarking approaches. Zhen et al. in their survey paper on fairness testing, reported the increased number of studies after 2018 (Petersen et al., 2015), which also aligns with the demographic of our selected papers Fig. 2. The authors show a relatively large number of studies on fairness testing for deep learning systems. In the education sector, the dominant evaluation is still on more traditional ML algorithms. Compared to the review by Memarian et al. we offer a more in-depth and comprehensive presentation of both descriptive and technical definitions of fairness (Hardt et al., 2016). Our analysis for RQ2, RQ3 and RQ7 could contribute better to the link between a theoretical view and an operational view of AI fairness in education.

5.2. Threats to validity

Like other secondary types of research, this study can suffer from several threats to validity:

- Construct validity primarily addresses the potential bias in the selection of primary papers. In our efforts to mitigate this concern, we aimed to encompass all relevant portals. Google Scholar, recognized as the most extensive online repository for academic publications, offers a comprehensive array of scholarly literature, encompassing articles, theses, books, and conference papers. According to Zhang et al. (2011b), IEEE Xplore, and ACM Digital Library are identified as the main search portals in software and system engineering. Our approach involved following a systematic protocol to diminish threats of bias in the selection and extraction process. To minimize individual bias, the majority of the work was conducted collaboratively by at least two authors. The inter-rater agreement was tested between the first and the third author during the pilot test, which reached an acceptable score for Kappa test between two raters (0.78). When an indecisive case happens (during paper selection, data extraction or quality assessment), all three authors met to resolve the cases.
- Conclusion validity is associated with the risk of incorrect data extraction or missing studies. The accuracy of the final results hinges on the decisions made by the authors who conducted the search process. To address this limitation, we took measures by

Table A.15
Quality assessment.

Study	Ref.	Q1	Q2	Q3	Q4	Q5	Q6	Q7a	Q7b	Q8	Q9	Q10	Score
1	Saleiro et al. (2019)	3	3	3	3	3	3		3	3	3	3	3
2	Kulkarni et al. (2021)	3	3	3	3	3	3		3	3	3	3	3
3	Sha et al. (2021)	3	3	3	3	3	3		3	3	3	3	3
4	Du et al. (2021)	3	3	3	3	3	3		3	3	3	3	3
5	Marcinkowski et al. (2020)	3	3	3	3	3	3	3		3	3	3	3
6	Holstein et al. (2019)	3	3	3	3	3	3		3	3	3	3	3
7	Sha et al. (2022)	3	3	3	3	3	3		3	3	3	3	3
8	Sha et al. (2023)	3	3	3	3	3	3		3	3	3	3	3
9	Yu et al. (2021)	3	3	3	3	3	3		3	3	3	3	3
10	Loukina et al. (2019)	3	3	3	3	3	3		3	3	3	3	3
11	Han et al. (2024)	3	3	3	3	3	3		3	3	3	3	3
12	Verger et al. (2023)	3	3	3	3	3	3	3		3	3	3	3
13	Wei et al. (2021)	3	2	3	3	3	3		3	3	3	3	2.9
14	Hutchinson and Mitchell (2019)	3	3	2	3	3	3	3		3	3	3	2.9
15	Friedler et al. (2019)	3	3	3	3	2	3		3	3	3	3	2.9
16	Grari et al. (2020)	3	3	3	3	2	3		3	3	3	3	2.9
17	Baker and Hawn (2021)	3	3	3	3	3	3	3		2	3	3	2.9
18	Anderson et al. (2019)	3	3	2	3	3	3		3	3	3	3	2.9
19	Xiang et al. (2022)	3	3	2	3	3	3		3	3	3	3	2.9
20	Bridgeman et al. (2012)	3	3	3	3	2	3		3	3	3	3	2.9
21	Gardner et al. (2023)	3	3	3	3	3	3	2		3	3	3	2.9
22	Hutt et al. (2019)	3	3	3	3	3	3		3	3	3	2	2.9
23	Gardner et al. (2019b)	3	3	3	3	3	3		3	3	3	2	2.9
24	Kung and Yu (2020)	3	3	3	3	3	3		3	2	3	3	2.9
25	Hu and Rangwala (2020)	3	3	3	3	3	3		3	2	3	3	2.9
26	Dobesh et al. (2023)	3	3	3	3	3	3	3		2	3	3	2.9
27	Jeong et al. (2021)	3	2	3	3	3	3		3	3	3	3	2.9
28	Li et al. (2021b)	3	3	3	3	3	2	3		3	3	3	2.9
29	Deho et al. (2023a)	3	3	3	3	2	3		3	3	3	3	2,9
30	Feffer et al. (2023)	3	3	3	3	3	3	3		2	3	3	2.9
31	Deho et al. (2023b)	3	3	3	3	3	3		3	2	3	3	2.9
32	Gándara et al. (2023)	2	3	3	3	3	3	3		3	3	3	2.9
33	Yee et al. (2023)	3	3	3	2	3	3		3	3	3	3	2.9
34	Rzepka et al. (2023)	3	3	2	3	3	3		3	3	3	3	2.9
35	Mashhadi et al. (2022)	3	3	3	3	3	2	3		3	2	3	2.8
36	Bogina et al. (2022)	3	3	3	3	2	3		3	2	3	3	2.8
37	Holstein and Doroudi (2021b)	3	3	3	3	3	3	2		3	2	3	2.8
38	Arthurs and Alvero (2020)	3	3	3	3	3	3		3	1	3	3	2.8
39	Rzepka et al. (2022)	3	3	3	3	3	3	3		1	3	3	2.8
40	Lee and Kizilcec (2020)	3	3	3	3	3	3		2	2	3	3	2.8
41	Fenu et al. (2022)	3	3	3	3	3	2	3		2	3	3	2.8
42	Verdugo et al. (2022)	3	3	3	3	3	3		3	2	3	2	2.8
43	Riazy et al. (2020)	3	3	2	2	3	3		3	3	3	3	2.8
44	Larson (2017)	3	3	3	2	2	3	3		3	3	3	2.8
45	Huggins-Manley et al. (2022)	3	3	3	3	3	2	3		3	3	2	2.8
46	Yu et al. (2020)	3	3	3	2	3	3		3	2	3	3	2.8
47	Jiang and Pardos (2021)	3	3	3	2	3	3		3	2	3	3	2.8
48	Clauser et al. (2002b)	2	3	3	2	3	3	3		3	3	3	2.8
49	Alghamdi et al. (2022)	3	3	3	3	3	3		2	2	3	3	2.8
50	Di Carlo et al. (2023)	3	3	2	2	3	3		3	3	3	3	2.8
51	Nezami et al. (2024)	3	3	1	2		3	3	3	3	3	3	2.7
52	Manisha and Gujar (2018)	3	3	1	3	2	3		3	3	3	3	2.7
53	Kizilcec and Lee (2021)	3	3	3	3	2	3		3	3	1	3	2.7
54	Madaio et al. (2021)	3	3	3	3	3	3	2		3	1	3	2.7
55	Bridgeman et al. (2009)	3	3	3	3	3	3	0		0	3	3	2.7
56	Doroudi and Brunskill (2019)	3	3	2	3	2	2		3	3	3	3	2.7
57	Elglaly and Liu (2023)	3	3	3	3	3	2		2	2	3	3	2.7
58	Paquette et al. (2020)	3	3	3	2	2	3		3	2	3	3	2.7
59	Tschiatschek et al. (2022b)	3	3	3	1	3	1		3	3	3	3	2.6
60	Karumbaiah and Brooks (2021)	3	3	2	3	2	3	2		2	3	3	2.6
61	Akgun and Greenhow (2022)	3	3	2	2	2	3	0	2	2	2	3	2.4
62	Sahlgren (2023)	3	3	3	1	3	1	3		2	2	2	2.3
63	Matias and Zipitria (2023)	3	3	1	1	2	2	3		2	1	3	2.1

