European Economic Letters
ISSN 2323-5233
Vol 16, Issue 2 (2026)
http://eelet.org.uk

# Exploring Gender Bias Against Women in STEM: A Qualitative Case Study Analysis

**Kyra Anand,**

Jumeirah College, Dubai

anandkyra1@gmail.com

**ABSTRACT**

The increasing amount of artificial intelligence (AI) employed within STEM ecosystem are raising many concerns for businesses using algorithm-based decision-making systems that are shown to embed gender-related biases. The purpose of this investigation was to analyze whether AI methods of recruiting, evaluating, and promoting individuals may re-introduce or increase structural disparities among genders within STEM fields. This research is grounded in socio-technical systems theory; thus, the researcher defines algorithmic bias as being structural in nature and caused by historically biased datasets, proxy variables, and optimization models, placing predictive accuracy over distributional equality. The methodology included qualitative, integrative research techniques that included a structured review of literature, as well as an analytic review of case studies of biased AI implementation in industry. A comparative analysis of fairness-aware design methodologies, debiasing strategies, and governance modalities was used to help assess their effectiveness in reducing bias. While the results indicate that technical fairness interventions produce statistically significant reductions in measurement biases such as statistical parity and corresponding opportunities, the researchers conclude that they are inadequate without implementing corresponding institutional and governance reforms. The analysis determined that the long-term reduction of gender biases existing in AI systems requires a holistic strategy that includes technical redesigning, inclusive data governance practices, and accountability from organizations. Finally, the researcher presents a theoretically integrated conceptual model for creating an equity-based AI design and implementation framework within STEM ecosystem.

**Keywords:** Algorithmic Gender Bias, Artificial Intelligence in STEM, Algorithmic Fairness and Governance, Socio-Technical Systems Theory, Equitable AI Design

## I. Introduction

The increasing prevalence of AI in making important decisions raises an important question - what will happen when AI takes our existing societal biases against women, and exacerbates them? Gender bias is defined as a detrimental action or thought process that surfaces due to the presumption that women lack the equal rights and dignity of men (Botella et al., 2019). Such bias is not new and continues to exist and flourish within society through social constructs and norms that have affected the language used and data collected. It is important to note that gender has been created as an ongoing social construction to define what it means to be "feminine" or "masculine" based on power and sociocultural influences (Risberg et al., 2009). These sociocultural norms encourage negative stereotypes related to the roles women vs. men should play in society and to their expected behavior through socialization - for example, encouraging girls to be creative vs. boys being technical (Allegrini, 2014). Furthermore, language itself includes this bias. The use of terms such as "female lawyer" is an example of how bias is encoded into language and how, the use of masculine terms such as "chairman" to describe the head of an organization reinforces bias against women being leaders (Leavy et al., 2020). The gender biases present within society, as well as unconscious biases that still exist, can be found in the large datasets that AI uses to learn and develop (Leavy et al., 2020). Facial recognition systems have higher error rates for women and people with darker skin colours because of their underrepresentation in the training sets (Cirillo et al., 2020; Besse et al., 2020). A number of AI technologies, including predictive policing and hiring algorithms, have been found to perpetuate and reproduce the negative outcomes associated with bias. For example, women tend to be underrepresented in the hiring process and are underreported when it comes to job offers, while people of color tend to be over-policed. Because these systems often do not show the effects of bias, they can create records that appear to be neutral, when in fact they can have a very real impact on people's lives. For example, in the healthcare field, biases in how women are treated create significant differences in treatment that are unmotivated (Risberg et al., 2009), and there is insufficient research on women's biological conditions because most research is conducted on men. Similarly, in hiring and economic opportunities, AI systems with bias will have an impact on the number of women who are hired and the salaries offered (Kuppler, 2022); examples include Amazon's recruitment algorithms that showed biases toward women (Kordzadeh & Ghasemaghaei, 2021). These instances illustrate how the underlying biases continue to shape the structures we use, not only in language and behaviour but also in the systems we

create. Furthermore, because human biases are embedded in the fabric of human society and language, they create discrimination in hiring and healthcare (Cirillo et al., 2020; Risberg et al., 2009). Finally, because machine learning algorithms learn from human data, they will gain and propagate human bias (Xu et al., 2017).

The overall prevalence of gender-based discrimination in both society and in technical work environments spills over into its applicability to STEM disciplines and careers. There is a widespread perception that there an under-representation of women across STEM fields, known as the "Gender Gap in STEM". There exists a plethora of definitions of this perceived under-representation; Patterson et al.'s definition consists of the proportion of total working professionals in STEM fields being only 22% female (2020), while there is also a definition noting that females earned only 26% of all degrees in STEM fields as defined by their respective disciplines (Ceci et al., 2023). As a result of a long history of cultural and educational traditions that restrict females from pursuing careers in technical occupations, along with a lack of exposure to female role models in the field of science and/or technology, females have been subjected to negative implications being placed upon their ability to compete with their male counterparts in the fields of science and/or technology solely because of their gender, regardless of their technical abilities (Xu et al., 2017). Additionally, the negative use of artificial intelligence due to gender-based discrimination also continues to enhance existing disparities between men and women in technical fields (Botella et al., 2019). Despite the fact that women comprise the majority of all college graduates in the U.S. (57% Bachelors; 59% Masters; 53% PhDs) (Stewart et al., 2020), the proportion of women who are graduating with technical degrees is very small -35% of all Bachelor's degrees in STEM and 34% of all PhD's in STEM (Gómez-Talal et al., 2025). This lack of representation of women in technical fields is further reflected when examining that female students comprise only 3% of the IT sector workforce worldwide and 8% of the engineering sector workforce worldwide (Botella et al., 2019), and comprise only 8% of the total number of software engineers working in the U.S. (Allegrini, 2014). Even though there is no significant difference in the performance of females and males on test measuring knowledge/ability in science and/or technology, females continuously express a lack of confidence in comparing themselves with males who possess the same level of knowledge/ability (Stewart et al., 2020). Even when they perform at the same level as men, women are more likely to say that they have less confidence in themselves (Stewart et al., 2020). Academic and career pathways are impacted by entrenched gender biases and a lack of representation, leading to the continued existence of these disparities due to the existence of systems that favour certain groups over others (Hall & Ellis, 2023).

