A Systematic Review and Taxonomy of Bias Detection in Machine Learning for Scholarship Allocation in Educational Decision Systems

Manga Ibrahim, Tami Bitrus Small, Yusufu Gambo

Department of Computer Science, Adamawa State University, Mubi

DOI: https://doi.org/10.51584/IJRIAS.2026.11060047

Received: 24 May 2026; Accepted: 30 May 2026; Published: 20 June 2026

ABSTRACT

The rapid adoption of Machine Learning (ML) in educational decision-making, particularly scholarship allocation, has raised critical concerns about fairness, bias propagation, and institutional accountability. However, existing research on how bias is detected, measured, and mitigated in these systems remains fragmented across methods, domains, and evaluation practices. This study conducts a PRISMA-guided systematic review to explore machine learning-driven bias detection approaches in education and scholarship, integrating findings from Scopus, Web of Science, IEEE Xplore, and the ACM Digital Library. The synthesis examines bias variants, fairness metrics, methodological trends, and evaluation practices across selected studies. Findings show that representation and measurement bias dominate the literature, while label and deployment biases are less explored. Statistical and group fairness metrics are most commonly used, whereas causal inference and in-model fairness approaches remain underdeveloped. Major methodological limitations identified include the scarcity of high-quality datasets, inconsistent reporting practices, limited reproducibility, and inadequate evaluation of fairness across the entire model lifecycle. To address these issues, the study proposes a domain-specific taxonomy for scholarship allocation that structures fairness analysis across bias source, detection stage, method type, fairness metric, and educational context. The framework consolidates fragmented evidence and highlights research gaps in causal fairness, deployment monitoring, and longitudinal bias analysis.

Keywords: Algorithmic bias detection, Educational resource allocation, Fairness, taxonomy, Machine learning, Scholarship allocation, Educational decision systems

INTRODUCTION

ML has evolved from a supporting analytical technique into a core component of modern educational decision systems. Its applications now span admissions, grading, recommendation systems, and financial aid allocation, where algorithmic outputs increasingly determine access to educational opportunities and social mobility. In scholarship allocation, these systems play a particularly critical role because they directly influence equity, institutional diversity, and long-term socioeconomic outcomes. As a result, the integration of ML into scholarship decision-making introduces both efficiency gains and significant ethical responsibilities.

Recent developments in Artificial Intelligence (AI) and data-driven optimization have enabled systems to process complex educational datasets and extract meaningful patterns from large-scale information environments (Dritsas & Trigka, 2025; Wei et al., 2025). These capabilities support predictive modeling, academic performance estimation, and automated resource allocation. Consequently, ML-driven systems are often promoted as scalable and consistent alternatives to human decision-making in education (Baker & Hawn, 2022; Pham et al., 2025).

Despite these benefits, the application of machine learning in scholarship allocation continues to raise concerns related to fairness, transparency, and accountability. Educational datasets are inherently socio-technical and

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

often reflect historical inequalities, institutional biases, and uneven access to educational resources. When such datasets are used for model training, they may reproduce or amplify existing disparities. Prior research confirms that algorithmic systems frequently inherit biases embedded in training data, leading to systematic unfairness if not properly addressed (Mehrabi et al., 2021; Siddique et al., 2024; Pessach & Shmueli, 2022).

These challenges are particularly pronounced in developing and resource-constrained educational contexts, where data limitations, infrastructure gaps, and contextual variability complicate AI deployment (Bali, 2026; Siddique et al., 2024). In such environments, scholarship allocation systems must balance technical performance with institutional constraints, cultural expectations, and equity requirements. Therefore, fairness in ML-driven scholarship systems must be understood as both a technical and socio-ethical necessity (Alvarez et al., 2024; Venkatasubbu & Krishnamoorthy, 2022).

At the same time, advances in explainable AI (XAI) and human-centered modeling have improved transparency in algorithmic decision-making. Contemporary research increasingly emphasizes that fairness cannot be achieved at a single stage of development but must be integrated across the entire ML lifecycle, including data preparation, model design, evaluation, and deployment (Choraś et al., 2020; Lee et al., 2024; Yan & Liu, 2025). These developments motivate deeper investigation into how bias emerges, how it can be detected, and how equitable scholarship systems can be designed.

An effective ML-based scholarship allocation system should provide transparent and reliable decision support, ensuring fair candidate evaluation while accounting for socioeconomic diversity and contextual disadvantage. However, current implementations often fall short. Bias may arise at multiple stages of the pipeline, including data collection, feature engineering, model training, and deployment. Historical inequalities may lead models to favor privileged groups, while proxy variables may inadvertently encode sensitive attributes such as income level, gender, or geographic location (Fazelpour & Danks, 2021; Van Giffen et al., 2022).

A further challenge is the lack of consensus on how fairness should be defined in educational decision-making. Scholarship allocation typically requires balancing merit, financial need, and institutional diversity goals. However, general-purpose fairness frameworks often fail to capture this contextual complexity, resulting in inconsistent evaluation practices and limited applicability in real-world educational systems (Corbett-Davies et al., 2023).

Transparency and accountability remain additional concerns. Although explainability techniques exist, they are not consistently integrated into operational scholarship systems. This constrains stakeholders, including students, institutions, and policymakers, from effectively interpreting or contesting algorithmic decisions. Lack of transparency also reduces trust in automated allocation systems and raises concerns about procedural justice (Chinta et al., 2024; Bogina et al., 2022).

Existing literature on AI fairness and educational technology provides valuable insights but remains fragmented. Most fairness research focuses on general domains such as hiring, healthcare, or predictive analytics, with limited integration into education-specific scholarship allocation contexts (Albaroudi et al., 2024; Pessach & Shmueli, 2022). Educational AI studies, on the other hand, tend to emphasize learning outcomes, system performance, or personalization rather than fairness in high-stakes financial decisions (Baker & Hawn, 2022; Li et al., 2023).

More importantly, there is currently no unified, education-specific taxonomy that systematically integrates bias types, fairness metrics, detection approaches, dataset characteristics, and deployment stages within scholarship allocation workflows. Existing studies typically address isolated components of fairness rather than the full ML lifecycle. This fragmentation limits comparability, reproducibility, and methodological rigor across studies (Siddique et al., 2024; Van Giffen et al., 2022).

Additional gaps exist in how datasets and sensitive attributes are handled. Decisions regarding whether to include or exclude demographic variables, how to encode socioeconomic indicators, and how to balance

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

predictive accuracy with fairness remain inconsistent across studies. Such variability weakens the generalizability of findings and limits practical adoption in educational systems (Fazelpour & Danks, 2021; Pessach & Shmueli, 2022).

In response to these limitations, this study presents a PRISMA-compliant systematic review of fairness and bias detection in ML-based scholarship allocation systems. The study pursues three objectives: (i) to examine bias types, detection methods, and fairness metrics in scholarship allocation research; (ii) to develop an education-specific taxonomy integrating datasets, features, sensitive attributes, and ML lifecycle stages; and (iii) to identify methodological gaps, deployment risks, and future research directions for fair scholarship systems. Accordingly, the study addresses the following research questions: (i) What types of bias, detection methods, and fairness metrics are applied in ML-driven scholarship allocation systems? (ii) How are datasets, features, and sensitive attributes managed in these systems? (iii) What evaluation protocols, validation strategies, and deployment gaps exist in current implementations?