clearly defining inclusion and exclusion criteria. The inclusion and exclusion criteria were developed based on the third author's extensive experience in conducting systematic mapping studies and adjusted during the pilot search. Throughout the steps of selection and classification, the potential influence of bias on result interpretation was acknowledged. This limitation was mitigated by providing clear and comprehensive descriptions of every activity performed during these steps. Elberzhager et al. (2012).

- Internal validity pertains to the extraction and data analysis processes. A notable concern involves the selection of low-quality papers, specifically those lacking peer reviews in Arxiv. To mitigate this risk, we implemented a quality assessment check for all included papers. Another potential challenge lies in the possibility of misclassification during data extraction. However, we addressed this limitation by developing a classification scheme based on widely accepted guidelines.
- External threats to validity in a mapping study are associated with the generalizability of the results (Easterbrook et al., 2008). It is important to note that the validity of the conclusions drawn in

Table B.16
Type of fairness in primary studies.

Type of fairness	Definition	Reference	No. of primary studies
Group fairness	AI systems should treat all groups in the overall population equally. This means that the AI system should not discriminate against any group as a whole.	Du et al. (2021), Hutchinson and Mitchell (2019), Saleiro et al. (2019), Madaio et al. (2021), Verger et al. (2023), Gándara et al. (2023), Yu et al. (2020), Anderson et al. (2019), Sha et al. (2021), Kung and Yu (2020), Deho et al. (2023b), Mashhadi et al. (2022), Deho et al. (2023a), Elglaly and Liu (2023), Bridgeman et al. (2009), Larson (2017), Riazy et al. (2020), Alghamdi et al. (2022), Gómez et al. (2021), Islam et al. (2021), Rzepka et al. (2023, 2022)	22
Demographic parity	Measure of fairness in an AI system ensures that the model produces similar outcomes for different groups of people, regardless of their protected characteristics such as race, gender, or age.	Du et al. (2021), Baker and Hawn (2021), Lee and Kizilcec (2020), Manisha and Gujar (2018), Han et al. (2024), Sha et al. (2021), Gardner et al. (2019a), Verdugo et al. (2022), Deho et al. (2023b), Mashhadi et al. (2022), Jiang and Pardos (2021), Grari et al. (2020), Sha et al. (2023), Bayer et al. (2021)	14
Equalized odds	An AI system should have the same false positive and false negative rates for different subgroups. For example, if an AI system is used to predict whether a person will default on a loan, it should have the same false positive and false negative rates for people of different races.	Baker and Hawn (2021), Gándara et al. (2023), Anderson et al. (2019), Gardner et al. (2019a)–(Li et al., 2021b; Mashhadi et al., 2022; Jiang and Pardos, 2021; Elglaly and Liu, 2023; Riazy et al., 2020; Wei et al., 2021; Jeong et al., 2021; Nezami et al., 2024; Alghamdi et al., 2022)	13
Individual fairness	Similar individuals should be treated similarly by an AI system, i.e., people with similar backgrounds, genders, ages, should be treated similarly.	Hutchinson and Mitchell (2019), Saleiro et al. (2019), Kizilcec and Lee (2021), Madaio et al. (2021), Hutt et al. (2019), Hu and Rangwala (2020), Gardner et al. (2019a), Deho et al. (2023b), Mashhadi et al. (2022), Tschiatschek et al. (2022a), Islam et al. (2021)	11
Equal opportunity	An AI system should have the same true positive rate for different subgroups. For example, if an AI system is used to predict whether a person will be successful in a job, it should have the same true positive rate for men and women.	Baker and Hawn (2021), Lee and Kizilcec (2020), Han et al. (2024), Gándara et al. (2023), Anderson et al. (2019), Verdugo et al. (2022), Jiang and Pardos (2021), Nezami et al. (2024)	8
Statistical parity	An AI system should have the same rate of positive outcomes for different subgroups. For example, if an AI system is used to make loan approval decisions, it should have the same approval rate for people of different races.	Baker and Hawn (2021), Verdugo et al. (2022), Wei et al. (2021), Sahlgren (2023), Loukina et al. (2019), Nezami et al. (2024)	6
Fairness through unawareness	An AI system can be fair even if it is not aware of certain protected variables, such as race or gender. This means that the AI system can make decisions that are fair to all individuals or groups without considering these variables.	Kizilcec and Lee (2021), Hutt et al. (2019), Yu et al. (2021), Jiang and Pardos (2021), Sahlgren (2023)	5
Subgroup fairness	AI systems should treat different subgroups within the overall population equally. For example, if an AI system is used to make hiring decisions, it should not discriminate against certain subgroups, such as women or people of a certain race.	Hutchinson and Mitchell (2019), Gardner et al. (2023), Bayer et al. (2021)	3
Counterfactual fairness	An AI system should make the same decisions for individuals or groups that are similar in all relevant ways, except for a single protected variable. For example, if an AI system is used to predict whether a person will default on a loan, it should make the same prediction for two people who are identical in all relevant ways, except for their race.	Kizilcec and Lee (2021), Hutt et al. (2019)	2
Conditional procedure accuracy equality	An AI system should have the same accuracy for different subgroups when the same procedure is used. For example, if an AI system is used to predict whether a person will default on a loan, it should have the same accuracy for people of different races when the same prediction procedure is used.	Mashhadi et al. (2022), Loukina et al. (2019)	2
Procedural fairness	An AI system that makes decisions should be fair. This means that the process should be transparent, unbiased, and free from discrimination.	Doroudi and Brunskill (2019), Belitz et al. (2021)	2
Disparate Impact (dispIR)	AI systems can have an unfairly negative impact on certain groups, even if it is not intentionally biased against those groups. This can occur if the AI system is trained on biased data, which can lead to unfair outcomes for certain groups.	Verdugo et al. (2022)	1
Treatment equality	An AI system should provide the same treatment to individuals or groups that are similar in relevant ways. This means that the AI system should not discriminate against individuals or groups based on characteristics such as race or gender.	Loukina et al. (2019)	1
Conditional use accuracy equality	An AI system should have the same accuracy for different subgroups when the same use case is considered. For example, if an AI system is used to predict whether a person will be successful in a job, it should have the same accuracy for men and women when the same job is considered.	Loukina et al. (2019)	1
Overall accuracy equality	An AI system should have the same overall accuracy for different subgroups. This means that the AI system should make the same number of correct and incorrect decisions for all subgroups.	Loukina et al. (2019)	1

