International Journal Software Engineering and Computer Science (IJSECS)
6 (1), 2026, 129-139
Published Online April 2026 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN: 2776-4869, E-ISSN: 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v6i1.6746.

RESEARCH ARTICLE Open Access

Analysis of Gender Inequality in Artificial Intelligence-Based Recruitment Systems: A Systematic Literature Review (SLR)

Herdaning Sandra Kumalasari [1*], **Magdalena A. Ineke Pakereng** [2]

[1*,2] Informatics Engineering Study Program, Faculty of Information Technology, Universitas Kristen Satya Wacana, Salatiga City, Central Java Province, Indonesia.

**Corresponding author*: 672022077@student.uksw.edu.

Received: March 5, 2026; Accepted: April 1, 2026; Published: April 10, 2026.

Abstract: The increasing adoption of Artificial Intelligence (AI) in recruitment has raised concerns about algorithmic discrimination that may disadvantage certain groups, particularly women. This study analyzed gender inequality in AI-based recruitment systems by synthesizing evidence from both technical and ethical perspectives. A Systematic Literature Review (SLR) was conducted on studies published between 2020 and 2025, applying predefined inclusion and exclusion criteria, followed by screening, quality assessment, and thematic synthesis. The review retained 10 studies (n = 10) that met the eligibility and quality threshold. Historically imbalanced training data emerged as the most frequently reported driver of gender bias, often producing unfair screening, ranking, and selection outcomes. Fairness conclusions were found to depend strongly on how recruitment outcomes were defined and measured, and prior studies consistently called for multiple fairness metrics supported by auditing practices. The literature also identified mitigation strategies spanning data balancing, fairness-aware model evaluation, transparency and audit mechanisms, and human oversight in decision-making. Gender bias in AI-based recruitment is, at its core, a socio-technical problem that requires combined interventions across data governance, model evaluation, and organizational accountability, while research gaps remain for future empirical validation and responsible AI deployment.

Keywords: Artificial Intelligence; Recruitment; Gender Bias; Systematic Literature Review; Algorithmic Fairness.

1. Introduction

Mehrabi *et al*. (2021) established that bias and fairness problems in machine learning arise when models absorb discriminatory patterns from historical data and imperfect labels. In employment contexts, Raghavan *et al*. (2020), Cowgill (2019), and Köchling and Wehner (2020) showed that algorithmic hiring tools can produce disparate impacts when fairness is not explicitly evaluated. Mori *et al*. (2024) argued that AI use in recruiting carries ethical requirements — transparency, accountability, and auditability — precisely because hiring decisions are socio-technical rather than purely algorithmic, while Mujtaba and Mahapatra (2024) mapped how those requirements connect to specific fairness metrics and mitigation approaches in AI-driven recruitment.

Mori *et al*. (2024) and Kassir *et al*. (2023) reported that AI is increasingly deployed in recruitment for résumé screening, candidate ranking, and early-stage assessment, partly to manage the efficiency demands of processing large applicant pools. This trajectory is also visible in practitioner and institutional discussions of

AI adoption in talent management and recruitment workflows (Deloitte, 2023; Kristian, 2024; Montesa, 2025). Yet Köchling and Wehner (2020) cautioned that biased outcomes can surface at multiple points in the hiring pipeline — data imbalance, feature representation, labeling practices, and evaluation choices — and De Lima *et al*. (2023) documented how such gender bias patterns recur across AI systems with notable consistency. Soleimani *et al*. (2025) added that reducing bias in recruitment and selection requires attention to both technical controls and process-level governance. Chaturvedi and Chaturvedi (2025) extended this concern to newer paradigms, providing evidence that Generative AI can reproduce occupational segregation patterns that disadvantage female candidates in certain roles — which suggests fairness scrutiny must follow the technology as it evolves.

Although existing studies have produced valuable evidence, Mori *et al*. (2024) and Mujtaba and Mahapatra (2024) noted that the literature tends to treat bias mechanisms, fairness evaluation, and governance considerations as separate concerns, which limits coherent understanding for practitioners. Mehrabi *et al*. (2021) observed that fairness conclusions depend on definitional and metric choices, while Kassir *et al*. (2023) argued that mitigation claims must be assessed within real hiring workflows rather than through isolated model statistics. Zhou *et al*. (2023) and Dablain *et al*. (2024) showed that imbalance-handling strategies and fairness objectives can interact in non-trivial ways during training and evaluation — a point that makes consolidated synthesis necessary rather than optional. In the Indonesian context, related discussions have addressed the growing role of AI in recruitment and HR processes, though these tend to concentrate on implementation benefits rather than fairness evaluation or governance mechanisms (Zurnali & Wahjono, 2022; Pujianto & Jamaluddin, 2023; Iwan *et al*., 2023). Broader digital discourse studies also suggest that gender bias may remain present in adjacent sociotechnical environments beyond formal hiring systems (Meisda & Moedjahedy, 2024). Against this backdrop, the present review was designed to produce an integrated synthesis connecting bias mechanisms, fairness measurement and auditing practices, and mitigation or governance strategies within a single analytical frame.

This study conducts a Systematic Literature Review (SLR) of research on gender inequality in AI-based recruitment systems published between 2020 and 2025. The timeframe was selected to capture recent developments in AI-assisted recruitment, including increased attention to fairness evaluation, accountability, auditability, applicant interaction, and emerging concerns related to Generative AI in hiring contexts (Singh, 2025). Studies in other applied AI settings have also suggested that gendered responses to AI integration may remain relevant across domains, reinforcing the case for closer attention to fairness in adoption contexts (Fihris *et al*., 2024). The objectives of this study are:

1) to analyze reported forms and sources of gender bias in AI-based recruitment systems;
2) to identify technical factors associated with biased screening and selection outcomes;
3) to synthesize mitigation strategies and governance practices discussed in prior studies; and
4) to identify research gaps and practical implications for developing fair and ethically responsible AI-assisted recruitment.