Equality is defined as giving everyone the same treatment or resources, no matter what their circumstances are, whether they have had disadvantages for a long time or are any other reason, so that there can be equality of outcome between groups (Risberg et al., 2009; Kordzadeh & Ghasemaghaei, 2021). On the other hand, equity is providing individuals with what they need based on systemic barriers to an equitable outcome and their needs (Risberg et al., 2009). For example, equality reflects a commitment to all individuals having equal access and value as members of society regardless of gender (Risberg et al., 2009; Siddique et al., 2024). In systems relying on algorithms, this mostly relates to providing equal treatment and value to all individuals (Besse et al., 2020). In contrast, equity deals with the issue of discrimination head-on (Risberg et al., 2009).

The distinction between these approaches is visually captured in the “Perspectives on Gender Equity” model below (Risberg et al., 2009).

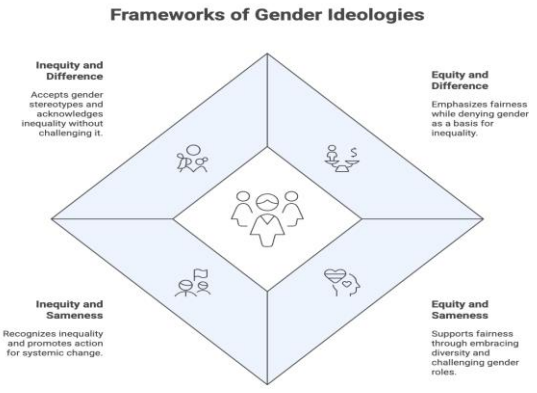


Figure 1: Model explaining perspectives on gender equality and equity

This framework provides an opportunity to look at how different outcomes emerge from addressing systemic inequity and gender-based differences along the axes of equity/inequity and sameness/difference across four quadrants. The concept of "equity and sameness, for instance, supports women's access to complete or full performance/participation opportunities by addressing traditional notions of gender role expectations." In contrast, "equity and difference" support women's access to complete or full performance/participation opportunities while taking into account the unique performance/participation requirements that arise because of gender differences (Risberg, et al., 2009). "Equal treatment of all persons in STEM does not adequately recognise the depth of the inequities faced by women in STEM fields, and, therefore, requires that other forms of intervention focused on equity be put into place" (Stewart et al., 2020; Ceci et al., in press). Socialisation based on gender from an early age reduces girls' self-confidence (Allegrini, 2014), and continued bias in (systemic) workplace environments perpetuate the "leaky pipeline" phenomenon (Balducci, 2023). Even seemingly neutral AI systems exacerbate bias by being trained against biased datasets (Duan et al., 2025). Equity is a societal concern because algorithm-based systems replicate social structure inequities (Cirillo et al., 2020), resulting in negative impact including, but not limited to, exclusion from the economy and stigmatisation (Manresa-Yee & Ramis, 2022). To achieve true justice, past barriers must be recognised and solutions tailored to those barriers are essential (Mulvey et al., 2022).

As stated in the research of Balducci (2023), current STEM institutions promote gender bias through social norms and stereotypes which dissuade women from entering the field and also by the absence of female role models (Allegrini, 2014; Kong et al., 2020). Environmental factors that are male-dominated, along with discrimination against women and the lack of authority in decision making, limit women (Balducci, 2023). Gender bias has also been found to exist in regards to teaching evaluations, salaries, and in the pathways to promotion (Stewart et al., 2020). The existence of gender bias in AI systems is perpetuated by the existence of trained algorithms that reflect the inequalities present in training data; this perpetuates gender bias by producing gender stereotype based personality types or job suggestions (Leavy et al., 2020). When AI systems produce gender-discriminatory outputs, they reproduce gender biases by creating visual representations of engineers as men and nurses as women (Hall & Ellis, 2023). The reproduction of gender bias in AI originates from the process through which societies build and maintain essentialist beliefs (Balducci, 2023).

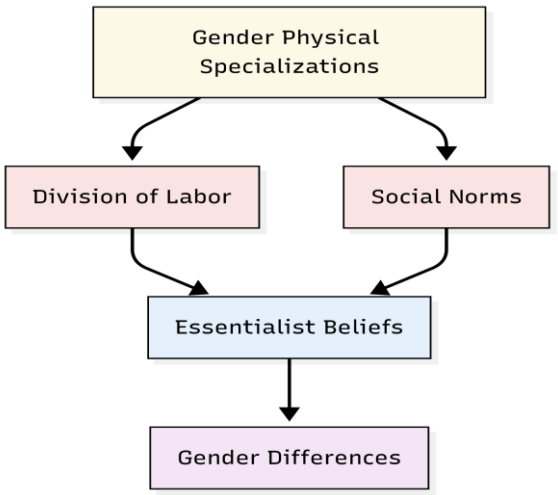


Figure 2: Causal Pathway from Physical Gender Differences to Structural Inequality

In science, technology, engineering and mathematics (STEM) the notion that men are inherently more capable of fulfilling technical and/or leadership functions reinforces stereo-typical gender bias in social and institutional frameworks (Balducci, 2023). While women may demonstrate equal levels of success through measurable academic performance, such as receiving grant funding and having research published; the existence of biased evaluation systems for student and faculty performance will continue to perpetuate inequity through wage disparities (Stewart et al., 2020). The systemic factors related to career interruption and gender expectations will help perpetuate inequities (Stewart et al., 2020). For example, many historical design standards, such as vehicle crash test dummies being based on a male body type contribute to creating unsafe conditions for women (Gómez-Talal et al., 2025). Many of the social constructs related to language and default are male-centric which provides implicit evidence of gender bias (Xu et al., 2017). Lastly, as the output produced by artificial intelligence follows the same trends, it reflects the continued presence of broad-scale systematic structures promoting a

male bias (Cirillo et al., 2020). The systemic foundations of these issues will continue to promote gender inequities in both higher education and technology (Balducci, 2023).