this paper is specific to this systematic mapping study. As such, this threat is not applicable in this context.

5.3. Implication for future research

Major issues that require researchers to focus more on improving fairness in machine learning systems in the field of education can be listed as:

1. Balancing the tradeoff between fairness and model performance: as a main target of fairness research, this direction should be continually explored to optimize both the performance and fairness of an AI model in education.
2. Combining fairness assurance for multiple sensitive variables: current empirical studies are skewed towards fairness assurance for a single. There is a research gap in studies that explore the combination of multiple sensitive attributes (e.g., race and gender simultaneously), evaluating how combined biases affect model outcomes and proposing integrated solutions
3. Combining different fairness constructs and metrics: there is currently a lack of comprehensive studies that address the combination of different fairness metrics, particularly in how they interact or conflict. Empirical research is needed to explore ways to handle combinations of fairness metrics or to implement a significant meta-review of these measures. Such studies would help categorize specific differences, ideological trade-offs, and preferences, which could guide the development of more universally acceptable fairness frameworks.
4. Lack of open-source benchmarking instrument — The field is currently deficient in open-source tools that can effectively identify a wide range of demographic biases across different technologies. Future research should aim to develop new datasets and benchmarks that can evaluate AI systems across diverse demographic groups.
5. Fairness testing and assurance for more modern ML/ AI models. There is a notable deficiency in the study of fairness in newer and more complex ML models, such as Large Language Models. Future empirical work should focus on developing specific fairness tests and assurance practices for these models, evaluating their biases in realistic settings, and proposing solutions to mitigate any identified unfairness

6. Conclusions

In our paper, we present an extensive review of scholarly works on machine learning software fairness within the education domain. Despite our exhaustive efforts to include recent research up until December 2023, we encountered a limitation in the volume of relevant studies, identifying 63 primary research papers. We have tried different approaches to ensure the comprehensiveness of our selection. Utilizing these studies, we aimed to address the seven questions introduced at the outset. Our approach involved synthesizing the machine learning algorithms applied in educational settings, identifying key issues highlighted in these studies, compiling the types of fairness approaches employed, examining the datasets referenced, focusing on the fairness evaluation methods discussed by the authors; and, finally, analyzing the outcomes reported in these studies to gain a comprehensive perspective on fairness in machine learning systems in education.

Given the identified gaps in research on the fairness of machine learning systems in the education domain, our upcoming focus will be on investigating ways to eliminate potential bias starting from the preprocessing stage. This involves evaluating the influence of attributes in the dataset on system fairness and exploring methods to alter their impact. This will be achieved by identifying and assigning suitable weights to the attributes in the dataset.

CRediT authorship contribution statement

Nga Pham: Writing – review & editing, Writing – original draft, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Pham Ngoc Hung:** Visualization, Supervision, Methodology, Conceptualization. **Anh Nguyen-Duc:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The author declare that we do not have any financial and personal relationships with any organizations that could inappropriately influence this work.

During the preparation of this work we used ChatGPT version 3.5 and Grammarly services to edit and improve our writing, i.e. rephrasing sentences, fixing typos and language mistakes. After using this tool/service, we reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Appendix A

See Table A.15.

Appendix B

See Table B.16.

Data availability

No data was used for the research described in the article.

References

Akgun, S., Greenhow, C., 2022. Artificial intelligence in education: Addressing ethical challenges in K-12 settings. AI Ethics 2 (3), 431–440. http://dx.doi.org/10.1007/s43681-021-00096-7.

Alghamdi, W., Hsu, H., Jeong, H., Wang, H., Michalak, P., Asoodeh, S., Calmon, F., 2022. Beyond adult and COMPAS: Fair multi-class prediction via information projection. Adv. Neural Inf. Process. Syst. 35, 38747–38760, URL https://papers.nips.cc/paper_files/paper/2022/hash/fd5013ea0c3f96931dec77174eaf9d80-Abstract-Conference.html.

Anderson, H., Boodhwani, A., Baker, R., 2019. Assessing the fairness of graduation predictions. In: 12th International Conference on Educational Data Mining.

Angelov, P., Soares, E., Jiang, R., Arnold, N., Atkinson, P., 2021. Explainable artificial intelligence: An analytical review. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. http://dx.doi.org/10.1002/widm.1424.

Arthurs, N., Alvero, A., 2020. Whose truth is the "ground truth"? College admissions essays and bias in word vector evaluation methods.

Baeza-Yates, R., 2018. Bias on the web. Commun. ACM 61, 54–61. http://dx.doi.org/10.1145/3209581.

Baker, R.S., Hawn, A., 2021. Algorithmic bias in education. Int. J. Artif. Intell. Educ. http://dx.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: Lecture Notes in Computer Science. vol. 12749, Springer International Publishing, pp. 71–75. http://dx.doi.org/10.1007/978-3-030-78270-12.

Belitz, C., Jiang, L., Bosch, N., 2021. Automating procedurally fair feature selection in machine learning. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 379–389. http://dx.doi.org/10.1145/3461702.3462585.

Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y., 2019. AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM J. Res. Dev. 63 (4/5), 4:1–4:15. http://dx.doi.org/10.1147/JRD.2019.2942287, URL https://ieeexplore.ieee.org/document/8843908. Conference Name: IBM Journal of Research and Development.