This study contributes in three key ways. First, it provides a systematic synthesis of fairness-related research in ML-driven scholarship allocation. Second, it introduces a structured, education-specific taxonomy linking bias sources, detection stages, and fairness metrics across the ML pipeline. Third, it offers practical recommendations for dataset handling, model evaluation, and deployment practices to support equitable and accountable scholarship decision-making systems.

METHODOLOGY

This study adopts a systematic literature review methodology to identify, evaluate, and synthesize existing research on bias detection in machine learning for educational decision-making systems. The review process was conducted in accordance with the PRISMA 2020 guidelines to ensure transparency, reproducibility, and methodological rigor.

A comprehensive and structured search was performed across four major academic databases: Scopus, Web of Science, IEEE Xplore, and Google Scholar. The search strategy combined keywords related to algorithmic bias, machine learning, and educational applications. Google Scholar was used only as a supplementary source. The validated search string used was: ("algorithmic bias" OR "fairness" OR "bias detection") AND ("machine learning" OR "artificial intelligence") AND ("education" OR "student selection" OR "scholarship allocation" OR "admission systems") AND (2020–2026). The search was limited to publications between January 2020 and March 2026 to capture recent advancements in fairness-aware machine learning and educational AI systems.

Studies were selected based on clearly defined inclusion and exclusion criteria to ensure methodological rigor and topical relevance. Eligible studies were limited to publications released between 2020 and 2026, reflecting recent developments in AI and ML applications. To qualify for inclusion, studies were required to employ ML or AI techniques and explicitly investigate issues related to bias detection, fairness assessment, or bias mitigation. Furthermore, the selected studies had to be situated within educational decision-making contexts, including scholarship allocation, student admission, academic performance prediction, and recommendation systems. Only peer-reviewed journal articles and conference proceedings were considered to maintain academic quality and reliability.

Conversely, studies were excluded if they addressed non-educational domains such as healthcare, finance, or autonomous systems, or if they applied ML in education without examining fairness or bias-related concerns. Research based solely on rule-based or traditional statistical approaches without machine learning components was also omitted. In addition, conceptual discussions, opinion papers, and studies without empirical implementation or validation were excluded, as were non–peer-reviewed sources, editorials, and technical reports deemed insufficiently rigorous for inclusion. This selection framework ensured that the final corpus remained empirically grounded, methodologically robust, and directly aligned with the study objectives.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

The study selection process followed a four-stage PRISMA workflow: identification, screening, eligibility, and inclusion. Initially, 820 records were identified, comprising 768 database records and 52 additional sources. After removing 180 duplicates, 640 unique records were retained for title and abstract screening. During the screening phase, 520 records were excluded for irrelevance, including studies outside the educational domain, those lacking fairness considerations, and those without machine learning applications. The remaining 120 studies proceeded to full-text review. Following full-text assessment, 70 studies were excluded for reasons such as absence of bias detection methods, lack of empirical validation, or insufficient methodological rigor. Ultimately, 50 studies met all inclusion criteria and were included in the final synthesis.

A structured data extraction protocol was implemented to ensure systematic and consistent collection of relevant information from all included studies. Main variables extracted comprised the study objectives and educational application domains, the types of machine learning models employed, bias detection techniques and fairness evaluation metrics, dataset characteristics, and the principal findings coupled with reported limitations. Following extraction, the data were analyzed using a thematic synthesis approach to identify recurring trends, methodological patterns, and emerging themes in bias detection and fairness assessment. The analysis further enabled the development of a structured taxonomy of bias detection approaches in educational machine learning systems, providing a coherent basis for comparing existing methods and identifying current research directions and gaps.

Although scholarship allocation is the main focus of this review, the scope was extended to encompass broader educational decision-making systems, given the scarcity of studies directly focusing on scholarship allocation. This expansion is methodologically justified, as these systems share key characteristics, including high-stakes decision-making, implications for resource allocation, and susceptibility to algorithmic bias.

To enhance methodological rigor and ensure the credibility of the review process, a predefined review protocol was established and systematically followed throughout the study selection and analysis stages. The inclusion and exclusion criteria were applied consistently across all retrieved studies to maintain objectivity and reduce selection bias, while ambiguous cases went through iterative evaluation to ensure accurate and justified inclusion decisions. Furthermore, only peer-reviewed journal articles and conference proceedings were considered to safeguard the quality and scholarly reliability of the evidence base. Collectively, these procedures strengthened the reliability, transparency, and reproducibility of the review.

A structured quality assessment framework was applied across four dimensions: methodological rigor, dataset transparency, reproducibility, and fairness evaluation validity. Methodological rigor was assessed based on the clarity of experimental design, the suitability of ML models, and the robustness of evaluation procedures. Dataset transparency was assessed based on whether datasets were publicly accessible, well-documented, and representative of the target populations. Reproducibility was evaluated based on the availability of code and data, as well as the level of methodological detail provided to support replication. Fairness evaluation validity deemed whether fairness metrics were appropriately selected and justified relative to the study context. Each study was scored using a standardized rubric to ensure comparability and to identify high-quality contributions, consistent with established practices in ML fairness and bias evaluation research (Alelyani, 2021; Huang et al., 2024).

The overall study selection process is summarized in Figure 1, which documents the records identified, screened, excluded, and ultimately included at each stage, thereby providing a transparent account of the study inclusion and exclusion decisions.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

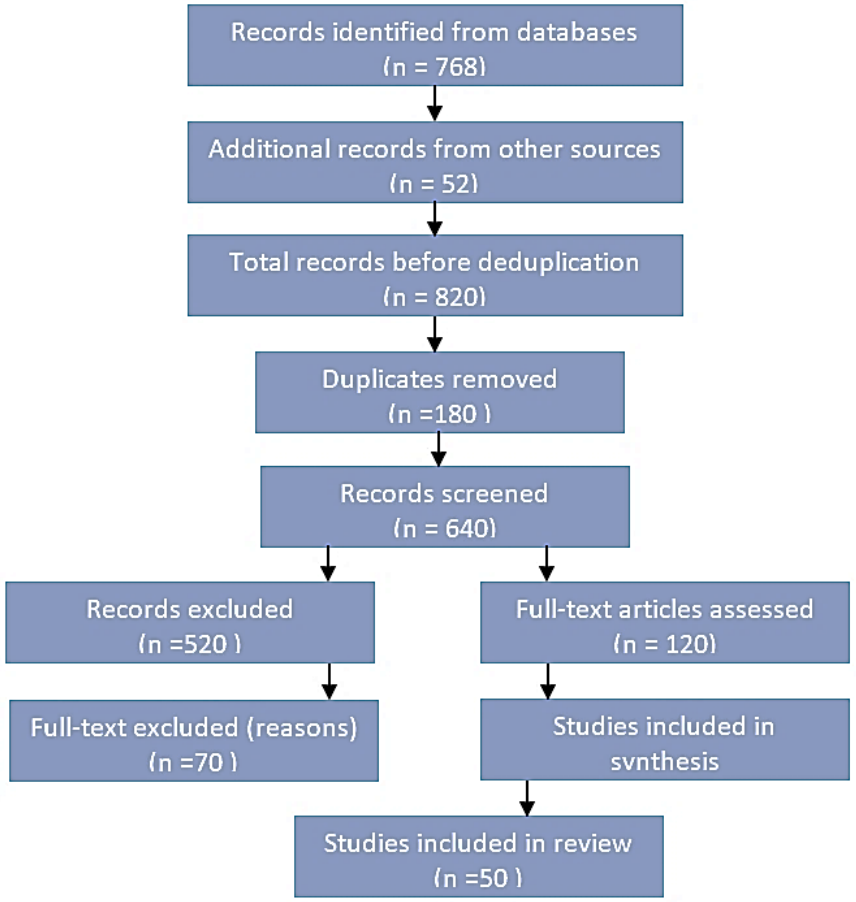

Figure 1: PRISMA-based study selection workflow for systematic review of bias detection in ML-driven scholarship allocation systems

