International Journal of Research in All Subject
www.multiarticlejournal.com
ISSN: 3107-5878
Received: 13-02-2026, Accepted: 16-03-2026, Published: 31-03-2026
Volume 2, Issue 1, 2026, Page No. 174-176

Bias and fairness in machine learning models: Detection and mitigation

Dr. Nikita Bahaley[1], Prathamesh Mane[2], Atharva Mhatre[2], Ashwin John[2]
[1] Department of Information Technology, Pillai College of Arts, Commerce and Science (Empowered Autonomous), New Panvel, Maharashtra, India
[2] Pillai College of Arts, Commerce and Science (Empowered Autonomous), New Panvel, Maharashtra, India

Abstract
In recent years, intelligent computing systems have become deeply integrated into everyday decision-making environments. These systems depend largely on previously collected data, which may contain hidden imbalances. When such patterns are learned, the system may produce outcomes that unintentionally favor or disadvantage certain groups. This raise concerns related to fairness and responsible use of technology. The present study explores how bias appears in machine learning-based systems and why it is important to address it. A structured survey involving 50 final-year Information Technology students was conducted to understand their awareness, opinions, and level of familiarity with fairness-related concepts. The responses were collected using a scaled questionnaire to capture different levels of understanding. The study highlights that improving fairness in intelligent systems requires both technical knowledge and awareness. Strengthening education in this area can help future developers build systems that are more balanced, transparent, and socially responsible.

Keywords: Machine learning, algorithmic bias, fairness, ethical AI, bias mitigation, data ethics

Introduction

In modern digital systems, automated decision-making tools are widely used to assist human activities. These tools function by identifying patterns from previously available information and using those patterns to generate predictions. However, if the underlying data contains uneven representation or hidden preferences, the resulting outputs may not be balanced. Bias in such systems does not always appear intentionally. It often develops due to historical patterns, incomplete datasets, or uneven sampling. When these systems are applied in areas like recruitment, banking, or healthcare, even small imbalances can influence important outcomes. Because of this, concerns related to fairness have gained attention in recent years. Fairness in intelligent systems focuses on ensuring that outputs remain consistent and do not negatively affect any specific group. Achieving this requires careful examination of how data is prepared, how models are trained, and how results are evaluated. Various techniques are used to improve fairness, including adjusting datasets, modifying learning processes, and reviewing final outputs.

To understand how well students are aware of these issues, a survey-based approach was used in this study. The aim was to examine their understanding of bias, their perception of its impact, and their views on the importance of fairness in future system development.

Literature Review

Earlier work in this field mainly concentrated on improving the performance of computational models, particularly in terms of prediction accuracy and processing speed. Over time, researchers began to notice that these systems could reflect patterns that were not always fair or balanced. This led to increased attention toward ethical concerns in data-driven technologies. Studies have shown that bias can enter systems through various stages, including data collection and feature selection. When datasets do not properly represent all groups, the resulting models tend to perform unevenly. This issue becomes more visible in applications where decisions directly affect individuals.

To study fairness, researchers introduced several evaluation methods that help compare outcomes across different groups. These methods are useful for identifying whether a system behaves consistently or shows variation in performance. Efforts to reduce such issues have resulted in different corrective approaches. Some methods focus on improving the quality of input data, while others adjust the internal learning process. Additional techniques are applied after model development to refine the outputs and reduce imbalance.

Recent discussions also emphasize the importance of interpretability. When system decisions can be understood clearly, it becomes easier to identify unexpected behavior. Researchers also highlight that fairness may vary depending on the application, which makes it necessary to choose appropriate evaluation methods based on context. Educational research further suggests that students should be exposed to these topics early in their learning process. Awareness of fairness and responsible system design can help future professionals develop more balanced and accountable technologies.

Purpose of Study

This study was conducted to understand how final-year IT students perceive issues related to bias and fairness in machine learning systems.

The objectives include

- Examining awareness about biased patterns in datasets
- Understanding perception of real-world impact
- Evaluating familiarity with fairness-related techniques
- Identifying the importance of corrective measures
- Highlighting the need for ethical learning in academics

Testing & Research Questions

This section focuses on evaluating how students interpret and understand fairness-related challenges in intelligent systems.

Research Questions

1. How aware are students about the presence of bias in training data?
2. Do students relate biased outputs to real-world consequences?
3. What level of familiarity exists regarding fairness evaluation methods?
4. How important do students consider bias correction techniques?
5. Do students support including ethical AI topics in their curriculum?