The paper proceeds as follows. Section 2 reviews background concepts on bias and fairness in AI-assisted hiring. Section 3 describes the SLR methodology, including search strategy, selection criteria, and synthesis procedure. Section 4 reports the results and discussion. Section 5 concludes with recommendations for future work.

2. Literature Review

This section reviews prior research relevant to gender inequality in AI-based recruitment systems. The literature is organized thematically to map state-of-the-art findings, common evaluation practices, and recurring limitations that motivate the need for an integrated synthesis.

2.1 State-of-the-art Research

Köchling and Wehner (2020) synthesized evidence on discrimination and fairness in algorithmic decision-making for HR recruitment and development, mapping recurring risks related to data representativeness and decision criteria. Mori *et al*. (2024) reviewed AI in recruiting and selection through an ethics lens, emphasizing governance requirements such as transparency, accountability, and auditability. Complementing these HR-focused reviews, De Lima *et al*. (2023) and Tronnier *et al*. (2024) mapped gender bias in AI more broadly and showed how gendered stereotypes become embedded in data and model behavior — findings that remain directly relevant to recruitment systems even when the original studies were not recruitment-specific. On the question of practice-grounded evaluation, Raghavan *et al*. (2020) critically examined bias mitigation claims in algorithmic hiring and argued that fairness cannot be inferred from automation alone, because real-world practices and evaluation design materially shape outcomes. Kassir *et al*. (2023) reinforced this position by

arguing that disparate impact assessment in hiring must be interpreted within employment research constraints and workflow context, not through isolated model metrics.

Regarding fairness metrics and mitigation directions, Mujtaba and Mahapatra (2024) compiled fairness challenges in AI-driven recruitment alongside the metrics and methods used to evaluate and address discrimination. Swaroop (2025) examined fairness audits in AI recruitment tools, stressing the role of monitoring, documentation, and accountability mechanisms in deployed systems, while Soleimani *et al*. (2025) argued for combined interventions spanning process-level governance and technical controls. At the data level, Mehrabi *et al*. (2021) provided foundational definitions of bias and fairness in machine learning and explained why group disparities persist when training data and labels encode historical inequities. Dablain *et al*. (2024) bridged algorithmic fairness and imbalanced learning, showing that imbalance-handling and fairness objectives can interact in non-trivial ways during training and evaluation. Zhou *et al*. (2023) examined oversampling approaches such as SMOTE as a data-level option that may affect fairness outcomes, though the broader literature treats such techniques as one component of a larger strategy rather than a standalone solution. Finally, Chaturvedi and Chaturvedi (2025) provided evidence that Generative AI can reproduce occupational segregation patterns, suggesting that fairness concerns extend beyond classical predictive screening to newer recommendation and conversational AI systems — a finding that warrants attention as generative tools become more common in hiring workflows.

2.2 Comparison with Previous Studies

Table 1 summarizes representative studies and reviews, comparing scope, methods, typical evidence type, and commonly reported limitations.

Table 1. Comparison of representative studies on gender bias and fairness in AI-based recruitment

Study	Type	Main Focus	Evidence/Method	Strengths	Limitations commonly reported
Köchling & Wehner (2020)	SLR	Discrimination & fairness in HR ADM	Systematic review	Strong mapping of discrimination issues	Broad HR scope, limited recruitment pipeline detail
Mori *et al*. (2024)	SLR	Ethics in AI recruiting/selection	Systematic review	Clear governance and ethics framing	Less emphasis on operational metric selection details
Raghavan *et al*. (2020)	Empirical/critical	Bias mitigation claims in hiring	Practice-grounded evaluation	Connects "fairness claims" to real hiring practice	Context-dependent; not a full taxonomy of mitigation techniques
Kassir *et al*. (2023)	Perspective	Hiring in context & disparate impact	Conceptual + employment research constraints	Identifies evaluation pitfalls and context needs	Limited prescriptive metric standardization
Mujtaba & Mahapatra (2024)	Survey	Fairness metrics & methods in recruitment	Review/synthesis	Compiles metrics, methods, and challenges	Often high-level; implementation details vary by setting
Zhou *et al*. (2023)	Technical study	Oversampling & fairness	Empirical + theoretical	Shows data-level intervention relevance	Not recruitment-specific; fairness depends on evaluation design
Dablain *et al*. (2024)	Technical synthesis	Fairness × imbalance	Conceptual + method framing	Clarifies interaction between imbalance and fairness	Domain-general; requires recruitment-specific operationalization

Chaturvedi & Chaturvedi (2025)	Empirical (preprint)	GenAI & gender bias	Audit-style evaluation	Extends bias discussion to GenAI	Early evidence; context boundaries and reproducibility vary

The comparison reveals that prior studies collectively provide substantial evidence of gender-bias risks and propose various mitigation directions, yet their contributions remain fragmented across ethics-focused reviews, metric-oriented surveys, and context-specific evaluations. Mori *et al*. (2024) and Swaroop (2025) stress governance requirements; Mujtaba and Mahapatra (2024) concentrate on metrics and methods; Raghavan *et al*. (2020) and Kassir *et al*. (2023) insist that fairness claims must be read within real hiring contexts. That fragmentation is not merely an academic inconvenience — it creates practical ambiguity for organizations trying to act on the literature. An integrated thematic synthesis connecting bias sources, fairness measurement and auditing practices, and mitigation and governance strategies across recruitment pipelines is therefore warranted.

