Bias and Unfairness in Machine Learning Models: A Systematic Literature Review

Authors: Tiago Palma Pagano, Rafael Bessa Loureiro, Fernanda Vitória Nascimento Lisboa, Gustavo Oliveira Ramos Cruz, Rodrigo Matos Peixoto, Guilherme Aragão de Sousa Guimarães, Lucas Lisboa dos Santos, Maira Matos Araujo, Marco Cruz, Ewerton Lopes Silva de Oliveira, Ingrid Winkler, Erick Giovani Sperandio Nascimento

1. **Summary of the Introduction**

The introduction highlights that machine learning (ML) systems are increasingly used in important decision-making areas (like lending, criminal justice, hiring, public health). However, there is a critical challenge: these systems can inherit and produce unfair and biased outcomes. These outcomes can systematically disadvantage individuals or groups based on sensitive attributes such as race, gender, class, etc. Given their real-world impact, bias and unfairness in ML models raise ethical, legal, and social concerns.

The review notes that while numerous techniques and tools exist to detect and mitigate bias (e.g., AIF360, FairLearn, Tensorflow Responsible AI, Aequitas), there is no consensus on standardized metrics or methods. This makes selecting appropriate fairness measurements and mitigation strategies difficult for practitioners. The review aims to consolidate current knowledge and highlight these issues systematically.

2. **Problem Statement**

The core problem is that although machine learning is widely adopted in critical domains, there is still no clear, standardized approach to identify and mitigate bias and unfairness in ML models. Existing detection and mitigation methods are scattered across different contexts, and developers must often choose among a large and inconsistent set of metrics and tools without guidance. This lack of standardization complicates reliable evaluation of fairness and leads to inconsistent or inappropriate fairness practices in ML applications

3. **Objectives of the Paper**

I. The paper's main objectives are:

II. To systematically review existing research on bias and unfairness in machine learning models.

III. To identify and summarize the main datasets, fairness metrics, tools, and mitigation methods used in ML fairness research.

IV. To analyze the current state of the art, including strengths and limitations of existing approaches.

V. To highlight challenges and opportunities for future research in bias and unfairness mitigation

4. **Summary of the Problem Statement (restated)**

ML systems frequently produce biased or unfair decisions due to underlying training data, model designs, or deployment contexts. Research has proposed many fairness metrics and mitigation techniques, but no unified standard exists. Developers often lack clear guidance on how to choose or apply these methods, resulting in inconsistent fairness handling — especially in ethical, legal, and societal contexts where fairness is crucial

5. **Research Gap Identified**

The paper identifies several gaps in the literature:

Lack of standardized definitions and practices for fairness and bias mitigation in ML — existing work often uses different criteria and metrics without consensus.

Tools and techniques tend to be use-case specific, rather than broadly applicable across various ML domains.

Developers often lack guidance, as current solutions require deep expertise in fairness and related metrics.

Need for clearer guidance on selecting appropriate fairness metrics for different contexts and applications, since many metrics exist with varied implications.