Berg, V., Birkeland, J., Nguyen-Duc, A., Pappas, I.O., Jaccheri, L., 2018. Software startup engineering: A systematic mapping study. J. Syst. Softw. 144, 255–274. http://dx.doi.org/10.1016/j.jss.2018.06.043.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A., 2021. Fairness in criminal justice risk assessments: The state of the art. Sociol. Methods Res. 50 (1), 3–44. http://dx.doi.org/10.1177/0049124118782533.

Beutel, A., Chen, J., Zhao, Z., Chi, E.H., 2017. Data decisions and theoretical implications when adversarially learning fair representations. http://dx.doi.org/10.48550/arXiv.1707.00075, arXiv.

Blyth, C.R., 1972. On Simpson's paradox and the sure-thing principle. J. Amer. Statist. Assoc. 67 (338), 364–366. http://dx.doi.org/10.1080/01621459.1972.10482387.

Bogina, V., Hartman, A., Kuflik, T., Shulner-Tal, A., 2022. Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics. Int. J. Artif. Intell. Educ. 32 (3), 808–833. http://dx.doi.org/10.1007/s40593-021-00248-0.

Bridgeman, B., Trapani, C., Attali, Y., 2009. Considering Fairness and Validity in Evaluating Automated Scoring. Tech. Rep..

Bridgeman, B., Trapani, C., Attali, Y., 2012. Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. Appl. Meas. Educ. 25 (1), 27–40. http://dx.doi.org/10.1080/08957347.2012.635502.

Casas-Roma, J., Conesa, J., 2021. A literature review on artificial intelligence and ethics in online learning. pp. 111–131. http://dx.doi.org/10.1016/B978-0-12-823410-5.00006-1.

Caton, S., Haas, C., 2020. Fairness in machine learning: A survey. URL http://arxiv.org/abs/2010.04053. arXiv:2010.04053 [cs, stat].

Chan, K.S., Zary, N., 2019. Applications and challenges of implementing artificial intelligence in medical education: Integrative review. JMIR Med. Educ. 5 (1), e13930. http://dx.doi.org/10.2196/13930.

Chen, L., Chen, P., Lin, Z., 2020a. Artificial intelligence in education: A review. IEEE Access 8, 75264–75278. http://dx.doi.org/10.1109/ACCESS.2020.2988510.

Chen, X., Xie, H., Zou, D., Hwang, G.-J., 2020. Application and theory gaps during the rise of artificial intelligence in education. Comput. Educ. Artif. Intell. 1, 100002. http://dx.doi.org/10.1016/j.caeai.2020.100002.

Chen, Z., Zhang, J.M., Hort, M., Harman, M., Sarro, F., 2024. Fairness testing: A comprehensive survey and analysis of trends. ACM Trans. Softw. Eng. Methodol. http://dx.doi.org/10.1145/3652155, URL https://dl.acm.org/doi/10.1145/3652155. Just Accepted.

Chen, Z., Zhang, J.M., Sarro, F., Harman, M., 2023. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. ACM Trans. Softw. Eng. Methodol. 32 (4), 106:1–106:30. http://dx.doi.org/10.1145/3583561, URL https://dl.acm.org/doi/10.1145/3583561.

Chouldechova, A., 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv URL http://arxiv.org/abs/1610.07524.

Cico, O., Jaccheri, L., Nguyen-Duc, A., Zhang, H., 2021. Exploring the intersection between software industry and software engineering education - a systematic mapping of software engineering trends. J. Syst. Softw. 172, http://dx.doi.org/10.1016/j.jss.2020.110736.

Clauser, B.E., Kane, M.T., Swanson, D.B., 2002a. Validity issues for performance-based tests scored with computer-automated scoring systems. Appl. Meas. Educ. 15 (4), 413–432. http://dx.doi.org/10.1207/S15324818AME1504-05.

Clauser, B., Kane, M., Swanson, D., 2002b. Validity issues for performance-based tests scored with computer-automated scoring systems. Appl. Meas. Educ. - APPL MEAS EDUC 15, 413–432. http://dx.doi.org/10.1207/S15324818AME1504-05.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A., 2017. Algorithmic decision making and the cost of fairness. http://dx.doi.org/10.1145/3097983.309809, arXiv.

Deho, O.B., Joksimovic, S., Li, J., Zhan, C., Liu, J., Liu, L., 2023a. Should learning analytics models include sensitive attributes? explaining the why. IEEE Trans. Learn. Technol. 16 (4), 560–572. http://dx.doi.org/10.1109/TLT.2022.3226474.

Deho, O.B., Joksimovic, S., Liu, L., Li, J., Zhan, C., Liu, J., 2023b. Assessing the fairness of course success prediction models in the face of (un)equal demographic group distribution. In: Proceedings of the Tenth ACM Conference on Learning @ Scale. In: L@S '23, Association for Computing Machinery, pp. 48–58. http://dx.doi.org/10.1145/3573051.3593381.

Di Carlo, F., Nezami, N., Anahideh, H., Asudeh, A., 2023. FairPilot: An explorative system for hyperparameter tuning through the lens of fairness. http://dx.doi.org/10.48550/arXiv.2304.04679, arXiv URL http://arxiv.org/abs/2304.04679.

Dobesh, S.J., Miller, T., Newman, P., Liu, Y., Elglaly, Y.N., 2023. Towards machine learning fairness education in a natural language processing course. In: Proceedings of the 54th ACM Technical Symposium on Computer Science Education. In: SIGCSE 2023, vol. 1, pp. 312–318. http://dx.doi.org/10.1145/3545945.3569802.

Doroudi, S., Brunskill, E., 2019. Fairer but not fair enough on the equitability of knowledge tracing. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge. pp. 335–339. http://dx.doi.org/10.1145/3303772.3303838.

Došilović, F.K., Brčić, M., Hlupić, N., 2018. Explainable artificial intelligence: A survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics. MIPRO, pp. 0210–0215. http://dx.doi.org/10.23919/MIPRO.2018.8400040.

Du, M., Yang, F., Zou, N., Hu, X., 2021. Fairness in deep learning: A computational perspective. IEEE Intell. Syst. http://dx.doi.org/10.1109/mis.2020.3000681.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R., 2011. Fairness through awareness. arXiv URL http://arxiv.org/abs/1104.3913.

Easterbrook, S., Singer, J., Storey, M.-A., Damian, D., 2008. Selecting empirical methods for software engineering research. In: Shull, F., Singer, J., Sjøberg, D.I.K. (Eds.), Guide to Advanced Empirical Software Engineering. Springer, pp. 285–311.

Elberzhager, F., Münch, J., Nha, V.T.N., 2012. A systematic mapping study on the combination of static and dynamic quality assurance techniques. Inf. Softw. Technol. 54 (1), 1–15.