2.3 Positioning of This Study

The reviewed literature shows strong but fragmented coverage of gender bias and fairness in AI-based recruitment. Some studies concentrate on ethics and governance in recruiting (Mori *et al*., 2024; Swaroop, 2025), others on fairness metrics and mitigation methods (Mujtaba & Mahapatra, 2024), and still others on practice-grounded evaluation and context sensitivity (Raghavan *et al*., 2020; Kassir *et al*., 2023). Domain-general fairness foundations and imbalance-focused research offer important technical explanations but are not always translated into recruitment-specific synthesis (Mehrabi *et al*., 2021; Dablain *et al*., 2024; Zhou *et al*., 2023). The present study is therefore positioned as an integrated SLR that draws these strands into a single thematic synthesis aligned with RQ1–RQ3, clarifying (i) bias sources and mechanisms in recruitment pipelines, (ii) fairness measurement and auditing practices, and (iii) mitigation and governance strategies, while identifying research gaps for future empirical validation.

2.4 Review of Technologies, Frameworks, and Algorithms

Mehrabi *et al*. (2021) described core fairness concepts and explained why fairness evaluation depends on outcome definitions and measurement choices. In recruitment contexts, Mujtaba and Mahapatra (2024) noted that fairness metrics must align with the decision stage — screening versus ranking — and be interpreted alongside hiring workflow constraints, a point reinforced by Kassir *et al*. (2023) and Raghavan *et al*. (2020), who cautioned that fairness claims may not transfer across settings without context-aware evaluation. On the governance side, Mori *et al*. (2024) stressed transparency and accountability requirements in AI recruiting, while Swaroop (2025) examined fairness audits as practical mechanisms for monitoring disparate impacts and improving traceability in recruitment tools — positioning audits and documentation as complements to metric-based evaluation, not substitutes for it. At the data level, Dablain *et al*. (2024) argued that imbalanced learning and algorithmic fairness are closely linked because under-representation can systematically degrade model performance for protected groups. Zhou *et al*. (2023) examined oversampling strategies such as SMOTE as a data-level approach that may affect fairness outcomes; the broader literature, however, consistently indicates that rebalancing alone does not guarantee fairness without appropriate metrics, audits, and governance practices (Mujtaba & Mahapatra, 2024; Kassir *et al*., 2023).

3. Methodology

This study applied a qualitative Systematic Literature Review (SLR) to synthesize evidence on gender inequality in AI-based recruitment systems. The SLR approach was selected to systematically identify, evaluate, and integrate findings from prior studies addressing gender bias, fairness measurement, and ethical governance in algorithmic hiring. The review covered publications from 2020 to 2025 to capture recent developments in AI-assisted recruitment, including emerging discussions on fairness auditing, accountability, and generative AI systems. The overall SLR procedure follows the workflow illustrated in Figure 1, comprising identification, deduplication, screening, full-text eligibility assessment, quality assessment (QA), and final inclusion. The selection process began with 47 initial records and was progressively filtered to 10 studies retained in the final synthesis.

3.1 Research Questions

The SLR was guided by the following research questions (RQs):

1) RQ1: What forms and sources of gender bias are reported in AI-based recruitment systems?

2) RQ2: How do studies measure and/or audit fairness with respect to gender in hiring contexts?
3) RQ3: What mitigation strategies — technical and governance-oriented — are proposed or evaluated to reduce gender bias?

3.2 Literature Search Strategy

The SLR workflow (Figure 1) consisted of identification, deduplication, screening, full-text eligibility assessment, and quality assessment leading to final inclusion. This structure was adopted to ensure a transparent and systematic review process and to align the literature selection procedure with the study objectives and research questions.