Machine learning (ML) and artificial intelligence (AI) (Duan et al., 2025; Besse et al., 2020) increase the likelihood of bias reproducing and even intensifying societal inequities, unless addressed directly. The use of biased data to develop AI models creates its own set of problems because the models replicate society's previously established inequities (Duan et al., 2025; Hall & Ellis, 2023). This is largely due to the fact that AI algorithms are predominantly reliant on training datasets that have pre-existing social biases that exist as a result of human conduct, language, and culture (Duan et al., 2025; Leavy et al., 2020). Thus, AI algorithms frequently produce results that systematically disadvantage specific segments of the population (Duan et al., 2025; Leavy et al., 2020). AI can also produce the same or an even greater level of bias due to the increased volume of data used to develop and train machine learning algorithms that contain pre-existing bias (Hall & Ellis, 2023; Müller, 2025). Many real-world examples highlight this issue: For example, when developing Amazon's CV screening tool, it was trained on a predominately male history of CVs (Feldman & Peake, 2021). Similarly, Google Translate defaulted to a gendered masculine for many occupational translations when translating from a gender-neutral language (Cirillo et al., 2020; Leavy et al., 2020). The Gender Shades project identified significant accuracy differences across race and ethnicity, as well as a significantly higher error rate in identifying women of color in automated facial recognition technologies than other categories, such as women and men of European descent or men of color (Besse et al., 2020). Artificial intelligence (AI) tools, including ChatGPT, have given female individuals commonly accepted (or stereotypically sent ascribed) personality traits and have furthermore disproportionately devalued their competency levels for the same position when compared with male candidates on CVs or resumes (Leavy et al., 2020). The potential for this AI-driven bias to negatively impact individuals who fall into the category of woman ("gender based discrimination") while making decisions in human-based, perspective taking roles (such as healthcare, employment and criminal justice) (Duan et al., 2025; Besse et al., 2020) is significant; it could result in a denial of services, reduced employment opportunities, or unwarranted convictions, all of which create greater socioeconomic disparity and injustice across our societies (Leavy et al., 2020).

Machine Learning (ML) uses algorithms trained on previously collected data to create models based on patterns found in the data (Manresa-Yee & Ramis, 2022; Ibarra-Vazquez et al., 2024). To achieve this goal, the typical process starts with collecting data, training the model, generating a prediction, and then deploying the model into real-world use cases (Manresa-Yee & Ramis, 2022). The ability for the algorithm to output any form of outputs depends primarily upon the quality of the data. Additionally, if there is an inherent bias present in the input data that influences the training of the algorithms used to build the ML model, then the resulting predictions will either reproduce or compound any expected bias leading to inequity/discrimination (Manresa-Yee & Ramis, 2022; Duan et al., 2025). The primary difference between traditional ML “black box” models which produce results based solely upon maximizing predictive accuracy without regard for transparency or explainability versus those that utilize XAI (explainable artificial intelligence) methods provide an understanding of the basis for the ml predictions, thus allowing humans to obtain an adequate explanation behind the predictions and the rationale for producing the output (Zanellati et al., 2024). The use of techniques such as LIME allow transparency of the input data by providing information regarding which of the inputs had the most influence on a particular ml model decision (Feldman & Peake, 2021). Interpretability is crucial for fostering fairness and building trust in these systems (Zanellati et al., 2024). This is particularly vital in high-stakes domains such as education, healthcare, human resources, and criminal justice, where algorithmic decisions can significantly impact individuals' lives and opportunities (Duan et al., 2025). To address these concerns, solutions include incorporating fairness constraints directly into ML algorithms and establishing comprehensive ethical AI frameworks and policies to ensure unbiased and equitable decision-making processes (Leavy et al., 2020; Manresa-Yee & Ramis, 2022).

There are huge ethical implications and concerns with ML models, specifically in the areas of fairness and accountability (Duan et al., 2025; Leavy et al., 2020). When we talk about fairness in algorithms and decision-making processes it often means trying to achieve equitable outcomes instead of just achieving equal treatment to prevent these algorithms from perpetuating society’s already existing bias (Kordzadeh & Ghasemaghaei, 2021; Leavy et al., 2020). The goal here is trying to ensure that the algorithms will not disproportionately put protected classes, such as gender and race, at a disadvantage due to their characteristics (Kordzadeh & Ghasemaghaei, 2021; Manresa-Yee & Ramis, 2022). Unfortunately, since AI is

trained on data generated by humans, they would also inherit and amplify pre-existing social bias due to the way they were built; therefore, the end result is typically discriminatory (Kordzadeh & Ghasemaghaei, 2021; Leavy et al., 2020). The implications of this are really detrimental, especially to women, who continue to experience the "glass ceiling," because when recruitment algorithms use historical data that is created by men, the result is a bias against women, which affects the likelihood of being hired or receiving job offers and associated salary (Leavy et al., 2020; Feldman & Peake, 2021).

The goal of this paper is to create a predictive model to identify and measure gender biases in AI systems that affect women's careers in STEM. The focus of the project will be on hiring algorithms and facial recognition technologies; investigate how career gaps, gender based expectations in life sciences, and historical issues in design (male-centric standards) continue to contribute to these disparities; evaluate algorithmically based mitigation strategies (such as adversarial debiasing) to achieve more equitable outcomes in tech recruitment and evaluation processes while preserving model utility.

Black-box models present serious dangers when they are neither accountable nor transparent. Although these models may produce accurate results, the process that occurs inside the black box (such as identifying and addressing bias) will not be able to occur if it is not visible (Kordzadeh & Ghasemaghaei, 2021; Besse et al., 2020). The lack of visibility will lead to systems which run on negative feedback loops (Cirillo et al., 2020). Thus, the development, deployment, and regulation of the system is ultimately the responsibility of the developers, the organisation who is using the system, and the policymakers who will establish accountability through either ethical AI frameworks or regulation (such as, "Right to Explanation" under the GDPR; Kordzadeh & Ghasemaghaei, 2021; Cirillo et al., 2020). Additionally, there is a moral obligation for tech companies to ensure that diverse voices are involved in the development of these systems, as it will enhance the inclusion and equitable use of ethical AI through principles of care, empathetic action, and social context (Leavy et al., 2020; Müller, 2025).

## II. Literature Review

The gaps between males and females in mathematics education are persistent and often occur because of gender stereotypes and gender bias; these barriers destroy girl's self-esteem, limit her opportunity to take advanced courses, and lower her willingness to persist through school. The impact of these factors is seen in a girl's choice of university, resulting in restricted access to math-intensive majors and ultimately fewer women being prepared to enter the growth industries of STEM careers. In addition to the economic impact on women, the effect on society as a whole is a loss of innovative capacity. To close these gaps, we need early math support/programs, strong anti-discrimination policies, inclusive practices (classroom and workplace), and better metrics to track/measure and remove bias (Wang & Degol, 2017). While there are clear calls to provide early support as part of policy reform, the remaining disparities continue to vary by context and developmental phase, contributing to the unevenness of progress across settings (Wang & Degol, 2017).