Elglaly, Y.N., Liu, Y., 2023. Promoting machine learning fairness education through active learning and reflective practices. ACM SIGCSE Bull. 55 (3), 4–6. http://dx.doi.org/10.1145/3610585.3610589.

Engberg, M.E., 2004. Improving intergroup relations in higher education: A critical examination of the influence of educational interventions on racial bias. Rev. Educ. Res. 74 (4), 473–524. http://dx.doi.org/10.3102/00346543074004473.

Farnadi, G., Babaki, B., Getoor, L., 2018. Fairness in relational domains. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. In: ACM, New Orleans LA USA, pp. 108–114. http://dx.doi.org/10.1145/3278721.3278733.

Feffer, M., Heidari, H., Lipton, Z.C., 2023. Moral machine or Tyranny of the majority? http://dx.doi.org/10.48550/arXiv.2305.17319, arXiv, May 26.

Fenu, G., Galici, R., Marras, M., 2022. Experts' view on challenges and needs for fairness in artificial intelligence for education. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (Eds.), Artificial Intelligence in Education. In: Lecture Notes in Computer Science, vol. 243–255, Springer International Publishing, http://dx.doi.org/10.1007/978-3-031-11644-5-20.

Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D., 2018. A comparative study of fairness-enhancing interventions in machine learning. Tech. Rep., http://dx.doi.org/10.48550/arXiv.1802.04422.

Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D., 2019. A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. In: FAT* '19, Association for Computing Machinery, pp. 329–338. http://dx.doi.org/10.1145/3287560.3287589.

Friedman, B., Nissenbaum, H., 1996. Bias in computer systems. ACM Trans. Inf. Syst. 14 (3), 330–347. http://dx.doi.org/10.1145/230538.230561.

Gándara, D., Anahideh, H., Ison, M.P., Tayal, A., 2023. Inside the Black Box: Detecting and Mitigating Algorithmic Bias across Racialized Groups in College Student-Success Prediction. Tech. Rep., http://dx.doi.org/10.48550/arXiv.2301.03784, arXiv.

Gardner, J., Brooks, C., Baker, R., 2019a. Evaluating the fairness of predictive student models through slicing analysis. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge. Tempe AZ USA, pp. 225–234. http://dx.doi.org/10.1145/3303772.3303791.

Gardner, J., Brooks, C., Baker, R., 2019b. Evaluating the fairness of predictive student models through slicing analysis. http://dx.doi.org/10.1145/3303772.3303791.

Gardner, J., Yu, R., Nguyen, Q., Brooks, C., Kizilcec, R., 2023. Cross-institutional transfer learning for educational models: Implications for model performance, fairness, and equity. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. FAccT '23, Association for Computing Machinery, pp. 1664–1684. http://dx.doi.org/10.1145/3593013.3594107.

Gómez, E., Zhang, C.S., Boratto, L., Salamó, M., Marras, M., 2021. The winner takes it all: Geographic imbalance and provider (un)fairness in educational recommender systems. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 1808–1812. http://dx.doi.org/10.1145/3404835.3463235.

Grari, V., Ruf, B., Lamprier, S., Detyniecki, M., 2020. Achieving fairness with decision trees: An adversarial approach. Data Sci. Eng. 5 (2), 99–110. http://dx.doi.org/10.1007/s41019-020-00124-2.

Grgić-Hlača, N., Zafar, M.B., Gummadi, K.P., Weller, A., 2016. The case for process fairness in learning: Feature selection for fair decision making. URL http://www.mlandthelaw.org/papers/grgic.pdf.

Han, X., Chi, J., Chen, Y., Wang, Q., Zhao, H., Zou, N., Hu, X., 2024. FFB: A fair fairness benchmark for in-processing group fairness methods. http://dx.doi.org/10.48550/arXiv.2306.09468, arXiv URL http://arxiv.org/abs/2306.09468.

Hardt, M., Price, E., Srebro, N., 2016. Equality of opportunity in supervised learning. http://dx.doi.org/10.48550/arXiv.1610.02413.

Holstein, K., Doroudi, S., 2021a. Equity and Artificial Intelligence in Education: Will 'AIEd' Amplify or Alleviate Inequities in Education?. Tech. Rep., http://dx.doi.org/10.48550/arXiv.2104.12920, arXiv.

Holstein, K., Doroudi, S., 2021b. Equity and artificial intelligence in education: Will 'AIEd' amplify or alleviate inequities in education? ArXiv URL https://www.semanticscholar.org/paper/Equity-and-Artificial-Intelligence-in-Education%3A-or-Holstein-Doroudi/bd31e6a1580143bf3605b1a6a762ca7461d57ada.

Holstein, K., Vaughan, J.W., III, H.D., Dudík, M., Wallach, H., 2019. Improving fairness in machine learning systems: What do industry practitioners need? In: Proc. 2019 CHI Conf. Hum. Factors Comput. Syst.. pp. 1–16. http://dx.doi.org/10.1145/3290605.3300830.

Hort, M., Chen, Z., Zhang, J.M., Harman, M., Sarro, F., 2024. Bias mitigation for machine learning classifiers: A comprehensive survey. ACM J. Responsib. Comput. 1 (2), 11:1–11:52. http://dx.doi.org/10.1145/3631326, URL https://dl.acm.org/doi/10.1145/3631326.

Hort, M., Kechagia, M., Sarro, F., Harman, M., 2022. A survey of performance optimization for mobile applications. IEEE Trans. Softw. Eng. 48 (8), 2879–2904. http://dx.doi.org/10.1109/TSE.2021.3071193, URL https://ieeexplore.ieee.org/document/9397392. Conference Name: IEEE Transactions on Software Engineering.

Hort, M., Zhang, J.M., Sarro, F., Harman, M., 2021. Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. In: ESEC/FSE 2021, Association for Computing Machinery, New York, NY, USA, pp. 994–1006. http://dx.doi.org/10.1145/3468264.3468565.

Hu, Q., Rangwala, H., 2020. Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students. Tech. Rep., International Educational Data Mining Society, URL https://eric.ed.gov/?q=bias+real+life+examples&id=ED608050.

Huggins-Manley, A.C., Booth, B.M., D'Mello, S.K., 2022. Toward argument-based fairness with an application to AI-enhanced educational assessments. J. Educ. Meas. 59 (3), 362–388. http://dx.doi.org/10.1111/jedm.12334.

Hughes, G., 2013. Racial justice, hegemony, and bias incidents in U.S. higher education. Multicult. Perspect. http://dx.doi.org/10.1080/15210960.2013.809301.