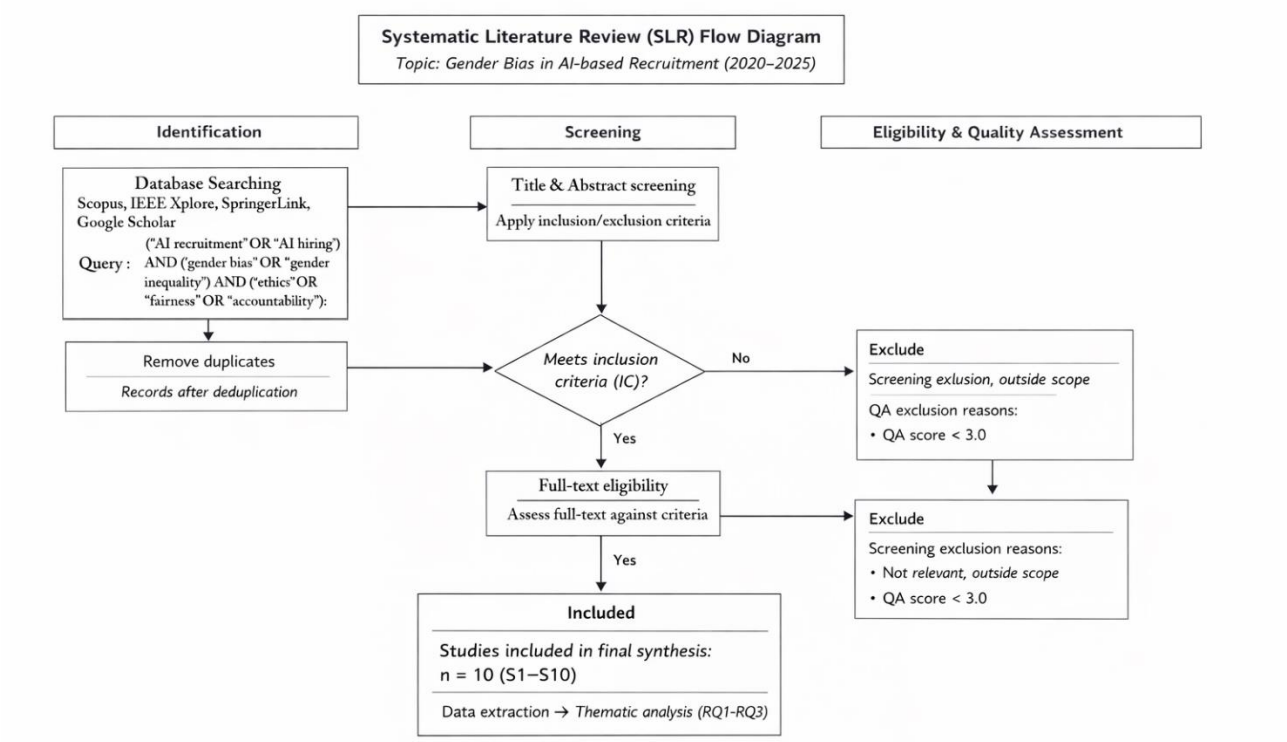

Figure 1. Systematic Literature Review (SLR) flow diagram.

3.2.1 Identification

The identification stage aimed to retrieve candidate studies relevant to gender inequality in AI-based recruitment systems, covering publications from 2020 to 2025. Records were identified using a structured Boolean search strategy adapted from Carrera-Rivera *et al.* (2022), with search terms constructed by combining three keyword groups: (1) AI/ML-related terms, (2) recruitment-related terms, and (3) gender bias and fairness-related terms, as follows:

1) AI/ML terms: "artificial intelligence", "machine learning", "algorithmic"
2) Recruitment terms: "recruitment", "hiring", "resume screening", "selection", "candidate ranking"
3) Gender bias/fairness terms: "gender bias", "fairness", "discrimination", "disparate impact", "audit", "accountability"

The initial search returned 47 records. Duplicate records were then removed by comparing titles, authors, publication years, and available bibliographic metadata, following the procedure described by Carrera-Rivera *et al.* (2022). Five duplicate records were removed, leaving 42 records for title and abstract screening.

3.2.2 Screening

The screening stage involved manual review of titles and abstracts before full-text assessment, guided by the predefined inclusion and exclusion criteria in Table 2. Studies were retained if they were relevant to AI-based recruitment systems and addressed gender bias, fairness, discrimination, auditing, or governance. Twenty records were excluded at this stage, leaving 22 studies for full-text eligibility review.

Table 2. Inclusion and Exclusion Criteria

Type	Code	Criteria
Inclusion	IC1	Published in 2020–2025
Inclusion	IC2	Written in English or Indonesian and available in full text

Inclusion	IC3	Discusses AI/ML or algorithmic decision-making in recruitment/selection (screening, ranking, assessment)
Inclusion	IC4	Includes discussion of gender bias, fairness, disparate impact, or algorithmic ethics related to gender
Exclusion	EC1	Opinion/editorial/news/blog; non-scholarly or unverifiable publication
Exclusion	EC2	AI domain not related to recruitment (e.g., education only; HR topics unrelated to hiring)
Exclusion	EC3	Duplicate record
Exclusion	EC4	No analyzable discussion of gender bias/fairness/audit/mitigation

3.2.3 Full-Text Eligibility

Studies passing the title and abstract screening were retrieved in full text and assessed for eligibility. Each article was required to clearly describe: (i) the recruitment setting or hiring stage addressed, (ii) the framing of gender bias or fairness in relation to recruitment outcomes, and (iii) sufficient methodological detail to support data extraction and thematic synthesis. Twenty-two full-text articles were assessed at this stage. Studies were excluded if the full text was unavailable, the context was unrelated to recruitment or selection, or the article lacked sufficient analytical or methodological detail for further review.

3.2.4 Eligibility and Quality Assessment (QA)

To ensure the robustness of the synthesized evidence, a quality assessment was applied to eligible full-text studies using the criteria in Table 3. Each QA item was scored on a three-level scale: 1.0 (Complete), 0.5 (Partial), and 0.0 (Absent), with a maximum total score of 5.0. Studies achieving a minimum QA score of 3.0 were included in the final synthesis. The assessment was conducted by a single reviewer using the predefined rubric; the possibility of subjective judgment is acknowledged as a limitation of this review. Following eligibility assessment and QA, 12 studies were excluded, and 10 studies were retained in the final synthesis.

Table 3. Quality Assessment Criteria for Included Studies

ID	Assessment Criteria	Score/Weight
QA1	Are the research objectives clearly and specifically defined and aligned with analyzing gender bias in AI-based recruitment?	1.0
QA2	Does the study explicitly focus on AI-based recruitment tasks (e.g., resume screening, candidate ranking, selection/assessment) rather than general AI/HR discussion?	1.0
QA3	Does the study identify/operationalize sources of gender bias (e.g., data imbalance, feature/label bias, evaluation criteria) and explain impacts on screening/ranking outcomes?	1.0
QA4	Does the study report fairness measurement and/or an audit approach (e.g., fairness metrics, disparate impact evaluation, audit considerations) suitable for hiring context?	1.0
QA5	Does the study discuss mitigation and governance practices (e.g., re-sampling/SMOTE, fairness interventions, transparency, audit, human oversight) and support claims with empirical evidence or credible literature (including limitations)?	1.0
Total Maximum Score		5.0

Scoring definition: 1.0 = Complete, 0.5 = Partial, 0.0 = Absent. Inclusion threshold: studies with a total QA score of 3.0 or higher were included in the final synthesis.