While the gender gap in math performance has been closing, women continue to be underrepresented in math-intensive STEM careers. The different levels of performance and persistence among females and males can be attributed to many factors, including individual ability profiles, relative cognitive strengths, individual interests, work-family preferences, and beliefs about ability in the specific field being pursued, but also to pervasive stereotypes that exist due to both biology and sociocultural contexts at different stages of development. Moreover, despite some narrowing of gender gaps in math performance and persistence, these gaps vary based on context in low-stakes situations, they tend to be very small; however, there is a significant widening in gender gaps when measured in high-stakes testing environments, where female students are subject to stereotype threat and performance pressure which negatively impacts their outcomes. Evidence-based responses to how these gaps can be reduced include the following: early math identity development, inclusive forms of pedagogy, flexible pathways to high-level math learning and career opportunities, and institutional policy changes that support continued persistence in math-intensive careers (Ceci et al., 2023). A theory-based framework can be used to help understand the reasons that gender gaps persist despite narrowing in the last decade, and to identify potential strategies for reducing them in classrooms and institutions (Ceci et al., 2023).

Various theories have been developed to provide explanations for the reproduction of gender biases and exclusions in STEM. Social cognitive theories emphasize how, through stereotypical assumptions, the association of STEM with male traits increases the confidence of boys and decreases the interest and perseverance of girls. Theories about sociocultural and ecological systems examine how classroom and school norms, teaching practices, and the formal and informal institutional framework (both micro and macro) that define how one behaves within these contexts reinforce the different

types of socialization that are representative of inequities. Historic notions of gender have also been critiqued in feminist frameworks as the result of the historic gendering of the sciences and as such, warrant an examination of performative notions of gender and a re-examination of science as discipline and entity. All of these theories suggest some form of intervention (e.g., relational or cooperative classroom experiences; pedagogy rooted in growth-mindset theory; mentorship; inclusive curriculum; and a re-evaluation of evaluation and policy structures) that would create environments that support women in their full participation in STEM (Allegrini, 2014). The combination of all these theories creates an environment that shapes the symbols associated with the construct of belonging and identity within male-dominated environments in a way that aligns with the continuity of the gaps identified above (Allegrini, 2014).

Social cognitive as well as sociocultural research demonstrates that male-dominated STEM environments signal low self-concept/goal fit/social fit for women. In addition, subtle stereotypes create consistent expectations that women should conform to male dominant norms, adopt competitive goal orientations, and navigate exclusion from male-centered networks. Even in the absence of explicit barriers, these signals undermine both belonging and authenticity to women resulting in greater numbers of women being less interested in math intensive careers and greater attrition from them. From a feminist perspective, these environments are systems designed by men and for men, leading both men and women to perceive them as “natural” places for men but not for women. This process helps perpetuate the belief that women do not belong in STEM rather than understanding that STEM does not have an inclusive structure (Botella et al., 2019).

Using a feminist lens, the biases and inequities that occur in STEM can be studied in two ways. Firstly, it's possible to look at the "gender gap in" participation and secondly, look at the "gender dimension of" STEM construction. Feminist viewpoints view gender as a dynamic and ongoing experience, which informs who is seen as being appropriate for math-heavy fields. Feminist science studies can examine how women are portrayed in science (i.e., what kind of stories, methods, and norms they are written about). Together, the two perspectives can explain educational choices and provide guidance for actions (curriculum, policy, and institutional reform) that can change STEM cultures instead of requiring women to adapt to them (Allegrini, 2014). We further understand the gendered nature of the biology/sociology dialogue with additional biological and sociological evidence showing that the impacts of small biological nudges are shaped by strong cultural forces. This also supports the context effect described above (Allegrini, 2014; Botella et al., 2019). Although there is evidence that prenatal exposure to androgens can influence some "masculine" behaviors (nudge developmental preferences), as a whole, all nudge developmental preferences are shaped and often outweighed

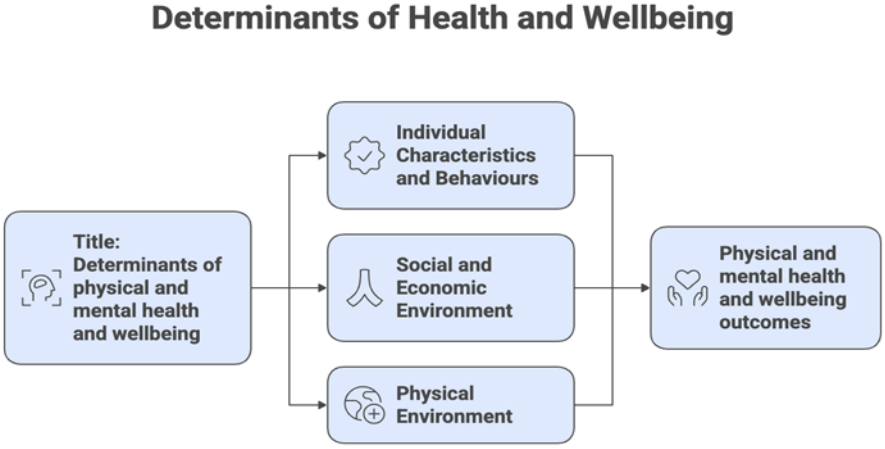


Figure *3*: Multi-level determinants influencing physical and mental health and wellbeing outcomes.

by socialization. The most common social influences such as parents, school systems, and cultural stereotypes not only shape girls/women's interests and self-perceptions of themselves but also result in girls/women thinking that mathematically oriented occupations are incompatible with people-oriented values. Thus, observed gender differences in math pathways reflect an interaction: small biological influences filtered through powerful sociocultural norms that shape motivation, identity, and choices (Balducci, 2023).

Several factors can influence how folks will do in their careers throughout their life and how those factors will affect their futures. Pivotal life changes such as marriage and having a child can lead to larger differences in both the persistence and advancement of women and men who have experienced the same transitional life events (Wang & Degol, 2017). Once again, re-calibrating one's career and lifestyle becomes important after a transitional life event, especially, in cases, where the gender of an individual will lead them to experience different types of career paths. The long hours and constant up-skilling associated with next-generation-type STEM jobs are often incompatible with the inequalities of caring for children associated with women. Mothers tend to reduce their working hours more than fathers and childless women do. Women who are unable to work and are working for low wages and have limited options regarding parental leave or meeting productivity expectations will find careers in math-based fields increasingly inaccessible (Ceci et al., 2023).