Huston, T., 2006. Race and gender bias in higher education: Could faculty course evaluations impede further progress toward parity? Seattle J. Soc. Justice 4 (2), URL https://digitalcommons.law.seattleu.edu/sjsj/vol4/iss2/34.

Hutchinson, B., Mitchell, M., 2019. 50 years of test (un)fairness: Lessons for machine learning. In: Proc. Conf. Fairness Account. Transpar.. pp. 49–58. http://dx.doi.org/10.1145/3287560.3287600.

Hutt, S., Gardner, M., Duckworth, A.L., D'Mello, S., 2019. Evaluating fairness and generalizability in models predicting on-time graduation from college applications. In: EDM.

Islam, M.T., Fariha, A., Meliou, A., 2021. Through the data management lens: Experimental analysis and evaluation of fair classification. ArXiv.

Jamieson, S., 2004. Likert scales: How to (ab)use them. Med. Educ. 38 (12), 1217–1218. http://dx.doi.org/10.1111/j.1365-2929.2004.02012.x.

Jeong, H., Wu, M.D., Dasgupta, N., Médard, M., Calmon, F., 2021. Who Gets the Benefit of the Doubt? Racial Bias in Machine Learning Algorithms Applied to Secondary School Math Education. In: Neural Inf. Process. Syst. NeurIPS 2021 Workshop Math AI Educ. MATHAI4ED.

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. http://dx.doi.org/10.1145/3461702.3462623.

Karumbaiah, S., Brooks, J., 2021. How colonial continuities underlie algorithmic injustices in education. In: 2021 Conference on Research in Equitable and Sustained Participation in Engineering, Computing, and Technology. RESPECT, pp. 1–6. http://dx.doi.org/10.1109/RESPECT51740.2021.9620605.

Kizilcec, R.F., Lee, H., 2021. Algorithmic fairness in education. arXiv URL http://arxiv.org/abs/2007.05443.

Kleinberg, J., Ludwig, J., Mullainathan, S., Rambachan, A., 2018. Algorithmic fairness. AEA Pap. Proc. 108, 22–27. http://dx.doi.org/10.1257/pandp.20181018.

Kulkarni, O.N., Patil, V., Singh, V.K., Atrey, P.K., 2021. Accuracy and fairness in pupil detection algorithm. In: 2021 IEEE Seventh International Conference on Multimedia Big Data. BigMM, pp. 17–24. http://dx.doi.org/10.1109/BigMM52142.2021.00011.

Kung, C., Yu, R., 2020. Interpretable models do not compromise accuracy or fairness in predicting college success. In: Proceedings of the Seventh ACM Conference on Learning @ Scale. In: L@S '20, Association for Computing Machinery, pp. 413–416. http://dx.doi.org/10.1145/3386527.3406755.

Kusner, M.J., Loftus, J., Russell, C., Silva, R., 2017. Counterfactual fairness. In: Adv. Neural Inf. Process. Syst., vol. 30.

Larson, B.N., 2017. Gender as a variable in natural-language processing: Ethical considerations. In: EthNLP@EACL. http://dx.doi.org/10.18653/v1/W17-1601.

Lee, H., Kizilcec, R.F., 2020. Evaluation of fairness trade-offs in predicting student success. arXiv.

Li, Y., Meng, L., Chen, L., Yu, L., Wu, D., Zhou, Y., Xu, B., 2022. Training data debugging for the fairness of machine learning software. In: Proceedings of the 44th International Conference on Software Engineering. ICSE '22, Association for Computing Machinery, New York, NY, USA, pp. 2215–2227. http://dx.doi.org/10.1145/3510003.3510091.

Li, L., Sha, L., Li, Y., Raković, M., Rong, J., Joksimovic, S., Selwyn, N., Gašević, D., Chen, G., 2023. Moral machines or tyranny of the majority? A systematic review on predictive bias in education. In: LAK23: 13th International Learning Analytics and Knowledge Conference. In: LAK2023, Association for Computing Machinery, New York, NY, USA, pp. 499–508. http://dx.doi.org/10.1145/3576050.3576119.

Li, X., Song, D., Han, M., Zhang, Y., Kizilcec, R.F., 2021a. On the limits of algorithmic prediction across the globe. ArXiv.

Li, C., Xing, W., Leite, W., 2021b. Yet another predictive model? Fair predictions of students' learning outcomes in an online math learning platform. In: LAK21: 11th International Learning Analytics and Knowledge Conference. In: LAK21, Association for Computing Machinery, pp. 572–578. http://dx.doi.org/10.1145/3448139.3448200.

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. http://dx.doi.org/10.18653/v1/W19-4401.

Madaio, M., Blodgett, S.L., Mayfield, E., Dixon-Román, E., 2021. Beyond 'Fairness:' Structural (In)justice Lenses on AI for Education. Tech. Rep., http://dx.doi.org/10.48550/arXiv.2105.08847, arXiv.

Mahmud, A., 2020. Racial disparities in student outcomes in British higher education: Examining mindsets and bias. Teach. High. Educ. http://dx.doi.org/10.1080/13562517.2020.1796619.

Mahmud, A., Gagnon, J., 2023. Racial disparities in student outcomes in British higher education: Examining mindsets and bias. Teach. Higher Educ. 28 (2), 254–269, Publisher: SRHE Website _eprint: http://dx.doi.org/10.1080/13562517.2020.1796619.

Manisha, P., Gujar, S., 2018. A Neural Network Framework for Fair Classifier. Tech. Rep., http://dx.doi.org/10.48550/arXiv.1811.00247, arXiv.

Marcinkowski, F., Kieslich, K., Starke, C., Lünich, M., 2020. Implications of AI (un-)fairness in higher education admissions: the effects of perceived AI (un-)fairness on exit, voice and organizational reputation. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. In: FAT* '20, Association for Computing Machinery, pp. 122–130. http://dx.doi.org/10.1145/3351095.3372867.

Mashhadi, A., Zolyomi, A., Quedado, J., 2022. A case study of integrating fairness visualization tools in machine learning education. In: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems. In: CHI EA '22, Association for Computing Machinery, New York, NY, USA, pp. 1–7. http://dx.doi.org/10.1145/3491101.3503568.

Matias, A., Zipitria, I., 2023. Promoting ethical uses in artificial intelligence applied to education. In: Frasson, C., Mylonas, P., Troussas, C. (Eds.), Augmented Intelligence and Intelligent Tutoring Systems. In: Lecture Notes in Computer Science, Springer Nature Switzerland, pp. 604–615. http://dx.doi.org/10.1007/978-3-031-32883-1-53.