RESULTS AND DISCUSSION

Research on AI fairness in education is extensive yet fragmented across technical, ethical, and pedagogical domains. While existing studies provide important insights into explainability, fairness definitions, and robustness, they rarely capture end-to-end decision pipelines in high-stakes applications such as scholarship allocation. Foundational work in fairness and explainable AI establishes key principles of interpretability, accountability, and transparency (Chorąś et al., 2020), but remains largely domain-agnostic and insufficient for capturing the operational complexity of educational allocation systems. Similarly, broader AI governance literature emphasizes robustness, privacy, and ethics but gives limited attention to distributive fairness in education-specific decision environments (Fazelpour & Danks, 2021; Siddique et al., 2024).

In educational AI, recent studies highlight persistent challenges in learning analytics, predictive modeling, and system integration, particularly in contexts constrained by limited data and infrastructure (Baker & Hawn, 2022; Bali et al., 2024). Emerging human-centered and context-aware approaches further emphasize interpretability and alignment with educational environments (Bali, 2026; Barnes & Hutson, 2024). However, fairness is still largely treated as a single-stage evaluation problem rather than a continuous process across data, model development, evaluation, and deployment. This fragmented treatment limits holistic understanding and weakens lifecycle accountability in educational decision systems.

Pipeline-oriented perspectives in broader AI domains demonstrate the importance of structured decision architectures where each stage is explicitly modeled and evaluated (Dritsas & Trigka, 2025; Lee et al., 2024). This motivates a similar lifecycle-based framing for scholarship allocation systems, where fairness must be embedded across interconnected operational stages rather than treated as an isolated evaluation criterion.

Scholarship allocation can therefore be conceptualized as a multi-stage machine learning decision pipeline involving screening, scoring, ranking, and allocation. Screening enforces eligibility rules, scoring estimates

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

candidate merit or need, ranking prioritizes applicants under scarcity constraints, and allocation distributes limited resources. Each stage introduces distinct bias risks: screening may exclude disadvantaged groups through rigid thresholds, scoring may amplify proxy indicators of privilege, ranking can magnify small score differences under competitive conditions, and allocation reflects institutional policy constraints. These interdependencies demonstrate that fairness must be evaluated across the entire pipeline rather than at isolated stages (Kanwal et al., 2024; Pham et al., 2025).

Bias in educational machine learning emerges through interactions among data, model design, and deployment contexts. Representation bias arises when certain groups are underrepresented in datasets, affecting downstream learning outcomes. Measurement bias occurs when proxy variables fail to accurately reflect constructs such as academic ability or financial need. Label bias reflects historical decision patterns embedded in training data, while historical bias originates from structural inequalities in society. Collectively, these biases interact and may compound unfairness when systems are deployed without corrective safeguards (Mehrabi et al., 2021; Boateng & Boateng, 2025).

Fairness in machine learning is commonly framed through group and individual perspectives. Group fairness evaluates parity across demographic groups, while individual fairness focuses on similar treatment of similar individuals. In scholarship allocation, these notions often conflict due to competing objectives of merit, need, and equity under constrained resources (Roshanaei, 2024). Consequently, fairness cannot be defined universally but must be interpreted within institutional and policy contexts.

Fairness metrics operationalize these concepts but introduce additional complexity. Statistical parity evaluates equality of selection rates across groups, equal opportunity ensures equal true positive rates among qualified individuals, equalized odds balances both false positives and true positives, and calibration ensures consistent probabilistic predictions across groups. However, these metrics are often mathematically incompatible, creating unavoidable trade-offs between fairness definitions and predictive performance (Corbett-Davies et al., 2023; Huang et al., 2024). In scholarship systems, metric selection is therefore a normative decision shaped by institutional priorities rather than purely technical optimization.

Bias detection methods rely on statistical, metric-based, and explainability-driven approaches. Statistical methods identify disparities across groups, while fairness metrics quantify inequities in model outcomes. Explainable AI enhances detection by exposing feature-level contributions to predictions, improving interpretability of bias mechanisms (Choraś et al., 2020; Bali, 2025). Bias mitigation strategies are typically categorized into pre-processing, in-processing, and post-processing techniques. Although mitigation dominates existing research, detection remains foundational because incorrect bias identification can lead to ineffective interventions. In scholarship allocation systems, detection also strengthens accountability and auditability (Ghani et al., 2023).

Fairness research in education has expanded significantly since 2020, reflecting broader adoption of ethical AI frameworks (Baker & Hawn, 2022). Early studies focused primarily on predictive accuracy, whereas recent work increasingly emphasizes transparency and fairness. However, research remains concentrated in computer science domains, with limited representation from developing regions despite their heightened relevance (Bali et al., 2024).

Across application contexts, three dominant systems emerge: scholarship allocation, admission prediction, and financial aid systems. Scholarship allocation emphasizes distributive fairness, admission systems prioritize ranking accuracy, and financial aid systems focus on socioeconomic need. These variations highlight the absence of a unified, scholarship-specific fairness framework capable of integrating diverse objectives (Idowu, 2024; Pham et al., 2025).

Representation bias remains the most frequently reported bias type, followed by measurement and label bias. These biases often interact, producing compounded algorithmic unfairness. Statistical disparity analysis, fairness metric evaluation, and explainability-based techniques are the most commonly used detection methods. However, statistical approaches lack causal depth, while explainability techniques remain largely post hoc and are not fully integrated into model design (Kesgin et al., 2025).

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

Traditional models such as logistic regression and decision trees remain widely used due to interpretability, while deep learning models offer improved predictive performance but reduce transparency, complicating fairness assessment. Hybrid approaches attempt to balance interpretability and accuracy (Du et al., 2020). Dataset design strongly influences fairness outcomes, as exclusion of sensitive attributes limits bias detection, while proxy variables introduce indirect discrimination. Limited dataset transparency further reduces reproducibility and weakens evaluation rigor (Alelyani, 2021; Boateng & Boateng, 2025).

Finally, evaluation practices remain inconsistent across studies. Cross-validation is widely used, but external validation is rare. Reporting of fairness–accuracy trade-offs is often incomplete, limiting comparability across studies. Reproducibility remains a major challenge due to missing datasets, incomplete methodological reporting, and limited implementation transparency (Huang et al., 2024).