The figure 3 illustrates the multi-level determinants (both individual, social, and environmental) that can influence outcomes in both physical and mental health/well-being. Although the multi-level determinants were previously focused primarily on health outcomes, the multi-level determinants may be beneficial to understand how the interaction of multiple levels of individual, social, and environmental determinants have shaped women's persistence and success in STEM fields (Risberg et al., 2009). In educational research, there has traditionally been an assumption of inclusion within hands-on learning environments, which may be a contributing factor to the ongoing inequities seen in some progressive forms of learning (Allegrini, 2014; Botella et al., 2019). In early STEM education research in makerspaces, researchers used constructionism, constructivism and experiential learning theories to frame their designs, with an emphasis on "learning by doing" through collaboration, creativity, and designing projects. However, as STEM education researchers have discovered later from the situative approach and community of practice framework, learning is typically shaped through social contexts and is used to develop one's identity. Yet, these frameworks of STEM education research have primarily assumed all students to be included within the communities; they also have failed to examine the possibility that some students may feel excluded or disconnected (Andrews & Boklage, 2023). A critical framework is defined as an anti-deficit, justice-oriented orientation for understanding non-dominant learners and examining how various forms of power and normative structure STEM. A critical framework would interrogate how institutions maintain the status quo through reproduction of inequities in STEM and consider whose knowledge is legitimized or counted as valid within non-dominant learners in STEM spaces by interrogating the role that nominally hierarchical status distinctions have played in establishing power relationships. In terms of research in makerspaces and STEM, a critical approach situates that work within both its historical and social context, and provides actionable recommendations that would disrupt existing power structures (Andrews & Boklage, 2023).

While various institutions are attempting to find solutions to remove these barriers, one potential area for addressing these barriers is to use predictive modelling or machine learning in tandem with artificial intelligence to inform institutions on where to target their support resources. However, any predictive modelling or AI used for the purposes of finding solutions to inequity needs to be designed and governed with an equity lens to avoid perpetuating these biases described previously (Wang & Degol, 2017; Ceci et al, 2023; Allegrini, 2014). Machine learning is becoming more prevalent as a means of personalizing learning and of predicting future academic performance based on a student's academic record, which aids educators in creating targeted interventions prior to a student graduating (Guo et al., 2016). The loss of females from STEM disciplines has wide-ranging consequences for society and the economy. With the incorporation of academic performance, demographic characteristics, and key socio-economic variables, predictive ML models can identify first-year students who have a low chance of returning to college for a second year, allowing for timely, individualized interventions (Zanellati et al., 2024). The use of Educational Data Mining to predict student performance has seen great advances with the adoption of deep learning classifiers that use sparsely connected autoencoder networks. These new methods are significantly superior to previous approaches and can also support the development of early warning systems for at-risk students (Xu et al., 2017). The predictive accuracy of ensemble classifiers applied to Spanish PISA-2022 math data is higher than that obtained from traditional methodologies, while also revealing additional variables (e.g., socio-economic status, ownership of a device, and number of additional lessons) that help support understanding gender patterns (Gómez-Talal et al., 2025).

The potential for advancement of equity with artificial intelligence (AI) tools is present. However, if these tools are not properly audited, they can reinforce bias (Allegrini, 2014; Botella et al., 2019). Most AI applications reproduce the existing bias embedded in training datasets, resulting in further marginalization of women and gender diverse individuals (O'Connor & Liu, 2023). Successfully addressing gender bias in AI requires a socio-technical synthesis of social structures and technical choices (Hall & Ellis, 2023). AI has been shown to encode gender bias in a systematic way across all application

domains, indicating that the adoption of inclusive data standards along with participatory design processes, regulation, and evaluation focused on marginalized end-users must occur to effectively close existing gender gaps caused by AI (Bartl et al., 2024). There are three key stages where gender bias in AI can be reduced: data preprocessing, algorithms in process, and results of the algorithms post-processing (Siddique et al., 2024). Equity in AI must begin with addressing bias in data through structured documentation, balanced datasets, and counterfactual data augmentation (Siddique et al., 2024). Approaches for algorithmic debiasing include the use of adversarial training, integrated fairness-based optimization, and post-processing adjustments, such as equalized odds, usually combined to enhance results (Siddique et al., 2024).In order to close gender gaps due to AI, education systems reform along with either complete governance or thorough auditing must occur (Wang & Degol, 2017; Ceci et al., 2023; Allegrini, 2014; Siddique et al., 2024).

**Research Questions**

1. How do predictive hiring algorithms undervalue female candidates in STEM fields, particularly when factoring in career breaks and life sciences expectations?

2. To what extent do facial recognition systems, reliant on electrical engineering signal processing, exhibit higher error rates for women, amplifying surveillance risks in professional settings?

3. Can adversarial debiasing effectively reduce gender bias in these AI models while preserving predictive accuracy for STEM career assessments?

**Hypotheses:**

H1: Predictive recruitment models based on historical data will exhibit significant bias against female candidates with career interruptions, replicating the existing disparity in the life sciences.

H2: Facial recognition errors of women will be influenced by biased signal processing assumptions built on male norms, resulting in a higher rate of false negatives in STEM identity verification.

H3: Adversarial debiasing will reduce gender bias signals in AI models by at least as much as baseline methods, maintaining overall utility in recruitment predictions.

## III. Research Methodology

Using a mixed-method, multi-stage research design, the purpose of this study was to address the three research questions surrounding AI-based Circular Economy transitions in the steel industry. Scopus, Web of Science, and IEEE Xplore were the only sources used to support a reputable (peer-reviewed), scholarly, and quality body of evidence. The empirical results from the studies referenced above (McKinsey & Company, Deloitte, Harvard Business Review, Statista) were used to triangulate and enhance the credibility of the robustness of the set of data. Research Question #1 was answered through the PRISMA systematic review protocol, which composed of structured search instructions for identification of relevant studies on Machine Learning as it pertains to Material Design, Waste Sorting, and Reverse Logistics aspects of Steel Circularity, a Screening process to assess eligibility as well as synthesize the relevant studies in question. Research Question #2 and #3 were studied using a Comparative Study methodology, in which two of the world's largest steel manufacturers were analyzed and compared with one another, in addition to comparing their respective recycling ecosystems with regards to both leverage of AIoT enabled through RFID Block Chain systems and Digital Monitoring Systems as part of their operations.With this method of research data collection and synthesis, the current status of their remanufacturing/remaking/reusing and sustainability was thoroughly examined as it occurred in real-time. The cross-case synthesis method was applied to determine the key themes related to the technologies, operational processes, and associated sustainability achievements, such as CO2 emission reductions, energy efficiency improvements, and enhanced circularity in the supply chain. The use of systematic reviews provides the highest level of methodological validity through the integration of systematic review-based evidence and empirically obtained evidence from case studies. This also provides them with an opportunity to perform an in-depth analysis of these key areas; and has the greatest potential for generalisability.