Mayfield, E., Madaio, M., Prabhumoye, S., Gerritsen, D., McLaughlin, B., Dixon-Román, E., Black, A.W., 2019. Equity beyond bias in language technologies for education. In: Yannakoudakis, H., Kochmar, E., Leacock, C., Madnani, N., Pilán, I., Zesch, T. (Eds.), Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, Florence, Italy, pp. 444–460. http://dx.doi.org/10.18653/v1/W19-4446, URL https://aclanthology.org/W19-4446.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2019. A survey on bias and fairness in machine learning. arXiv URL http://arxiv.org/abs/1908.09635.

Memarian, B., Doleck, T., 2023. Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. Comput. Educ. Artif. Intell. 5, 100152. http://dx.doi.org/10.1016/j.caeai.2023.100152.

Mester, T., 2023. Statistical bias types explained (with examples) - part1. URL https://data36.com/statistical-bias-types-explained/. (Accessed: 13 Nov 2023).

Minnaert, A., Janssen, P.J., 1997. Bias in the assessment of regulation activities in studying at the level of higher education. Eur. J. Psychol. Assess. 13 (2), 99–108. http://dx.doi.org/10.1027/1015-5759.13.2.99.

Nematzadeh, A., Ciampaglia, G.L., Menczer, F., Flammini, A., 2018. How algorithmic popularity bias hinders or promotes quality. Sci. Rep. 8 (1), 15951. http://dx.doi.org/10.1038/s41598-018-34203-2.

Nezami, N., Haghighat, P., Gándara, D., Anahideh, H., 2024. Assessing disparities in predictive modeling outcomes for college student success: The impact of imputation techniques on model performance and fairness. Educ. Sci. 14 (2), 136. http://dx.doi.org/10.3390/educsci14020136, URL https://www.mdpi.com/2227-7102/14/2/136.

Nguyen-Duc, A., Cruzes, D.S., Conradi, R., 2015. The impact of global dispersion on coordination, team performance and software quality – a systematic literature review. In: Inf. Softw. Technol.. 57, pp. 277–294. http://dx.doi.org/10.1016/j.infsof.2014.06.002.

Okur, E., Aslan, S., Alyuz, N., Esme, A.A., Baker, R.S., 2018. Role of socio-cultural differences in labeling students' affective states. In: Artificial Intelligence in Education. In: Lecture Notes in Computer Science, vol. 10947, Springer International Publishing, pp. 367–380. http://dx.doi.org/10.1007/978-3-319-93843-1-27.

Olteanu, A., Castillo, C., Diaz, F., Kıcıman, E., 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. Front. Big Data 2, URL https://www.frontiersin.org/articles/10.3389/fdata.2019.00013.

Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., Baker, R., 2020. Who's learning? Using demographics in EDM research. J. Educ. Data Min. 12 (3), 1–30. http://dx.doi.org/10.5281/zenodo.4143612.

Pessach, D., Shmueli, E., 2022. A review on fairness in machine learning. ACM Comput. Surv. 55 (3), 51:1–51:44. http://dx.doi.org/10.1145/3494672.

Petersen, K., Vakkalanka, S., Kuzniarz, L., 2015. Guidelines for conducting systematic mapping studies in software engineering: An update. Inf. Softw. Technol. 64, http://dx.doi.org/10.1016/j.infsof.2015.03.007.

Riazy, S., Simbeck, K., Schreck, V., 2020. Fairness in learning analytics: Student at-risk prediction in virtual learning environments. In: Proceedings of the 12th International Conference on Computer Supported Education. SCITEPRESS - Science and Technology Publications, Prague, Czech Republic, pp. 15–25. http://dx.doi.org/10.5220/0009324100150025.

Rzepka, N., Fernsel, L., Müller, H.-G., Simbeck, K., Pinkwart, N., 2023. Unbias me! Mitigating Algorithmic Bias for Less-studied Demographic Groups in the Context of Language Learning Technology. OSF, http://dx.doi.org/10.35542/osf.io/qa9vz, URL https://osf.io/qa9vz.

Rzepka, N., Simbeck, K., Müller, H.-G., Pinkwart, N., 2022. Fairness of in-session dropout prediction:. In: Proceedings of the 14th International Conference on Computer Supported Education. SCITEPRESS - Science and Technology Publications, Online Streaming, — Select a Country —, pp. 316–326. http://dx.doi.org/10.5220/0010962100003182, URL https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010962100003182.

Sahlgren, O., 2023. The politics and reciprocal (re)configuration of accountability and fairness in data-driven education. Learn. Media Technol. 48 (1), 95–108. http://dx.doi.org/10.1080/17439884.2021.1986065.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K.T., Ghani, R., 2019. Aequitas: A bias and fairness audit toolkit. http://dx.doi.org/10.48550/arXiv.1811.05577, arXiv URL http://arxiv.org/abs/1811.05577.

Saxena, N.A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D.C., Liu, Y., 2019. How do fairness definitions fare?: Examining public attitudes towards algorithmic definitions of fairness. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. In: ACM, Honolulu HI USA, pp. 99–106. http://dx.doi.org/10.1145/3306618.3314248.

Sculley, D., Snoek, J., Wiltschko, A., Rahimi, A., 2018. Winner's curse? On pace, progress, and empirical rigor. URL https://openreview.net/forum?id=rJWF0Fywf.

Sha, L., Gašević, D., Chen, G., 2023. Lessons from debiasing data for fair and accurate predictive modeling in education. Expert Syst. Appl. 228, 120323. http://dx.doi.org/10.1016/j.eswa.2023.120323.

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 Trans. Learn. Technol. 15 (4), 481–492. http://dx.doi.org/10.1109/TLT.2022.3196278.

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: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (Eds.), Artificial Intelligence in Education. Springer International Publishing, Cham, pp. 381–394. http://dx.doi.org/10.1007/978-3-030-78292-4_31.

Smith, H., 2020. Algorithmic bias: Should students pay the price? AI Soc. 35 (4), 1077–1078. http://dx.doi.org/10.1007/s00146-020-01054-3.

Soremekun, E., Papadakis, M., Cordy, M., Traon, Y.L., 2022. Software fairness: An analysis and survey. http://dx.doi.org/10.48550/arXiv.2205.08809, arXiv URL http://arxiv.org/abs/2205.08809.

Suresh, H., Guttag, J., 2019. A framework for understanding unintended consequences of machine learning. arXiv URL https://www.semanticscholar.org/paper/A-Framework-for-Understanding-Unintended-of-Machine-Suresh-Guttag/61c425bdda0e053074e96c3e6761ff1d7e0dd469.