3.2.5 Included Studies, Data Extraction, and Synthesis

Studies passing the full-text eligibility stage and meeting the QA threshold were included in the final review for data extraction and thematic synthesis (n = 10; S1–S10). Data extraction followed a structured extraction sheet covering publication metadata, recruitment stage addressed, reported sources of bias, fairness metrics or audit approaches, mitigation strategies, and key findings or limitations. The extracted data were then compared across studies to identify recurring patterns, similarities, and differences. Synthesis was conducted through thematic analysis, organizing findings into three analytical themes aligned with the research questions: (1) sources and mechanisms of gender bias, (2) fairness measurement and auditing practices, and (3) mitigation and governance strategies.

4. Result and Discussion

This section presents the findings of the SLR and discusses their implications in relation to the research questions stated in Section 1. As this study is based on literature synthesis rather than system implementation, the results are organized into three parts: (i) study selection and quality appraisal outcomes, (ii) descriptive

characteristics of the included studies, and (iii) thematic findings addressing RQ1–RQ3. The discussion then interprets these findings in relation to prior literature, practical implications, limitations, and research gaps.

4.1 Results

4.1.1 Study Selection and Quality Appraisal Outcomes

Literature selection proceeded through the staged filtering steps described in Section 3: identification, deduplication, title and abstract screening, full-text eligibility assessment, and quality appraisal. The initial search returned 47 records. After removing 5 duplicates, 42 studies remained for title and abstract screening, at which stage 20 studies were excluded for not meeting the thematic focus of the review. The remaining 22 full-text articles were assessed for eligibility and quality; following QA, 12 studies were excluded, and the final corpus comprised 10 studies (S1–S10). Figure 2 presents a PRISMA-style flow diagram summarizing the selection process, while Table 4 reports the QA scores and categories of the included studies.

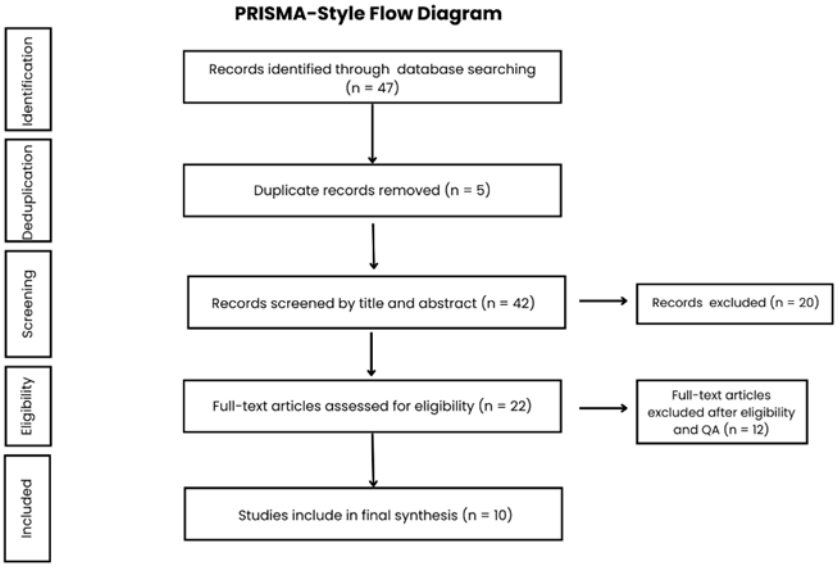

Figure 2. PRISMA-style flow diagram of study selection

As shown in Figure 2, the selection process progressively narrowed the literature set from 47 initial records to 10 studies retained in the final synthesis.

Table 4. Quality Assessment Results of Included Studies

Study ID	Main Reference	QA1	QA2	QA3	QA4	QA5	Total	Category
S1	Zhou *et al.* (2023)	1.0	1.0	1.0	1.0	1.0	5.0	High
S2	Mujtaba & Mahapatra (2024)	1.0	1.0	1.0	1.0	1.0	5.0	High
S3	Chaturvedi & Chaturvedi (2025)	1.0	1.0	1.0	0.5	1.0	4.5	High
S4	Dablain *et al.* (2024)	1.0	1.0	0.5	1.0	1.0	4.5	High
S5	Tronnier *et al.* (2024)	1.0	0.5	1.0	0.5	1.0	4.0	High
S6	Kassir *et al.* (2023)	1.0	0.5	1.0	0.5	1.0	4.0	High
S7	Raghavan *et al.* (2020)	1.0	1.0	1.0	0.5	1.0	4.5	High
S8	De Lima *et al.* (2023)	1.0	0.5	1.0	0.5	0.5	3.5	Medium
S9	Firdaus (2024)	1.0	1.0	0.5	0.5	1.0	4.0	High
S10	Shenbhagavadivu *et al.* (2024)	1.0	0.5	0.5	0.5	0.5	3.0	Medium

As shown in Table 4, most included studies were categorized as high quality, while a smaller number met the minimum inclusion threshold and were categorized as medium quality.