Table 1. Comparative studies on fairness in ML-driven educational decision systems with emphasis on bias types, methodologies, and relevance to scholarship allocation

Author(s) / Year	Domain	ML Method	Bias / Fairness Focus	Methodological Approach	Main Contribution / Findings	Relevance to Scholarship Allocation Context
Fazelpour & Danks (2021)	ML fairness	Algorithmic bias	Bias sources & contextual fairness	Philosophical review	Highlights socio-technical nature of bias	Supports contextual scholarship fairness framing
Baker & Hawn (2022)	Education	Educational AI systems	Algorithmic bias in education	Narrative review	Shows how bias affects learning systems	Direct relevance to educational scholarship systems
Bali et al. (2024)	Education	AI in education	Educational bias trends	Empirical review	Identifies AI adoption trends in education	Contextual grounding for scholarship bias studies
Bali (2025)	Education	Explainable AI	Transparent assessment	Conceptual framework	Human-centered XAI for education systems	Supports interpretable scholarship allocation
Bali (2026)	Education	Context-aware ML	Personalization & equity bias	Framework design	Adaptive learning in developing contexts	Supports equitable scholarship personalization
Barnes & Hutson (2024)	Higher education	Ethical AI systems	Fairness & accountability	Policy review	AI ethics strategies in education	Governance framework for scholarship systems
Siddique et al. (2024)	ML systems	Bias mitigation survey	Dataset & model bias	Systematic review	Taxonomy of ML bias types and mitigation	Direct relevance to scholarship bias taxonomy
Alelyani (2021)	ML systems	Bias detection	Detection techniques	Empirical study	Evaluates ML bias detection methods	Supports scholarship bias auditing
Huang et al. (2024)	Biomedical ML	Fair ML techniques	Fairness evaluation	Scoping review	Reviews fairness metrics	Supports evaluation

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

					in real data	design in scholarship systems
Dritsas & Trigka (2025)	ML systems	Big data ML	System-level bias	Survey	ML and big data integration challenges	Analogous to large-scale scholarship datasets
Lee et al. (2024)	Education	LLM lifecycle	Lifecycle bias propagation	Framework study	Bias across model lifecycle stages	Strong relevance to scholarship pipeline bias
Pham et al. (2025)	Education ML	Fair ML software	Educational fairness design	Mapping study	Fairness in educational ML software	Direct scholarship system relevance
Van Giffen et al. (2022)	ML systems	Bias classification	Pipeline bias types	Review	Categorizes ML bias and mitigation methods	Supports structured scholarship taxonomy
Corbett-Davies et al. (2023)	ML fairness	Fairness metrics	Metric limitations	Analytical review	Shows incompatibility of fairness metrics	Critical for scholarship fairness trade-offs
Oneto & Chiappa (2020)	ML fairness	Fair ML theory	Individual fairness	Tutorial review	Explains individual fairness concepts	Supports student-level fairness assessment
Choraś et al. (2020)	ML systems	Explainable AI	Transparency & fairness	Conceptual study	Links XAI with fairness and security	Supports explainable scholarship decisions
Ghani et al. (2023)	ML systems	Fair ML practice	Bias detection & mitigation	Practical guide	Hands-on fairness implementation methods	Supports real-world scholarship deployment
Li et al. (2023)	Education	Predictive analytics	Student bias prediction	Systematic review	Predictive bias in education systems	Direct scholarship prediction relevance
Gándara et al. (2024)	Education ML	Student success prediction	Racial bias	Empirical study	Bias in student success models	Scholarship allocation fairness relevance
Verger et al. (2024)	Education ML	Fair ML metrics	Evaluation bias	Empirical study	Fairness metrics in education data mining	Supports evaluation framework design
Idowu (2024)	Education ML	Debiasing models	Algorithmic bias in education	Analytical study	Education-specific bias mitigation	Direct scholarship fairness relevance
Roshanaei (2024)	Higher education	AI fairness governance	Diversity & inclusion bias	Policy analysis	AI fairness in higher	Scholarship policy

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

					education governance	alignment
Raftopoulos et al. (2024)	Education ML	Admission prediction	Fairness in ranking	Empirical study	Fair student admission prediction	Closely aligned with scholarship ranking
Pham et al. (2025)	Education ML	Regression-based fairness	Fair prediction models	Experimental study	Enhances fairness in educational ML models	Scholarship fairness optimization
Mehrabi et al. (2021)	ML systems	Bias taxonomy	Dataset & model bias	Survey	Comprehensive ML fairness survey	Foundational for scholarship bias taxonomy
Boateng & Boateng (2025)	Education AI	Educational decision systems	Structural bias	Analytical review	AI-driven educational inequality impacts	Strong scholarship allocation relevance
Mavrogiorgos et al. (2024)	ML systems	General ML bias	Dataset bias	Literature review	Overview of ML bias types	Supports dataset bias classification
Fazelpour et al. (2022)	ML fairness	Algorithmic justice	Contextual fairness	Philosophical analysis	Situated fairness dynamics	Supports policy-driven scholarship fairness
Pagano et al. (2023)	ML systems	Fair ML tools	Bias detection tools	Systematic review	Tools and metrics for fairness evaluation	Supports operational scholarship auditing

Table 1 presents a synthesized overview of prior research on algorithmic fairness in machine learning and education, highlighting variant bias, methodological approaches, and implications for scholarship allocation systems.

TAXONOMY OF BIAS DETECTION IN SCHOLARSHIP ALLOCATION SYSTEMS

The proposed taxonomy adopts a lifecycle-oriented perspective on machine learning–driven scholarship allocation systems, framing bias as a dynamic phenomenon that may emerge and intensify across different stages of algorithmic decision-making. Rather than treating unfairness as an isolated modeling problem, this framework recognizes bias as a systemic condition distributed across data preparation, feature construction, model development, prediction generation, and deployment environments. Fairness scholarship increasingly shows that inequitable outcomes rarely stem from a single technical component but instead arise through interactions among datasets, optimization processes, institutional assumptions, and operational contexts (Barocas et al., 2019; Fazelpour & Danks, 2021; Mehrabi et al., 2021). This perspective is particularly relevant in educational AI, where algorithmic systems influence access to scholarships, academic opportunities, and resource distribution (Baker & Hawn, 2022; Li et al., 2023; Chinta et al., 2024).

The taxonomy is guided by three principles: interpretability, extensibility, and contextual applicability. Interpretability ensures conceptual clarity among different forms of bias and fairness assessment. Extensibility accommodates evolving detection and mitigation techniques, particularly as generative and adaptive AI systems become more prevalent in education (Afreen et al., 2025; Lee et al., 2024). Contextual applicability ensures relevance to operational scholarship systems where institutional objectives, ethical expectations, and policy considerations intersect (Alvarez et al., 2024; Barnes & Hutson, 2024). These principles align with responsible AI frameworks emphasizing transparency, accountability, and socially responsive algorithmic design (Bogina et al., 2022; Chan, 2023; Williamson et al., 2023).

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

The taxonomy comprises five interconnected dimensions. *Bias source* identifies where unfairness originates within the machine learning pipeline. Data-level bias occurs when imbalanced sampling, incomplete records, and demographic underrepresentation distort the representativeness of the population (Mehrabi et al., 2021; Siddique et al., 2024). Educational datasets are particularly susceptible because historical academic and participation records often reflect broader structural inequalities (Baker & Hawn, 2022; Boateng & Boateng, 2025). Feature-level bias arises when seemingly neutral variables act as proxies for sensitive attributes such as socioeconomic status or geographic location (Fazelpour & Danks, 2021; Van Giffen et al., 2022). Label bias occurs when historical decisions or institutional evaluations are treated as objective ground truth despite containing embedded inequities (Pessach & Shmueli, 2022; Garcia et al., 2024). Model bias reflects inequities introduced through algorithmic design and optimization processes that prioritize predictive performance over distributive fairness (Du et al., 2020; Jui & Rivas, 2024). Deployment bias develops after implementation as institutional feedback, behavioral adaptation, or environmental changes reshape model outcomes over time (Pagano et al., 2023; Lee et al., 2024).