### *A. Research Design*

Using qualitative analysis and case study methodology informed by interpretive analysis and socio-technical theory, this research seeks to explore gender bias embedded in artificial intelligence technologies being utilized in STEM fields such

as electrical engineering, hiring algorithms, face recognition systems, etc. The objective of this research is to use a structured thematic synthesis of published case studies (as opposed to primary statistical modelling) of high-impact cases of bias to identify structural mechanisms, institutional mechanisms, and algorithmic mechanisms that contribute to the reproduction of bias. The research methodology is influenced by multi-level gender bias frameworks and socio-technical AI governance literature, providing an opportunity to integrate feminist theory, engineering design logic, and algorithmic fairness literature into a single theory-based analytic lens. This interpretive methodology is appropriate because bias in algorithms is not only a computational error but a systemic phenomenon as a result of historical data, institutional norms, and the environment in which they are deployed.

### *B. Data Sources*

This research used peer-reviewed journals found in all major databases such as Scopus, Web of Science, and IEEE Xplore so as to provide (a) a methodological framework and (b) a base for academic credibility. There were also references to other types of literature, such as the reports of (i) government policy and (ii) international organizations, such as the World Bank and OECD, to establish the macro context of the labour market and governance. The sources of literature referenced were: systematic reviews, empirical experimental studies, frameworks for mitigating bias and socio-technical analyses of AI. Case evidence is based only on research results documented in the literature, not on anecdotal accounts. By triangulating the literature and institutional reports, the ability to test or cross-verify the structural bias mechanisms across both technological and policy spheres can be strengthened through the reliability of the thematic conclusions.

### *C. Case Study 1: AI-Driven Hiring Algorithms in STEM*

This first case study looks at the impacts of predictive hiring algorithms in STEM-related recruitment systems while assessing whether or not these systems systematically undervalue female candidates. To support this case study, studies on AI recruitment bias (Kordzadeh & Ghasemaghaei, 2021), fairness mitigation methods (Siddique et al., 2024), and human trust experiments in algorithmic decision-making processes (Islam et al., 2022) were referenced. Structural descriptions of how gender discrepancies exist in the STEM workforce are provided by Wang and Degol (2017), and Ceci et al. (2023), who discussed how context/culture impact performance and participation disparities. The literature indicates that factors, such as career breaks, nonlinear trajectories, and traditionally male-based performance datasets, often play a role in how scoring systems are created through algorithms. Additionally, research indicated that even if gender was removed from scoring variables, then the use of proxy features and prior historically biased data could still be created with equality issues. Lastly, empirical evidence has suggested that humans occasionally prefer algorithmic outputs that correlate with their historical biases, regardless of whether or not a fairness adjustment occurs.

**(RQ1): How do predictive hiring algorithms undervalue female candidates in STEM fields, particularly when accounting for career breaks and nonlinear career trajectories?**

This case study forms part of the broader literature on algorithmic hiring bias, socio-technical AI governance, and structural gender inequality in professional STEM pathways.

### *D. Case Study 2: Facial Recognition Systems and Electrical Engineering Bias*

This is a second case study about facial recognition systems which are based on electrical engineering such as signal processing, pattern matching and feature extraction. The case discusses aspects of social and technical considerations about AI bias (Hall and Ellis 2023), global labor impact report (OECD 2022), and discussions from Cambridge Handbook of Facial Recognition in the Contemporary State. Documented studies have shown that there are a higher number of errors with facial recognition systems when attempting to identify female individuals, but especially for females that have dark skin tones. This inequality is caused by the lack of information in training datasets and optimizations that are conducted to those datasets that give preference to the majority individuals; therefore, the creation of these biased training datasets and optimizations arise from the way that data collects, the calibration processes for extracting feature points from captured images, and the ways the systems are deployed such as in surveillance and identity verification systems. The case demonstrates how bias in engineering design workflows can exist and be compounded when the systems are deployed as institutions.

**(RQ2): To what extent do facial recognition systems, reliant on electrical engineering signal processing and pattern recognition, exhibit higher error rates for women, thereby amplifying structural and professional risks?**

This case forms part of the broader discourse on algorithmic discrimination, electrical engineering design bias, and AI governance in public and professional environments.

*E. Case Study 3: Adversarial Debiasing and Fairness Interventions*

Algorithmic fairness interventions are evaluated in the third case study, which focuses on adversarial debiasing, pre-process data balancing, and post-process parity adjustments. Findings from bias mitigation survey literature (Siddique et al., 2024), fairness-aware modeling research literature (Feldman & Peake, 2021), and human-algorithm interaction literature (Islam et al., 2022) are synthesized. Literature indicates that, while adversarial and fairness aware technologies can provide reductions in measurable bias metrics, the effectiveness of those systems is also impacted significantly by social acceptability, institutional readiness, and organizational incentives. There is also some indication from the literature that debiasing techniques may actually reduce predictive performance under certain conditions, leading to a trade-off between fairness and accuracy. Therefore, both computational effectiveness and governance feasibility will be evaluated in this case.

**(RQ3): Can adversarial debiasing and fairness-aware modeling effectively reduce gender bias in AI-driven STEM assessments while preserving predictive accuracy and institutional viability?**

This case is situated within broader AI fairness, ethical AI governance, and socio-technical implementation literature.

*F. Analytical Procedure*

An analytical process uses a structured approach to thematic synthesis. The first step is to classify bias sources into three categories: (1) bias on the data level, (2) design bias in algorithmic processes and (3) bias in the deployment context. Then, using a multi-level gender framework, these bias mechanisms are mapped against various individual, institutional and structural determinants. In the third, cross-case comparisons of patterns of re-enforcement across cases of different hiring systems (i.e., facial recognition technologies or measures of fairness), systematic re-enforcement is evaluated. Finally, evaluation of mitigation strategies occurs based on technical efficacy, evidence of social acceptance and governance congruence. Using this multi-level and layered analytical process supports the integration of both engineering analysis of the data and institutional as well as feminist theoretical perspectives.