4.1.2 Study Characteristics of Included Studies

The included studies covered both recruitment-specific and supporting technical perspectives relevant to gender bias and fairness in AI-assisted hiring. Recruitment stages addressed in the literature included résumé screening, candidate ranking or shortlisting, recommendation or callback decisions, and broader end-to-end workflow evaluation. Methodologically, the corpus comprised surveys or SLRs, empirical audits, and conceptual or technical syntheses. Table 5 summarizes the characteristics of the included studies together with their QA totals.

Table 5. Study Characteristics of Included Studies

Study ID	Reference	Year	Study Type	Recruitment Stage Focus	Main Topic	QA Total	Category
S1	Zhou *et al.* (2023)	2023	Empirical/method	Training stage (data-level)	SMOTE/oversampling & fairness	5.0	High
S2	Mujtaba & Mahapatra (2024)	2024	Survey/review	End-to-end recruitment pipeline	Fairness challenges, metrics, methods	5.0	High
S3	Chaturvedi & Chaturvedi (2025)	2025	Empirical audit (preprint)	Recommendation/callback-like	Generative AI & gender bias	4.5	High
S4	Dablain *et al.* (2024)	2024	Research/synthesis	General ML (relevant to hiring)	Fairness & imbalanced learning	4.5	High
S5	Tronnier *et al.* (2024)	2024	SLR	General AI (gender bias)	Inclusiveness & gender bias mapping	4.0	High
S6	Kassir *et al.* (2023)	2023	Perspective	Hiring workflow (contextual)	Disparate impact & evaluation in context	4.0	High
S7	Raghavan *et al.* (2020)	2020	Critical empirical	Algorithmic hiring practices	Mitigation claims & real-world evaluation	4.5	High
S8	De Lima *et al.* (2023)	2023	SLR	General AI (gender bias)	Bias categories & mitigation overview	3.5	Medium
S9	Firdaus (2024)	2024	Empirical/qualitative	Recruitment (general)	Benefits & ethical challenges	4.0	High
S10	Shenbhagavadivu *et al.* (2024)	2024	Review/conceptual	HR recruitment (general)	AI in HR + general challenges	3.0	Medium

4.1.3 Thematic Findings (RQ1–RQ3)

The thematic synthesis organized the evidence into three themes aligned with the research questions: (1) sources and mechanisms of gender bias, (2) fairness measurement and auditing practices, and (3) mitigation and governance strategies.

1) RQ1: Forms and Sources of Gender Bias in AI-Based Recruitment

Across the included studies, the most frequently reported driver of gender bias was imbalanced or historically biased training data encoding prior hiring patterns — most notably the under-representation of women in certain occupational categories. Such imbalance increases the likelihood that models learn correlations that disadvantage female candidates in screening or ranking decisions (Raghavan *et al.*, 2020; Mujtaba & Mahapatra, 2024; De Lima *et al.*, 2023). Studies also described bias arising from feature design and proxy variables — educational history or career interruptions, for instance — that correlate with gender, as well as labeling practices reflecting subjective human judgments (Mehrabi *et al.*, 2021; Raghavan *et al.*, 2020; De Lima *et al.*, 2023). Several papers stressed that bias can emerge at multiple points of the hiring pipeline: data collection, preprocessing, model training, thresholding and ranking decisions, and downstream human–AI interaction. Evidence also indicated that human reliance on algorithmic recommendations can amplify disparities through automation bias or selective trust in model outputs (Raghavan *et al.*, 2020; Kassir *et al.*, 2023). Chaturvedi and Chaturvedi (2025) added that Generative AI systems may reproduce occupational segregation patterns that translate into gendered differences in recommendation or callback-like outcomes — a finding that warrants attention as generative tools become more common in hiring workflows.

2) RQ2: Fairness Measurement and Auditing Practices

The included literature consistently found that fairness conclusions depend strongly on how outcomes are defined and measured in hiring contexts (Raghavan *et al.*, 2020; Kassir *et al.*, 2023). Studies referenced group-based fairness metrics — selection rate parity and disparate impact ratios — alongside error-based measures such as differences in false positive and false negative rates across groups, drawing on broader algorithmic fairness foundations (Mehrabi *et al.*, 2021; Mujtaba & Mahapatra, 2024). Recruitment-specific discussions stressed that metric choice must align with the decision stage and the operational meaning of outcomes, whether that is a selection decision, a score, or downstream performance (Mujtaba & Mahapatra, 2024; Kassir *et al.*, 2023). Governance-oriented work treated auditing practices as necessary safeguards: documentation, transparency reporting, ongoing monitoring for disparate impact, and accountability structures for deployed tools (Raghavan *et al.*, 2020; Mori *et al.*, 2024; Swaroop, 2025).

Multiple studies cautioned that reporting a single fairness metric is rarely sufficient, because metrics can conflict and apparently "fair" model statistics may not reflect fairness in real hiring workflows (Mujtaba & Mahapatra, 2024; Kassir *et al.*, 2023). That caution deserves emphasis — a model that passes one fairness test while failing another is not, in any meaningful sense, a fair model.

3) RQ3: Mitigation Strategies and Governance Practices
Mitigation strategies discussed in the literature clustered into data-level, model-level, and process/governance interventions. Data-level approaches included rebalancing and resampling to address representation imbalance; oversampling methods such as SMOTE were discussed as one preprocessing option to increase minority-group representation in feature space (Zhou *et al.*, 2023). Model-level approaches included fairness-aware objective adjustments or constraints, as well as post-processing methods to reduce group disparities (Mehrabi *et al.*, 2021; Mujtaba & Mahapatra, 2024). Process- and governance-oriented interventions — transparency, routine audits, stakeholder review, and human oversight — appeared as recurring practices in responsible AI-assisted hiring across the reviewed studies (Raghavan *et al.*, 2020; Mori *et al.*, 2024; Kassir *et al.*, 2023).