The detection stage situates fairness assessment across the operational workflow. Pre-model detection focuses on dataset auditing, representational analysis, and preprocessing diagnostics before training (Alelyani, 2021; Pagano et al., 2023). Such interventions are essential because biased training data can disproportionately disadvantage marginalized student groups (Idowu, 2024; Omughelli et al., 2024). In-model detection integrates fairness objectives and constrained optimization directly into learning procedures (Friedler et al., 2019; Tizpaz-Niari et al., 2022). Post-model detection evaluates trained systems through disparity analysis, subgroup performance evaluation, and fairness auditing (Jones et al., 2020; Huang et al., 2024). Together, these stages support continuous fairness monitoring rather than isolated validation.

Method type categorizes bias detection techniques according to analytical strategy. Statistical approaches remain dominant, relying on group comparisons and disparity measures to identify unequal outcomes (Corbett-Davies et al., 2023; Pessach & Shmueli, 2022). Their computational efficiency and interpretability make them attractive for institutional auditing and policy reporting (Pagano et al., 2023). Causal approaches examine whether outcomes remain stable under hypothetical interventions or alternative conditions, allowing deeper investigation of discrimination mechanisms beyond observed correlations (Fazelpour et al., 2022; Kim & Kim, 2025). However, these methods require stronger assumptions and contextual knowledge, limiting widespread adoption in educational settings (Jui & Rivas, 2024). Explainability-based methods complement both approaches by revealing how model inputs influence predictions through feature attribution and interpretable reasoning techniques (Choraś et al., 2020; Kesgin et al., 2025). Such transparency is increasingly important because educational decision systems must maintain stakeholder trust alongside predictive performance (Raftopoulos et al., 2024).

Fairness metric family addresses competing normative definitions of fairness. Group fairness evaluates whether outcomes are equitably distributed across demographic groups and remains central to many fairness frameworks (Corbett-Davies et al., 2023). Individual fairness emphasizes consistent treatment among similarly situated students (Oneto & Chiappa, 2020; Pessach & Shmueli, 2022). Counterfactual fairness extends these approaches by assessing whether predictions remain unchanged under hypothetical changes to sensitive attributes (Kim & Kim, 2025). Because these fairness definitions may conflict mathematically and ethically, selecting appropriate metrics depends on institutional priorities and policy obligations rather than universal standards (Wachter et al., 2020; Corbett-Davies et al., 2023).

Educational context represents the fifth dimension and situates fairness within scholarship allocation objectives. Merit-based systems prioritize academic achievement and may generate tensions between efficiency and equitable access (Kanwal et al., 2024; Raftopoulos et al., 2024). Need-based models emphasize financial vulnerability and socioeconomic disadvantage, raising distinct distributive concerns (Bhat, 2024; Idowu, 2024). Hybrid systems combine merit and need, introducing more complex trade-offs among excellence, inclusion, and institutional priorities (Huang, 2022; Pham et al., 2025). Educational context, therefore, shapes both fairness interpretation and bias evaluation.

The taxonomy may be represented as a layered matrix aligned with the machine learning lifecycle, where pipeline stages form the horizontal dimension and bias sources define the vertical structure. Detection stages

function as embedded layers, while method types and fairness metrics operate as analytical overlays. Educational context spans the framework as a cross-cutting dimension. This structure enables systematic examination of bias propagation while preserving interpretability and practical relevance (Lee et al., 2024; Pagano et al., 2023). Mapping prior studies onto this taxonomy reveals important patterns. Existing literature places strong emphasis on pre-model auditing and post-model statistical evaluation, particularly in educational prediction and admission systems (Li et al., 2023; Gándara et al., 2024; Verger et al., 2024). In contrast, fairness-aware optimization and deployment-stage monitoring remain comparatively limited despite recognition that bias evolves after implementation (Raftopoulos et al., 2025; Zhang et al., 2026). Consequently, fairness research still emphasizes disparity identification more than sustained lifecycle governance.

Several insights emerge. Statistical methods dominate current scholarship, whereas causal and explainability-driven approaches remain less developed despite their stronger diagnostic potential (Pessach & Shmueli, 2022; Kesgin et al., 2025). Fairness monitoring at the deployment stage remains underexplored, despite the continuously evolving nature of educational environments (Lee et al., 2024). Moreover, most studies rely on static and single-attribute fairness evaluations, overlooking intersectional and longitudinal inequities (Madaio et al., 2022; Williamson et al., 2023). A persistent divide, therefore, remains between fairness theory and deployable scholarship systems, highlighting the need for integrated, lifecycle-aware, and education-specific fairness frameworks to support both predictive performance and distributive justice (Baker & Hawn, 2022; Chinta et al., 2024; Idowu, 2024).

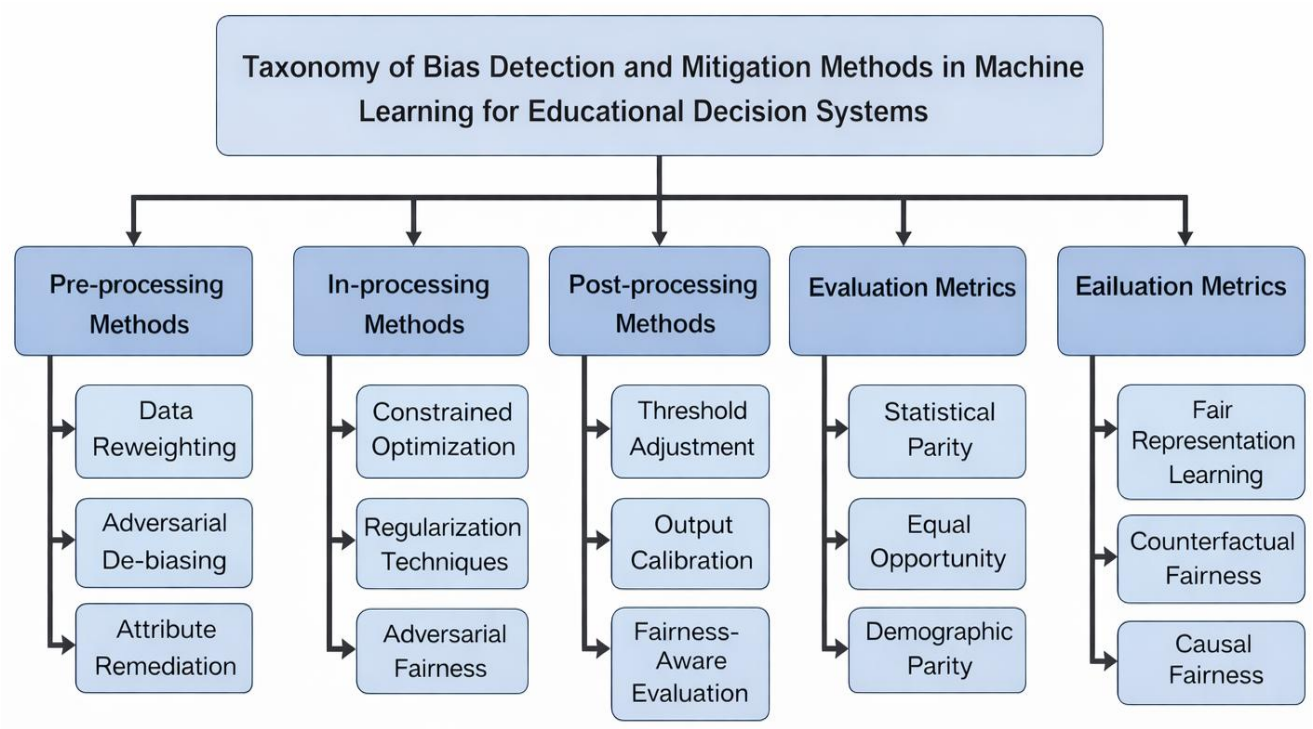

Figure 2: A Taxonomy of Bias Detection and Fairness in ML–Driven Scholarship Allocation Systems