Testing Objectives

Based on the above research questions, the study aims to test the following:

- The level of conceptual awareness regarding bias in machine learning datasets.
- Students' understanding of ethical implications arising from biased algorithmic decisions.
- Familiarity with quantitative fairness metrics and evaluation approaches.
- Perceived necessity of bias mitigation strategies during model development and deployment.
- Academic acceptance of integrating ethical AI and fairness-focused learning into higher education.

Data Collection

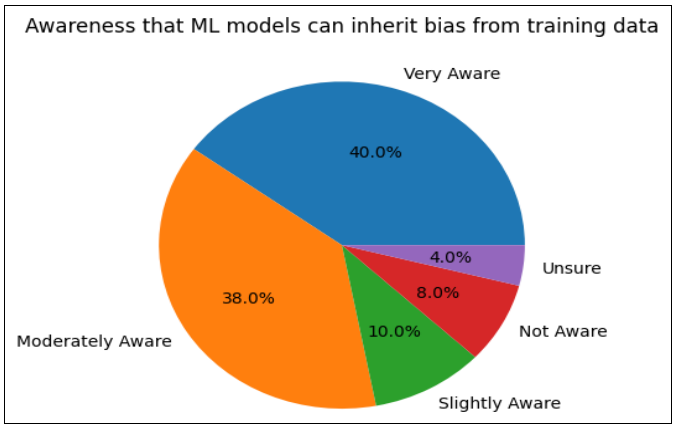


Fig 1: Awareness of Bias in Data

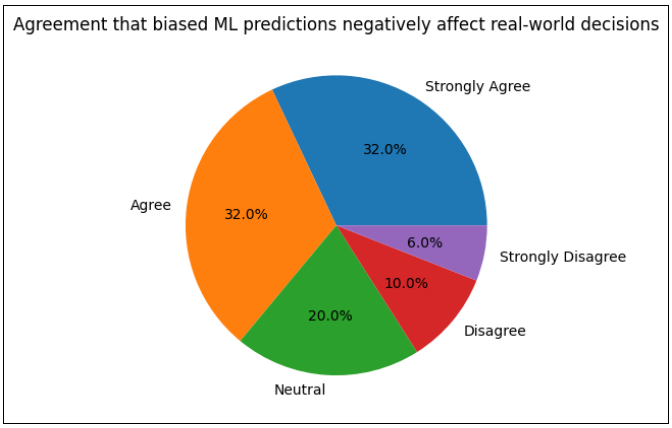


Fig 2: Understanding of Real-World Impact

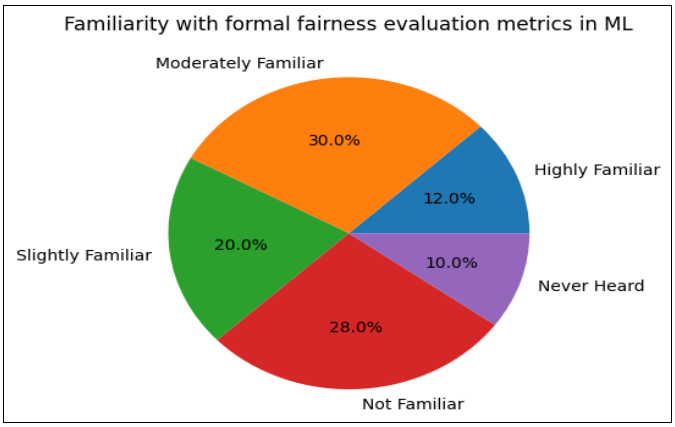


Fig 3: Familiarity with Evaluation Methods

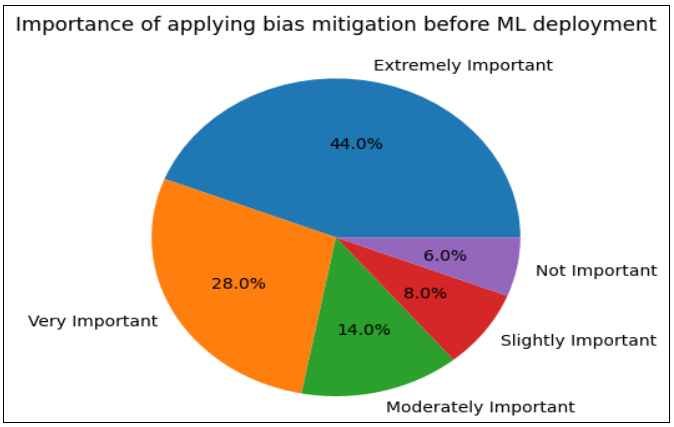


Fig 4: Importance of Bias Correction

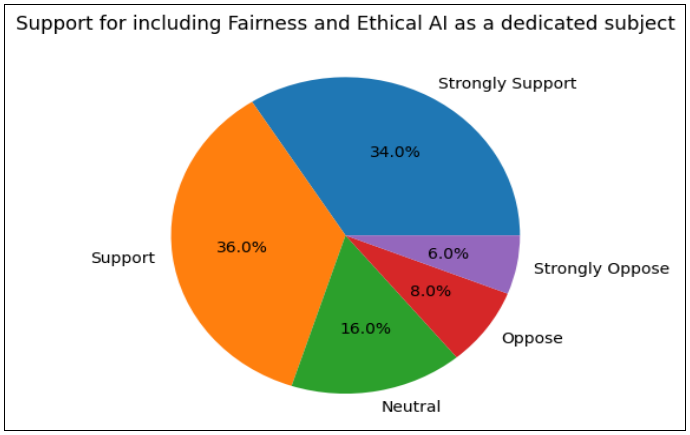
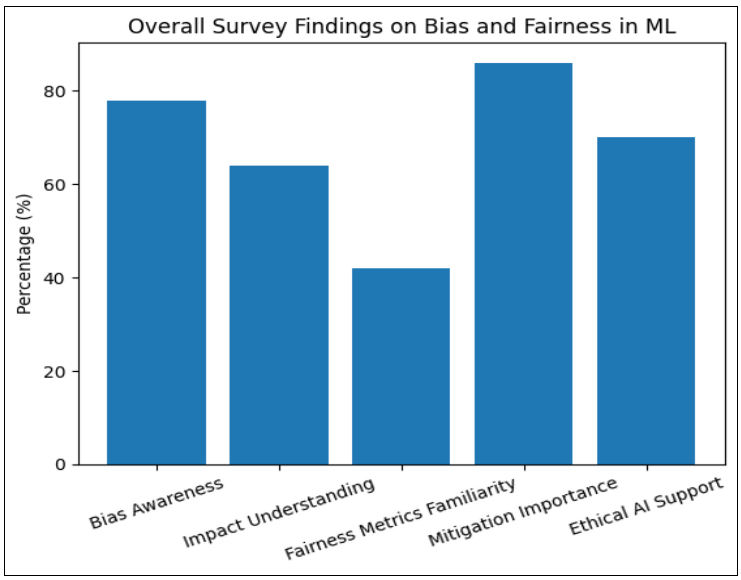


Fig 5: Support for Ethical AI Education

Data Interpretation & Findings



The collected responses indicate that a large number of participants already understand that data-driven systems may reflect uneven patterns. Many students were able to connect this concept with practical situations where decisions are influenced by automated systems.

It was also observed that participants generally agree that such issues can affect real-life processes. However, when asked about formal techniques used to measure fairness, responses showed lower confidence levels.

Most respondents strongly supported the idea that corrective steps should be taken before systems are used in real-world environments. There was also strong agreement regarding the inclusion of fairness-related topics in academic learning.

Overall, the results suggest that awareness exists at a conceptual level, but deeper technical understanding is still developing.

Conclusion