Suresh, H., Guttag, J.V., 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. http://dx.doi.org/10.1145/3465416.3483305.

Tang, Z., Zhang, J., Zhang, K., 2023. What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. ACM Comput. Surv. 55 (13s), 299:1–299:37. http://dx.doi.org/10.1145/3597199, URL https://dl.acm.org/doi/10.1145/3597199.

Thomas, P.S., da Silva, B.C., Barto, A.G., Giguere, S., Brun, Y., Brunskill, E., 2019. Preventing undesirable behavior of intelligent machines. Science 366 (6468), 999–1004. http://dx.doi.org/10.1126/science.aag3311.

Tschiatschek, S., Knobelsdorf, M., Singla, A., 2022a. Equity and Fairness of Bayesian Knowledge Tracing. In: Proc. 15th Int. Conf. Educ. Data Min., http://dx.doi.org/10.48550/ARXIV.2205.02333.

Tschiatschek, S., Knobelsdorf, M., Singla, A., 2022b. Equity and fairness of Bayesian knowledge tracing. In: The 15th International Conference on Educational Data Mining. EDM 2022, Durham, UK, pp. 578–582. http://dx.doi.org/10.48550/arXiv.2205.02333.

Verdonk, P., Benschop, Y.W.M., de Haes, H.C.J.M., Lagro-Janssen, T.L.M., 2009. From gender bias to gender awareness in medical education. Adv. Health Sci. Educ. Theory Pract. 14 (1), 135–152. http://dx.doi.org/10.1007/s10459-008-9100-z.

Verdugo, J.V., Gitiaux, X., Ortega, C., Rangwala, H., 2022. FairEd: A systematic fairness analysis approach applied in a higher educational context. In: LAK22: 12th International Learning Analytics and Knowledge Conference. In: LAK22, Association for Computing Machinery, pp. 271–281. http://dx.doi.org/10.1145/3506860.3506902.

Verger, M., Lallé, S., Bouchet, F., Luengo, V., 2023. Is your model 'MADD'? A novel metric to evaluate algorithmic fairness for predictive student models. http://dx.doi.org/10.5281/zenodo.8115786.

Verma, S., Rubin, J., 2018. Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. FairWare '18, Association for Computing Machinery, New York, NY, USA, pp. 1–7. http://dx.doi.org/10.1145/3194770.3194776.

Wan, M., Zha, D., Liu, N., Zou, N., 2023. In-processing modeling techniques for machine learning fairness: A survey. ACM Trans. Knowl. Discov. Data 17 (3), 35:1–35:27. http://dx.doi.org/10.1145/3551390, URL https://dl.acm.org/doi/10.1145/3551390.

Wang, Z., Zechner, K., Sun, Y., 2018. Monitoring the performance of human and automated scores for spoken responses. Lang. Test. 35 (1), 101–120. http://dx.doi.org/10.1177/0265532216679451.

Wei, D., Ramamurthy, K.N., Calmon, F.d.P., 2021. Optimized score transformation for consistent fair classification. http://dx.doi.org/10.48550/arXiv.1906.00066, arXiv URL http://arxiv.org/abs/1906.00066.

Xiang, F., Zhang, X., Cui, J., Carlin, M., Song, Y., 2022. Algorithmic bias in a student success prediction models: Two case studies. In: 2022 IEEE International Conference on Teaching, Assessment and Learning for Engineering. TALE, pp. 310–315. http://dx.doi.org/10.1109/TALE54877.2022.00058.

Yee, M., Roy, A., Perdue, M., Cuevas, C., Quigley, K., Bell, A., Rungta, A., Miyagawa, S., 2023. AI-assisted analysis of content, structure, and sentiment in MOOC discussion forums. Front. Educ. 8, http://dx.doi.org/10.3389/feduc.2023.1250846, URL https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2023.1250846/full. Publisher: Frontiers.

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. Virtual Event Germany, pp. 91–100. http://dx.doi.org/10.1145/3430895.3460139.

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. Tech. Rep., International Educational Data Mining Society, URL https://eric.ed.gov/?ft=on&q=decent&id=ED608066.

Zhai, X., Chu, X., Chai, C.S., Jong, M.S.Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., Li, Y., 2021. A review of artificial intelligence (AI) in education from 2010 to 2020. Complexity 2021 (1), 8812542. http://dx.doi.org/10.1155/2021/8812542, URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/8812542.

Zhang, K., Aslan, A., 2021. AI technologies for education: Recent research and future directions. Comput. Educ. Artif. Intell. 2, 100025. http://dx.doi.org/10.1016/j.caeai.2021.100025.

Zhang, H., Babar, M.A., Tell, P., 2011b. Identifying relevant studies in software engineering. Inf. Softw. Technol. 53 (6), 625–637.

Zhang, B.H., Lemoine, B., Mitchell, M., 2018. Mitigating unwanted biases with adversarial learning. CoRR abs/1801.07593, arXiv:1801.07593.

Nga Pham is a lecturer in the Faculty of Information Technology at Dainam University in Hanoi, Vietnam. She is currently pursuing her Ph.D. in computer science at the Faculty of Information Technology, VNU University of Engineering and Technology, also located in Hanoi, Vietnam. Her diverse research interests include Software Engineering, Machine Learning, Artificial Intelligence (AI), Data Science, and Education Applications.

Hung Pham Ngoc received his B.S. degree from the University of Engineering and Technology, Vietnam National University, Hanoi in 2002, M.S. and PhD. degrees from Japan Advanced Institute of Science and Technology in 2006 and 2009, respectively. He is now an Associate Professor at the University of Engineering and Technology, Vietnam National University, Hanoi. His research interests include software verification and testing, quality assurance for machine learning systems, model checking, program analysis, and software evolution. He has published more than 50 articles in his field in both international and domestic journals and conference proceedings. In addition to his professional activities, he is actively involved in social initiatives aimed at promoting learning, research, and the application of Information Technology among young people. He is an experienced expert in managing and implementing numerous IT projects, as well as in the national digital transformation process. His contributions to the IT field have been recognized with numerous prestigious awards.

Anh Nguyen-Duc is a Professor at the University of South Eastern Norway and the Norwegian University of Science and Technology. His research interests span various human and business aspects of software development. His latest work focuses on software engineering for artificial intelligence systems, including generative AI and AI fairness. He has authored over 150 peer-reviewed publications in top-tier journals, conferences, and workshops related to software engineering and closely related fields. He has chaired ten international conferences and served on more than 30 program committees. From 2017 to 2020, he was recognized as one of the Top 500 Researchers in Norway across all fields.