*G. Methodological Contribution*

This research contributes methodologically, by combining structural theories on gender with social-technical approaches to the governance of artificial intelligence (AI) and analysis of electrical engineering systems into a single qualitative framework. It connects studies on algorithms being fair and the disparity in the science, technology, engineering, and mathematics (STEM) workforce, and develops criteria for evaluating AI systems and algorithms beyond simply using statistical parity measures to include issues of institutional fit and social legitimacy. By synthesizing structural, institutional, and computational aspects of analysis, this study provides a complete framework for evaluating AI systems in gendered environments within STEM fields.

## IV. Results and Discussions

The study's findings indicate a strong correlation between gender bias in AI technologies used in STEM disciplines as a problem of technical failure (or lack of) but, rather, an example of a socially constructed, socio-technical phenomenon that is embedded within the data infrastructures and systems that develop and implement the technology. There are signs of this systematically occurring across three distinct types of AI systems; hiring algorithms, facial recognition technologies and fairness-based modelling frameworks. Each time we see a pattern of systematic encoding of historical gender disparities into these systems' data infrastructures, the engineering processes used to create these systems are driven optimally and therefore are legitimized through institutional use. We can conclude then that AI systems function primarily to reinforce existing social hierarchies rather than as impartial measures of performance.

Research about gender disparities in STEM participation show that they are the product of institutional norms that have developed over time; stereotype-based expectations for performance; and unequal child-rearing responsibilities (Wang & Degol, 2017; Ceci et al., 2023). These factors have an impact on education pipelines through career development and evaluation systems. When machine learning (ML) algorithms use data from historical performance in an environment like

this, ML algorithms will apply the same structural distortions in their predictive modelling. Predictive modelling builds on criteria from "merit" or "performance likelihood" through historical reference points based on male-dominated career paths, where men have had more continuous employment, and a very narrow definition of productivity. Therefore, the output produced by an ML algorithm will look statistically sound while reinforcing the historically contingent standards of success.

This pattern is an example of what is referred to as layered reinforcement in multi-level bias frameworks (Risberg et al., 2009). Individual bias (confidence-related gaps between genders; self-selection patterns) occur simultaneously with biased institutional criteria (evaluation processes; hiring norms) and structural bias (policy frameworks; gendered child-care expectations). ML's interface with the three levels of bias works by numerically assessing individual attributes; developing institutional metrics; and functioning within a larger social and economic context. The results of the analysis provide evidence that ML systems do not create new inequality, but instead create future inequity through computational distance and therefore exacerbate existing forms of inequities.

There is another dimension to the neutrality assumption: the engineering dimension. In machine learning and other technology disciplines, optimization processes tend toward maximizing overall aggregate accuracy and minimizing loss and maximizing performance. An example is in facial recognition technology or signal processing systems, which have been designed to optimize the overall classification of given data. However, by choosing to use aggregate optimization, harm done to sub-groups can be hidden or obscured. In addition, if a given dataset is imbalanced, the model architectures used in developing the model will create representational dominance that will generate systematically greater error rates in relation to under-represented groups than would otherwise exist. This situation demonstrates how technical rationality, or logic that focuses on efficiency and accuracy, can create or exacerbate distributive inequity. Thus, the issue is not just the existence of biased data, but also the interaction between data imbalances and the optimization of that data.

According to socio-technical governance research (Hall & Ellis, 2023), reasons such as the existence of power hierarchy among institutions through which Built (baked) in biases continue even with efforts to mitigate. Artificial Intelligence (AI) contains systemic and institutional forces/exercise of power that define how it will be designed, used and held accountable. The decision to remove systemically incorporated bias from AI via fairness versus accuracy and/or operational efficiency may be difficult for actors in organizations who opt for an alternative level of priority in these areas, since even when using techniques to debias models there is often perceived trade-offs due to oppositional forces or resources limits on accuracy/efficiency. When examined collectively, the above literature supports an ideology that fairness is a political commitment and not a strictly mathematical objective.

As a result of these findings, we conclude that purely technical concepts of algorithmic fairness are limited. While the mathematical redistribution of representation in model outputs is achievable through pre- processing, in-processing, and post-processing methods, these strategies do not address historical conditions that generated the training data or institutional logics defining merit. In the context of hiring, for example, debiasing a model does not change the definition of merit. Similarly, enhancing the accuracy of subgroups in a surveillance context does not eliminate the potential harm resulting from disproportionate monitoring. Therefore, using technical adjustments to correct for bias without scriptural reform undermines the potential for procedural fairness to yield true equity. This proof of concept is reinforced by feminist structural theories of inequality, which assert that professional norms, evaluation criteria and definitions of competence are constructed through historically gender-frames institutions (see Allegrini, 2014). The way we construct computational metrics from those norms is to take previously discretionary judgements and embed them into algorithmic systems. While the use of automation eliminates explicit bias from decision-making processes, it has the potential to shield the effects of structural inequity via mathematical abstraction. Because algorithms are perceived as "objective" evaluations, institutions can use their result to demonstrate fairness and equity even when there is a significant imbalance between the foundation on which the algorithm was built and the result of their use.

## V. Conclusion and Future Scope

The conclusion of this study is that gender inequality in STEM due to AI systems is a combination of socio-technology and embedded structure rather than just a flaw in computation. AI technologies replicate the historical labour market inequities, encode these inequities into the datasets on which they function, and implement these inequities using engineering approaches that are driven by the goals of optimising efficiency rather than equity. Fairness-seeking modelling approaches and adversarial debiasing approaches demonstrate statistically significant improvement in fair outcomes but do not

individually address the influences of institutional norms, standards on evaluation, and structural constraints upon the level of gender inequity in STEM. Thus, it is clear that in order to understand how to achieve sustainable algorithmic fairness there must be a combination of responsive design/category, institutional accountability methods, and structural reform. AI systems should be assessed not only for their predictive accuracy but also for their ability to address the historical inequities that exist in the ecosystems upon which their data are sourced.

Future research into fairness interventions requires longitudinal and interdisciplinary approaches in engineering validation, governance analysis, and labor market studies, as well as the need for performance auditing frameworks that are disaggregated by sub-group, inclusive protocols for constructing datasets, and institutional models that align technical fairness objectives with decision-making practices within organizations. An empirical field study that uses participatory AI design, i.e., developing systems with input from a diverse set of stakeholders, may lead to new methods for deploying AI systems in a more equitable manner. Additionally, examining the intersectionality of race, class, and geography will provide greater analytical rigor than examining only two genders. As such, future research will need to shift from reactive to proactive structural inclusion and ensure that consideration of equity is embedded early in the design, validation, and governance processes of STEM AI systems.