4.2 Discussion

This review reinforces the position that gender bias in AI-based recruitment is a socio-technical phenomenon rather than a purely algorithmic one. Consistent with HR-focused SLRs, the evidence indicates that discriminatory outcomes are often rooted in historical data and decision criteria reflecting labor market inequalities (Köchling & Wehner, 2020; De Lima *et al.*, 2023), while ethics-focused reviews stress transparency, accountability, and governance as recurrent requirements for responsible deployment (Mori *et al.*, 2024; Mujtaba & Mahapatra, 2024). By drawing technical and governance perspectives into a single synthesis, the analysis clarifies how bias sources — data, feature, and label — evaluation choices, and organizational practices jointly shape fairness outcomes. In relation to algorithmic hiring critiques, the findings align with arguments that mitigation claims should be assessed in context and that evaluation choices can materially determine whether a tool appears to reduce disparate impact (Raghavan *et al.*, 2020; Kassir *et al.*, 2023). The prevalence of data imbalance as a reported driver also echoes work bridging algorithmic fairness and imbalanced learning, which showed that rebalancing and fairness objectives may interact in non-trivial ways during training and evaluation (Dablain *et al.*, 2024; Zhou *et al.*, 2023). The reviewed literature makes clear that technical mitigations alone do not guarantee fair hiring outcomes — rebalancing without careful metric-based evaluation, monitoring, and process controls may produce the appearance of fairness without its substance (Mujtaba & Mahapatra, 2024; Kassir *et al.*, 2023). From a practical standpoint, organizations adopting AI-assisted hiring should: (i) audit training data and outcomes for representation and proxy-driven bias; (ii) report multiple fairness metrics aligned to recruitment stages; and (iii) put in place governance mechanisms — documentation, routine audits, and human oversight — to reduce automation bias and ensure accountability (Mori *et al.*, 2024; Mujtaba & Mahapatra, 2024; Kassir *et al.*, 2023; Swaroop, 2025).

Several research gaps emerge from the synthesis. Empirical validation of mitigation strategies in real-world hiring settings remains limited relative to the volume of conceptual recommendations — a gap that is, at this point, difficult to ignore. There is also a need for more consistent reporting of outcome definitions, demographic group handling, and audit protocols to enable comparability across studies (Mujtaba & Mahapatra, 2024; Kassir *et al.*, 2023). Emerging evidence on Generative AI further suggests that future work should extend fairness scrutiny beyond classical predictive screening to include generative recommendation and conversational hiring tools, with careful attention to context boundaries and reproducibility (Chaturvedi & Chaturvedi, 2025). Although the search strategy was designed to be systematic, relevant studies may exist outside the selected search scope or may use different terminology, and the QA process depended on methodological detail as reported in publications — applied sources with limited transparency may have been excluded. These limitations should be considered when generalizing the conclusions, and they motivate continued efforts toward standardized reporting and open evaluation in AI hiring research (Raghavan *et al.*, 2020; Kassir *et al.*, 2023).

5. Conclusion and Recommendations

This study synthesized evidence on gender inequality in AI-based recruitment through a qualitative Systematic Literature Review (SLR) covering publications from 2020 to 2025. Gender bias is most consistently associated with historically imbalanced training data, with additional contributions from proxy features, labeling practices, and human–AI interaction. Fairness conclusions depend strongly on how hiring outcomes are defined and measured — which is precisely why the reviewed literature calls for multiple fairness metrics supported by auditing practices, while mitigation approaches span data-level, model-level, and governance or process-

oriented interventions. Effective bias reduction, the evidence suggests, requires both technical measures and organizational safeguards working in tandem; neither is sufficient on its own.

The main contribution of this paper is an integrated, QA-filtered thematic synthesis aligned with RQ1–RQ3, based on a final corpus of 10 included studies (S1–S10). The synthesis provides a structured reference connecting bias sources, fairness evaluation practices, and governance implications in AI-based recruitment systems — bringing together strands that prior literature has largely treated in isolation. Several limitations should be acknowledged: the review scope is bounded by the selected search strategy and the 2020–2025 timeframe; the synthesis relies on methodological detail as reported in the included studies; and QA scoring may carry some subjectivity given that the assessment was conducted by a single reviewer.

Future work should expand empirical validation in real-world hiring settings, where conceptual recommendations remain largely untested. More standardized reporting of outcomes and audit protocols would also improve comparability across studies — a practical step the field can take without waiting for new empirical data. Fairness analysis should further extend to emerging Generative AI hiring tools, where occupational segregation patterns may reproduce in ways that classical screening models do not fully capture. Stronger transparency and auditing standards — reporting checklists, metric disclosure requirements, monitoring frequency specifications, and defined accountability structures — may help consolidate both practice and policy in responsible AI-assisted recruitment.