A structured framework summarizing the key components of fairness in ML-based scholarship allocation, organized into five layers: bias sources, detection stages, methodological approaches, fairness metric families, and educational application contexts. The taxonomy provides a clear, high-level synthesis of how bias is introduced, detected, mitigated, and evaluated to support equitable and transparent scholarship decision-making.

Empirical Mapping and Validation of the Taxonomy

The proposed taxonomy is further validated through empirical alignment with findings from existing educational machine learning studies and scholarship allocation models. For instance, datasets used in student success prediction and scholarship eligibility systems (Gándara et al., 2024; Raftopoulos et al., 2024) demonstrate clear representation and measurement bias patterns, which map directly to the data-level and

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

feature-level bias dimensions of the taxonomy. Similarly, regression-based fairness models in educational systems (Pham et al., 2025) correspond to the in-model detection stage, where fairness constraints are integrated during optimization. Explainability-based approaches (Choraś et al., 2020; Kesgin et al., 2025) align with post-model detection by enabling feature-level attribution of bias sources. This mapping confirms the practical relevance and applicability of the proposed taxonomy across real-world educational datasets and decision systems.

FINDINGS

The synthesis of systematically reviewed studies reveals a persistent misalignment between current ML fairness research and the operational realities of scholarship allocation systems. Most literature treats fairness as a static, model-level evaluation problem focused on post-hoc disparity measurement, rather than a dynamic governance issue spanning the full decision pipeline. This creates a conceptual divide between controlled experimental fairness settings and real-world educational systems, where decisions are embedded in institutional, socioeconomic, and policy-driven environments. Scholarship allocation instead functions as a lifecycle-driven socio-technical process in which bias emerges, evolves, and propagates across interconnected stages, consistent with the taxonomy of bias sources, detection stages, methods, metrics, and educational context (Barocas et al., 2019; Fazelpour & Danks, 2021; Mehrabi et al., 2021).

Across the reviewed literature, bias is primarily concentrated at the data and representation level, with limited attention to downstream propagation effects. Representation-related distortions dominate due to imbalanced and historically skewed educational datasets that fail to reflect equitable populations (Siddique et al., 2024; Mavrogiorgos et al., 2024). In scholarship contexts, such distortions are amplified by entrenched socioeconomic inequalities reflected in historical award distributions (Baker & Hawn, 2022). Measurement and proxy effects further contribute to inequity by encoding socioeconomic status indirectly through variables such as school ranking or geographic indicators (Fazelpour & Danks, 2021; Van Giffen et al., 2022). Additionally, historical labeling practices embedded in prior decisions introduce inherited bias patterns (Pessach & Shmueli, 2022). Although deployment-stage bias is less frequently studied, emerging evidence indicates that feedback loops in educational systems may reinforce and amplify inequality over time (Lee et al., 2024).

Regarding bias detection, the literature is heavily skewed toward pre-model and post-model auditing, while in-model approaches remain underdeveloped. Pre-model methods focus on dataset inspection, imbalance quantification, and representational analysis prior to training (Alelyani, 2021; Pagano et al., 2023). Post-model techniques dominate scholarship allocation studies, relying on fairness metrics and subgroup performance evaluation after model training (Jones et al., 2020; Ghani et al., 2023). In contrast, in-model detection, where fairness constraints are incorporated directly into optimization, remains limited despite its potential to reduce structural bias earlier in the pipeline (Tizpaz-Niari et al., 2022). This imbalance indicates that fairness is still largely treated as an evaluation layer rather than an integrated design principle in educational decision systems.

Methodologically, statistical approaches remain the most widely adopted due to their simplicity and interpretability, typically relying on group disparity comparisons across sensitive attributes (Corbett-Davies et al., 2023). However, these methods provide limited insight into causal mechanisms of discrimination. Causal and counterfactual approaches offer stronger theoretical grounding for identifying discrimination pathways but remain underutilized due to strong assumptions and data requirements (Fazelpour et al., 2022). Explainability-based techniques are increasingly applied in educational contexts to enhance interpretability of model decisions (Choraś et al., 2020; Kesgin et al., 2025), yet most remain post hoc rather than embedded within model design, limiting their operational impact in scholarship allocation systems.

In fairness evaluation, group-based metrics such as demographic parity, equal opportunity, and equalized odds dominate the literature, while individual and counterfactual fairness are less frequently applied despite their relevance to student-level decision equity (Oneto & Chiappa, 2020). A key limitation is the known incompatibility among fairness definitions, which leads to unavoidable trade-offs between competing fairness objectives and predictive accuracy (Wachter et al., 2020). In scholarship allocation, these trade-offs are further complicated by institutional constraints, where merit-based, need-based, and hybrid allocation policies impose

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

differing normative priorities. Consequently, fairness is better understood as a context-dependent policy decision rather than a purely technical optimization objective.

The educational context significantly influences how fairness is defined and applied in scholarship allocation systems. Merit-based approaches emphasize academic performance, need-based systems prioritize socioeconomic disadvantage, while hybrid models attempt to balance both goals, creating inherent tensions between equity, efficiency, and institutional policy objectives (Kanwal et al., 2024; Idowu, 2024). However, most existing fairness frameworks are domain-agnostic and do not adequately incorporate the specific constraints of educational allocation, limiting their practical applicability.

A major gap in current research is the lack of integrated bias modeling across the full machine learning lifecycle. Studies typically examine isolated stages such as data preprocessing, model training, or evaluation, without accounting for how bias propagates across these interconnected phases. This fragmented approach fails to capture feedback loops in scholarship systems, where decisions shape future data distributions and can reinforce inequality over time. This highlights the need for lifecycle-aware fairness frameworks that unify bias sources, detection, and mitigation strategies (Lee et al., 2024; Fazelpour & Danks, 2021).

Overall, fairness research in scholarship allocation remains largely focused on representation bias and post-hoc evaluation, with limited attention to causal mechanisms, deployment contexts, and feedback dynamics. Explainability and causal approaches are underused, and evaluation practices remain inconsistent and weakly aligned with educational policy goals. Collectively, the evidence suggests that fairness cannot be achieved through isolated interventions but requires an integrated, lifecycle-oriented governance framework that aligns technical fairness methods with educational realities, or risk reinforcing structural inequality under the guise of objectivity.