**References**

1. Allegrini, A. (2014). Gender, STEM studies and educational choices: Insights from feminist perspectives. In Understanding student participation and choice in science and technology education (pp. 43–59). Springer. https://doi.org/10.1007/978-94-007-7793-4_4
2. Andrews, M. E., & Boklage, A. (2023). Supporting inclusivity in STEM makerspaces through critical theory: A systematic review. Journal of Engineering Education. https://doi.org/10.1002/jee.20546
3. Balducci, M. (2023). Linking gender differences with gender equality: A systematic-narrative literature review of basic skills and personality. Frontiers in Psychology, 14, 1105234. https://doi.org/10.3389/fpsyg.2023.1105234
4. Bartl, M., Mandal, A., Leavy, S., & Little, S. (2024). Gender bias in natural language processing and computer vision: A comparative survey. ACM Computing Surveys. https://doi.org/10.1145/3700438
5. Besse, P., del Barrio, E., Gordaliza, P., Loubes, J.-M., & Risser, L. (2020). A survey of bias in machine learning through the prism of statistical parity for the adult data set. arXiv Preprint. https://doi.org/10.48550/arXiv.2003.14263
6. Botella, C., Rueda, S., López-Iñesta, E., & Marzal, P. (2019). Gender diversity in STEM disciplines: A multiple factor problem. Entropy, 21(1), 30. https://doi.org/10.3390/e21010030
7. Breda, T., Jouini, E., Napp, C., & Thebault, G. (2020). Do female role models reduce the gender gap in science? Evidence from French high schools. IZA Discussion Paper No. 13163. https://www.iza.org/publications/dp/13163
8. Ceci, S. J., Kahn, S., & Williams, W. M. (2023). Exploring gender bias in six key domains of academic science: An adversarial collaboration. Psychological Science in the Public Interest, 24(1), 15–73. https://doi.org/10.1177/15291006231163179
9. Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Santuccione Chadha, A., & Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. npj Digital Medicine, 3, Article 81. https://doi.org/10.1038/s41746-020-0288-5
10. Duan, W., Li, L., Freeman, G., & McNeese, N. (2025). A scoping review of gender stereotypes in artificial intelligence. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (pp. 1–20). ACM. https://doi.org/10.1145/3706598.3713093
11. Feldman, T., & Peake, A. (2021). End-to-end bias mitigation: Removing gender bias in deep learning. arXiv Preprint. https://doi.org/10.48550/arXiv.2104.02532
12. Gómez-Talal, I., Bote-Curiel, L., & Rojo-Álvarez, J. L. (2025). Interpretable machine learning models for PISA results in mathematics. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3538585
13. Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2016). Predicting students performance in educational data mining.

IEEE. https://ieeexplore.ieee.org/document/7439649

14. Hall, P., & Ellis, D. (2023). A systematic review of socio-technical gender bias in AI algorithms. Online Information Review, 47(7), 1264–1279. https://doi.org/10.1108/OIR-08-2021-0452
15. Ibarra-Vazquez, G., Ramírez-Montoya, M. S., & Buenestado-Fernández, M. (2024). Forecasting gender in open education competencies: A machine learning approach. IEEE Transactions on Learning Technologies, 17.
16. Kong, S. M., Carroll, K. M., Lundberg, D. J., Omura, P., & Lepe, B. A. (2020). Reducing gender bias in STEM. MIT Science Policy Review. https://doi.org/10.38105/spr.11kp6lqr0a
17. Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: Review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212
18. Kuppler, M. (2022). Predicting the future impact of computer science researchers: Is there a gender bias? Scientometrics, 127(11), 6695–6732. https://doi.org/10.1007/s11192-022-04337-2
19. Li, L., Srivastava, N., Rong, J., Guan, Q., Gašević, D., & Chen, G. (2025). When and how biases seep in: Enhancing debiasing approaches for fair educational predictive analytics. British Journal of Educational Technology. https://doi.org/10.1111/bjet.13575
20. Manresa-Yee, C., & Ramis, S. (2022). Assessing gender bias in predictive algorithms using explainable AI. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 74, 383–406. https://doi.org/10.1007/s11577-022-00839-2
21. Mulvey, K. L., Hoffman, A. J., & McGuire, L. (2022). Fairness and opportunity in STEM contexts. In Routledge handbook chapter (pp. 236–252). Routledge. https://doi.org/10.4324/9781003047247-19
22. O'Connor, S., & Liu, H. (2023). Gender bias perpetuation and mitigation in AI technologies: Challenges and opportunities. AI & Society, 39(4), 2045–2057. https://doi.org/10.1007/s00146-023-01675-4
23. Patterson, L., Varadarajan, D. S., & Salim, B. S. (2020). Women in STEM/SET: Gender gap research review of the United Arab Emirates (UAE)—A meta-analysis. Gender in Management: An International Journal, 36(8), 881–911. https://doi.org/10.1108/gm-11-2019-0201
24. Risberg, G., Johansson, E. E., & Hamberg, K. (2009). A theoretical model for analysing gender bias in medicine. International Journal for Equity in Health, 8(1), 28. https://doi.org/10.1186/1475-9276-8-28
25. 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. https://doi.org/10.3390/digital4010001
26. Stewart, J., Henderson, R., Michaluk, L., Deshler, J., Fuller, E., & Rambo-Hernandez, K. (2020). Using the social cognitive theory framework to chart gender differences in the developmental trajectory of STEM self-efficacy in science and engineering students. Journal of Science Education and Technology, 29(6), 758–773. https://doi.org/10.1007/s10956-020-09853-5
27. Tandrayen-Ragoobur, V., & Gokulsing, D. (2022). Gender gap in STEM education and career choices: What matters? Journal of Applied Research in Higher Education, 14(3), 1021–1040. https://doi.org/10.1108/JARHE-09-2019-0235
28. Wang, M.-T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. Educational Psychology Review, 29, 119–140. https://doi.org/10.1007/s10648-015-9355-x
29. Xu, J., Moon, K. H., & van der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. IEEE. https://ieeexplore.ieee.org/document/7894238
30. Zanellati, A., Zingaro, S. P., & Gabbrielli, M. (2024). Balancing performance and explainability in academic dropout prediction. IEEE Transactions on Learning Technologies, 17. https://ieeexplore.ieee.org/document/10612222