References

Carrera-Rivera, A., Ochoa, W., Larrinaga, F., & Lasa, G. (2022). How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, *9*, 101895. https://doi.org/10.1016/j.mex.2022.101895

Chaturvedi, S., & Chaturvedi, R. (2025). *Who gets the callback? Generative AI and gender bias* (arXiv:2504.21400). arXiv. https://doi.org/10.48550/arXiv.2504.21400

Cowgill, B. (2019). *Bias and productivity in humans and machines: Theory and evidence from résumé screening* (Working Paper). W. E. Upjohn Institute for Employment Research. https://doi.org/10.2139/ssrn.3433737

Dablain, D., Krawczyk, B., & Chawla, N. V. (2024). Towards a holistic view of bias in machine learning: Bridging algorithmic fairness and imbalanced learning. *Machine Learning and AI*, *2*(1), 1–20. https://doi.org/10.1007/s44248-024-00007-1

De Lima, R. M., Corrêa, V., & Pisker, B. (2023). Gender bias in artificial intelligence: A systematic review of the literature. *SN Computer Science*, *4*(1), 1–18. https://doi.org/10.1007/s42979-022-01377-6

Deloitte. (2023). *Scaling AI across talent management in financial services organizations*. https://www.deloitte.com

Fihris, Alfianika, N., & Nasikhin, N. (2024). Differences in male and female responses to artificial intelligence integration for education faculty: Study of Thailand international students at Islamic universities in Indonesia. *eL-HIKMAH: Jurnal Kajian dan Penelitian Pendidikan Islam*, *18*(1), 31–60. https://doi.org/10.20414/elhikmah.v18i1.10037

Firdaus, A. (2024). Implementasi artificial intelligence dalam rekrutmen: Manfaat dan tantangan di industri 4.0. *J-MAS (Jurnal Manajemen dan Sains)*, *9*(2), 1615. https://doi.org/10.33087/jmas.v9i2.2083

Iwan, C., Putra, C. K., Zabdi, D., Boy, E. I., Chandra, M. A., & Febrianti, L. Y. (2023). Analisis pemanfaatan artificial intelligence dalam membantu proses perekrutan karyawan perusahaan. *Jurnal Sains dan Teknologi*, *2*(2), 161–168. https://doi.org/10.58169/saintek.v2i2.248

Kassir, S., Baker, L., Dolphin, J., & Polli, F. (2023). AI for hiring in context: A perspective on overcoming the unique challenges of employment research to mitigate disparate impact. *AI and Ethics*, *3*(2), 199–214. https://doi.org/10.1007/s43681-022-00173-9

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, *13*(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w

Kristian, W. (2024, August). *AI dalam rekrutmen: Dari sumber daya manusia ke robot pencari kerja*. BINUS University. https://binus.ac.id/bekasi/2024/08/ai-dalam-rekrutmen-dari-sumber-daya-manusia-ke-robot-pencari-kerja/

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Meisda, R., & Moedjahedy, J. (2024). Sentiment analysis of the most viewed YouTube video: Exploring gender bias in the discussion of women workers in Indonesia. *YUME: Journal of Management*, *7*(1), 186–197. https://doi.org/10.37531/yum.v7i1.6311

Montesa, M. (2025). *AI recruiting in 2025: The definitive guide*. Phenom. https://www.phenom.com/blog/recruiting-ai-guide

Mori, M., Sassetti, S., Cavaliere, V., & Bonti, M. (2024). A systematic literature review on artificial intelligence in recruiting and selection: A matter of ethics. *Personnel Review*. https://doi.org/10.1108/PR-03-2023-0257

Mujtaba, D. F., & Mahapatra, N. R. (2024). *Fairness in AI-driven recruitment: Challenges, metrics, methods, and future directions* (arXiv:2405.19699). arXiv. https://doi.org/10.48550/arXiv.2405.19699

Pujianto, W. E., & Jamaluddin, M. (2023). Pengaruh artificial intelligence (AI) terhadap rekrutmen karyawan. *Jurnal Manajemen dan Teknologi*, *12*(2), 45–56.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency* (pp. 469–481). ACM. https://doi.org/10.1145/3351095.337282

Shenbhagavadivu, T., Poduval, K., & V. V. (2024). Artificial intelligence in human resource: The key to successful recruiting and performance management. *ShodhKosh Journal of Visual and Performing Arts*, *5*(3). https://doi.org/10.29121/shodhkosh.v5.i3.2024.1351

Singh, R. (2025, January). *How AI-driven chatbots enhance candidate experience in 2025* [LinkedIn post]. LinkedIn. https://www.linkedin.com

Soleimani, M., Intezari, A., Arrowsmith, J., Pauleen, D. J., & Taskin, N. (2025). Reducing AI bias in recruitment and selection: An integrative grounded approach. *The International Journal of Human Resource Management*. https://doi.org/10.1080/09585192.2025.2480617

Swaroop, N. (2025). The bias detection and fairness audits in AI recruitment tools. *International Journal of Modern Science and Research Technology (IJMSRT)*, *3*(4), 323–329. https://ijmsrt.com/articles/view/the-bias-detection-and-fairness-audits-in-ai-recruitment-tools

Tronnier, F., Löbner, S., Azanbayev, A., & Walter, M. L. (2024). A systematic literature review on gender bias in AI — Towards inclusiveness in machine learning. In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS)*. AIS Electronic Library. https://aisel.aisnet.org/pacis2024/track01_aibussoc/track01_aibussoc/3

Zhou, Y., Kantarcioglu, M., & Clifton, C. (2023). On improving fairness of AI models with synthetic minority oversampling techniques. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)* (pp. 1134–1145). SIAM. https://doi.org/10.1137/1.9781611977653.ch98

Zurnali, C., & Wahjono, A. (2022). Artificial intelligence dalam rekrutmen. *Jurnal Ilmiah INFOKAM*, *9*(2), 50–57.