Implementation, Computational, and Institutional Challenges

Despite progress in fairness-aware machine learning, deployment in scholarship allocation remains constrained by practical barriers. Computational demands of causal and multi-objective fairness models limit adoption in low-resource institutions. Many organizations also lack infrastructure for continuous bias monitoring and lifecycle auditing. Institutional challenges such as weak regulatory frameworks, inconsistent policies, and limited transparency standards further hinder implementation. In addition, stakeholder trust is a major obstacle, as educators and students may resist algorithmic decisions without clear explainability and accountability. Overall, effective deployment requires alignment between technical design, institutional capacity, governance structures, and ethical transparency requirements.

Table 2: Comparative Analysis of Fairness Techniques in ML-based Educational Decision Systems

Method Type	Strengths	Limitations	Computational Cost	Educational Applicability
Statistical fairness methods	Simple to implement; widely used for disparity measurement; strong interpretability for institutional auditing and reporting	Metric incompatibility across definitions (e.g., demographic parity vs equalized odds); limited causal insight into discrimination mechanisms	Low	Highly applicable for scholarship auditing and policy reporting, but insufficient for full fairness assurance in complex decision pipelines (Corbett-Davies et al., 2023; Pessach & Shmueli, 2022; Huang et al., 2024)
Causal fairness methods	Capture underlying discrimination mechanisms beyond correlations; enable	Require strong structural assumptions and high-quality causal data; difficult to operationalize	High	Limited but growing applicability in scholarship allocation due to data constraints and

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

	counterfactual reasoning and deeper fairness diagnosis	in real educational datasets		institutional complexity (Fazelpour et al., 2022; Kim & Kim, 2025; Jui & Rivas, 2024)
Explainability-based methods	Improve transparency of model decisions; support feature-level attribution; enhance trust and interpretability in high-stakes educational decisions	Mostly post-hoc; do not inherently guarantee fairness; limited integration into model training pipelines	Medium	Strong applicability in educational scholarship systems for accountability and stakeholder trust (Choraś et al., 2020; Kesgin et al., 2025; Raftopoulos et al., 2024)
Pre-processing methods	Address bias at dataset level before training; include rebalancing, sampling correction, and dataset auditing	May not fully eliminate downstream bias; risk of information loss or distortion of real-world distributions	Low–Medium	Highly applicable in resource-constrained educational datasets where bias originates from historical records (Alelyani, 2021; Pagano et al., 2023; Idowu, 2024; Omughelli et al., 2024)
In-processing methods	Integrate fairness constraints directly into model training; allow joint optimization of accuracy and fairness objectives	Increased algorithmic complexity; requires careful tuning of fairness–accuracy trade-offs; harder to implement in standard ML pipelines	High	Moderately applicable in advanced scholarship systems with sufficient computational and technical capacity (Friedler et al., 2019; Du et al., 2020; Tizpaz-Niari et al., 2022)
Post-processing methods	Adjust model outputs after training to improve fairness; flexible and model-agnostic; easier to deploy without retraining	Does not correct root causes of bias; may lead to instability or reduced model calibration	Medium	Commonly used in educational decision systems for quick fairness adjustments in scholarship allocation and ranking systems (Jones et al., 2020; Ghani et al., 2023; Pagano et al., 2023)

CONCLUSION

This study presents a comprehensive synthesis of bias detection methods in machine learning-based scholarship allocation systems and introduces a structured taxonomy spanning bias sources, detection stages, methodological approaches, fairness metrics, and educational contexts. The findings demonstrate that current research remains dominated by statistical, model-centric approaches that insufficiently address the systemic and causal nature of bias in educational decision pipelines. The proposed taxonomy provides a unified conceptual framework that integrates data-level, model-level, and deployment-level considerations, enabling systematic comparison across fragmented studies. By situating fairness within the full lifecycle of scholarship allocation systems, the framework advances both theoretical clarity and practical relevance. Ultimately, achieving fairness in scholarship allocation requires a shift from isolated algorithmic fixes toward holistic, governance-aware, and context-sensitive AI systems. As educational institutions increasingly rely on automated decision-making, ensuring transparency, accountability, and equity becomes not only a technical requirement but a foundational ethical imperative for sustainable and just educational development.

REFERENCES

1. Afreen, J., Mohaghegh, M., & Doborjeh, M. (2025). Systematic literature review on bias mitigation in generative AI. *AI and Ethics*, *5*(5), 4789-4841.
2. Albaroudi, E., Mansouri, T., & Alameer, A. (2024). A comprehensive review of AI techniques for addressing algorithmic bias in job hiring. *Ai*, *5*(1), 383-404.
3. Alelyani, S. (2021). Detection and evaluation of machine learning bias. *Applied Sciences*, *11*(14), 6271.
4. Alvarez, J. M., Colmenarejo, A. B., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., ... & Ruggieri, S. (2024). Policy advice and best practices on bias and fairness in AI. *Ethics and Information Technology*, *26*(2), 31.
5. Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International journal of artificial intelligence in education*, *32*(4), 1052-1092.
6. Bali, B., Garba, E. J., Ahmadu, A. S., Takwate, K. T., & Malgwi, Y. M. (2024). Analysis of emerging trends in artificial intelligence for education in Nigeria. *Discover Artificial Intelligence*, *4*(1), 110. https://doi.org/10.1007/s44163-024-00163-y
7. Bali, B. (2025). *Cognitively-informed explainable AI for education: Enhancing student engagement and transparent assessment through human-centered modeling. Journal of Social and Scientific Education, 2*(3), 173–190. https://doi.org/10.58230/josse.v2i3.351
8. Bali, B. (2026). *A novel context-aware intelligent learning framework for personalized education in developing countries. Next Research, 8*, 101570. https://doi.org/10.1016/j.nexres.2026.101570
9. Barnes, E., & Hutson, J. (2024). Navigating the ethical terrain of AI in higher education: Strategies for mitigating bias and promoting fairness. In *Forum for Education Studies* (Vol. 2, No. 2).
10. Bhat, J. (2024). Responsible Machine Learning in Student-Facing Applications: Bias Mitigation & Fairness Frameworks. *American International Journal of Computer Science and Technology*, *6*(1), 38-49.
11. Boateng, O., & Boateng, B. (2025). Algorithmic bias in educational systems: Examining the impact of AI-driven decision making in modern education. *World Journal of Advanced Research and Reviews*, *25*(1), 2012-2017.
12. Bogina, V., Hartman, A., Kuflik, T., & Shulner-Tal, A. (2022). Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics. *International Journal of Artificial Intelligence in Education*, *32*(3), 808-833.
13. Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International journal of educational technology in higher education*, *20*(1), 38.
14. Chinta, S. V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Quy, T. L., & Zhang, W. (2024). FairAIED: Navigating fairness, bias, and ethics in educational AI applications. *arXiv preprint arXiv:2407.18745*.
15. Choraś, M., Pawlicki, M., Puchalski, D., & Kozik, R. (2020, June). Machine learning–the results are not the only thing that matters! what about security, explainability and fairness?. In *International Conference on Computational Science* (pp. 615-628). Cham: Springer International Publishing.
16. Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. *Journal of Machine Learning Research*, *24*(312), 1-117.
17. Dritsas, E., & Trigka, M. (2025). Exploring the intersection of machine learning and big data: A survey. *Machine Learning and Knowledge Extraction*, *7*(1), 13.
18. Du, M., Yang, F., Zou, N., & Hu, X. (2020). Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, *36*(4), 25-34.
19. Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, *16*(8), e12760.
20. Fazelpour, S., Lipton, Z. C., & Danks, D. (2022). Algorithmic fairness and the situated dynamics of justice. *Canadian Journal of Philosophy*, *52*(1), 44-60.
21. Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019, January). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 329-338).