The study explored how students perceive bias and fairness in machine learning systems through a structured survey. The observations show that students are aware of the presence of imbalance in data-driven systems and recognize its practical impact. At the same time, familiarity with technical evaluation methods remains limited. Participants expressed strong support for improving fairness and including ethical topics in academic programs. The study concludes that increasing awareness along with practical exposure can help future professionals design systems that are more balanced and reliable.

Limitations of Study

This study was conducted using responses from 50 students within a similar academic background, which may limit diversity in perspectives. The results are based on self-reported understanding, which may vary among individuals. Future work can involve larger groups and practical experiments for deeper analysis.

References

1. Bogduk N, Bartsch T, Silberstein S, Lipton R, Dodick D. Clinical Evaluation of Cervicogenic Headache: A clinical perspective. International Journal of Multidisciplinary Research and Development,2008:16(2):73-80.
2. Barocas S, Hardt M, Narayanan A. Fairness and Machine Learning: Limitations and Opportunities. MIT Press Journal,2019:1(1):1-35.
3. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys,2021:54(6):1-35.
4. Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning Fair Representations. International Conference on Machine Learning,2013:32(1):325-333.
5. Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. Advances in Neural Information Processing Systems,2016:29(1):3315-3323.
6. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through Awareness. Innovations in Theoretical Computer Science Conference,2012:3(1):214-226.
7. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L. Model Cards for Model Reporting. Conference on Fairness, Accountability and Transparency,2019:1(1):220-229.
8. Binns R. Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of Machine Learning Research,2018:81(1):1-11.