22. Gándara, D., Anahideh, H., Ison, M. P., & Picchiarini, L. (2024). Inside the black box: Detecting and mitigating algorithmic bias across racialized groups in college student-success prediction. *AERA open*, *10*, 23328584241258741.
23. Garcia, A. C. B., Garcia, M. G. P., & Rigobon, R. (2024). Algorithmic discrimination in the credit domain: what do we know about it?. *AI & society*, *39*(4), 2059-2098.
24. Ghani, R., Rodolfa, K. T., Saleiro, P., & Jesus, S. (2023, August). Addressing bias and fairness in machine learning: A practical guide and hands-on tutorial. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 5779-5780).
25. Huang, Y., Guo, J., Chen, W. H., Lin, H. Y., Tang, H., Wang, F., ... & Bian, J. (2024). A scoping review of fair machine learning techniques when using real-world data. *Journal of biomedical informatics*, *151*, 104622.
26. Huang, F. (2022). FAI: Toward Fair Decision Making and Resource Allocation with Application to AI-Assisted Graduate Admission and Degree Completion. *NSF Award Number 2147276. Directorate for Computer and Information Science and Engineering*, *21*(2147276), 47276.
27. Idowu, J.A. Debiasing Education Algorithms. *Int J Artif Intell Educ* 34, 1510–1540 (2024). https://doi.org/10.1007/s40593-023-00389-4
28. Jones, G. P., Hickey, J. M., Di Stefano, P. G., Dhanjal, C., Stoddart, L. C., & Vasileiou, V. (2020). Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. *arXiv preprint arXiv:2010.03986*.
29. Jui, T. D., & Rivas, P. (2024). Fairness issues, current approaches, and challenges in machine learning models. *International Journal of Machine Learning and Cybernetics*, *15*(8), 3095-3125.
30. Kanwal, B., Shoukat, R. S., Rehman, S. U., Kundi, M., AlSaedi, T., & Alahmadi, A. (2024). A New Framework for Scholarship Predictor Using a Machine Learning Approach. *Intelligent Automation & Soft Computing*, *39*(5).
31. Kesgin, K., Kiraz, S., Kosunalp, S., & Stoycheva, B. (2025). Beyond performance: Explaining and ensuring fairness in student academic performance prediction with machine learning. *Applied Sciences*, *15*(15), 8409.
32. Kim, W., & Kim, H. (2025, June). Counterfactual fairness evaluation of machine learning models on educational datasets. In *International Conference on Intelligent Tutoring Systems* (pp. 88-103). Cham: Springer Nature Switzerland.
33. Lee, J., Hicke, Y., Yu, R., Brooks, C., & Kizilcec, R. F. (2024). The life cycle of large language models in education: A framework for understanding sources of bias. *British Journal of Educational Technology*, *55*(5), 1982-2002.
34. Li, L., Sha, L., Li, Y., Raković, M., Rong, J., Joksimovic, S., ... & Chen, G. (2023, March). Moral machines or tyranny of the majority? A systematic review on predictive bias in education. In *LAK23: 13th international learning analytics and knowledge conference* (pp. 499-508).
35. Madaio, M., Blodgett, S. L., Mayfield, E., & Dixon-Román, E. (2022). Beyond "fairness": Structural (in) justice lenses on AI for education. In *The ethics of artificial intelligence in education* (pp. 203-239). Routledge.
36. Mavrogiorgos, K., Kiourtis, A., Mavrogiorgou, A., Menychtas, A., & Kyriazis, D. (2024). Bias in machine learning: A literature review. *Applied Sciences*, *14*(19), 8860.
37. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, *54*(6), 1-35.
38. Omughelli, D., Gordon, N., & Al Jaber, T. (2024). Fairness, bias, and ethics in AI: Exploring the factors affecting student performance. *Journal of Intelligent Communication*, *3*(2), 100-110.
39. Oneto, L., & Chiappa, S. (2020, April). Fairness in machine learning. In *Recent trends in learning from data: Tutorials from the inns big data and deep learning conference (innsbddl2019)* (pp. 155-196). Cham: Springer International Publishing.
40. Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., ... & Nascimento, E. G. (2023). Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, *7*(1), 15.
41. Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM computing surveys (CSUR)*, *55*(3), 1-44.

42. Pham, N., Do, M. K., Dai, T. V., Hung, P. N., & Nguyen-Duc, A. (2025). FAIREDU: A multiple regression-based method for enhancing fairness in machine learning models for educational applications. *Expert Systems with Applications*, *269*, 126219.
43. Pham, N., Ngoc, H. P., & Nguyen-Duc, A. (2025). Fairness for machine learning software in education: A systematic mapping study. *Journal of Systems and Software*, *219*, 112244.
44. Raftopoulos, G., Davrazos, G., & Kotsiantis, S. (2024). Fair and transparent student admission prediction using machine learning models. *Algorithms*, *17*(12), 572.
45. Raftopoulos, G., Davrazos, G., & Kotsiantis, S. (2025). Evaluating fairness strategies in educational data mining: A comparative study of bias mitigation techniques. *Electronics*, *14*(9), 1856.
46. Roshanaei, M. (2024). Towards best practices for mitigating artificial intelligence implicit bias in shaping diversity, inclusion and equity in higher education. *Education and Information Technologies*, *29*(14), 18959-18984.
47. Siddique, S., Haque, M. A., George, R., Gupta, K. D., Gupta, D., & Faruk, M. J. H. (2024). Survey on machine learning biases and mitigation techniques. *Digital*, *4*(1), 1-68.
48. Tizpaz-Niari, S., Kumar, A., Tan, G., & Trivedi, A. (2022, May). Fairness-aware configuration of machine learning libraries. In *Proceedings of the 44th International Conference on Software Engineering* (pp. 909-920).
49. Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*, 93-106.
50. Venkatasubbu, S., & Krishnamoorthy, G. (2022). Ethical considerations in AI addressing bias and fairness in machine learning models. *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online)*, *1*(1), 130-138.
51. Verger, M., Fan, C., Lallé, S., Bouchet, F., & Luengo, V. (2024). A comprehensive study on evaluating and mitigating algorithmic unfairness with the MADD metric. *Journal of Educational Data Mining*, *16*(1), 365-409.
52. Wachter, S., Mittelstadt, B., & Russell, C. (2020). Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *W. Va. L. Rev.*, *123*, 735.
53. Wei, X., Kumar, N., & Zhang, H. (2025). Addressing bias in generative AI: Challenges and research opportunities in information management. *Information & Management*, *62*(2), 104103.
54. Williamson, B., Eynon, R., Knox, J., & Davies, H. (2023). Critical perspectives on AI in education: Political economy, discrimination, commercialization, governance and ethics. In *Handbook of artificial intelligence in education* (pp. 553-570). Edward Elgar Publishing.
55. Yan, Y., & Liu, H. (2025). Ethical framework for AI education based on large language models. *Education and Information Technologies*, *30*(8), 10891-10909.
56. Zhang, F., Xing, W., Li, C., & Jiang, Y. (2026). Fair AI in educational predictions: A multi-group fairness approach using reinforcement learning. *The Internet and Higher Education*, 101074.