International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 07 | July - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Bias Detection and Mitigation in AI: A Framework for fair and inclusive Machine Learning Models

Sweta Sucharita
Email ID: swetas2023@gift.edu.in

Prof. Kamalakanta Shaw
Email ID: kkshaw@gift.edu.in

Abstract- As AI becomes increasingly integrated into sectors like healthcare, finance, and recruitment, concerns around algorithmic bias, fairness, and data privacy are rising. This study addresses these ethical issues by introducing the Bias Detection Dashboard—a Python-based tool built with Pandas, Seaborn, Matplotlib, Fairlearn, and Streamlit. The dashboard helps users identify data imbalances, compute fairness metrics like disparate impact, and detect sensitive attributes before model training. A case study in HR recruitment shows how it flagged gender bias and guided corrective steps. The research also introduces the Ethical AI Triangle—bias, fairness, and privacy—as key pillars that must be balanced. It calls for ethical practices across the AI lifecycle, diversity in teams, and transparent systems to build responsible, trustworthy AI.

Keywords- Artificial Intelligence, algorithmic bias, data privacy, Bias Detection Dashboard, fairness metrics, gender bias, Ethical AI Triangle, responsible AI.

I. INTRODUCTION

Artificial Intelligence (AI) has evolved from a futuristic concept into a critical tool in solving real-world problems across industries. From predicting patient outcomes in healthcare, detecting fraud in banking, to assisting recruitment in HR, AI is now embedded in countless daily decisions. Yet, this integration comes with new responsibilities. When AI models are trained on real-world data, they inevitably absorb patterns including biased ones, which may lead to harmful or unfair predictions.

Bias in AI can take many forms. For instance, if historical recruitment data shows a preference toward male candidates, an AI model trained on this data may continue to favor men over equally qualified women. Similarly, loan prediction models may unknowingly discriminate against people from certain regions or economic backgrounds. These biases, even when unintentional, have far-reaching consequences reinforcing inequality and undermining trust in technology.

Today's world is witnessing a rising concern about three core challenges in AI:

1. **Bias** – When AI treats individuals unfairly because of their gender, skin color, language, or background.
2. **Privacy** – When AI systems collect or use people's personal data without consent or proper protection.
3. **Fairness** – When AI creates advantages for one group while ignoring or harming another.

Many datasets contain Personally Identifiable Information (PII) like names, email IDs, phone numbers, or addresses. If not properly handled, AI systems can inadvertently expose or misuse this data, violating privacy laws such as GDPR or India's Digital Personal Data Protection Act.

To tackle these challenges, this research introduces the Bias Detection Dashboard—a user-friendly, web-based tool designed to assess datasets before they are used for AI training. It focuses on three major areas: bias detection, fairness evaluation, and privacy protection. The dashboard scans data for representation gaps calculates fairness metrics like Disparate Impact, and flags PII to ensure responsible data use.

This approach empowers developers to correct data issues proactively rather than reacting after a biased model is deployed. The dashboard promotes a new standard where fairness and transparency are considered from the very beginning of the AI pipeline—not as an afterthought but as a core requirement.

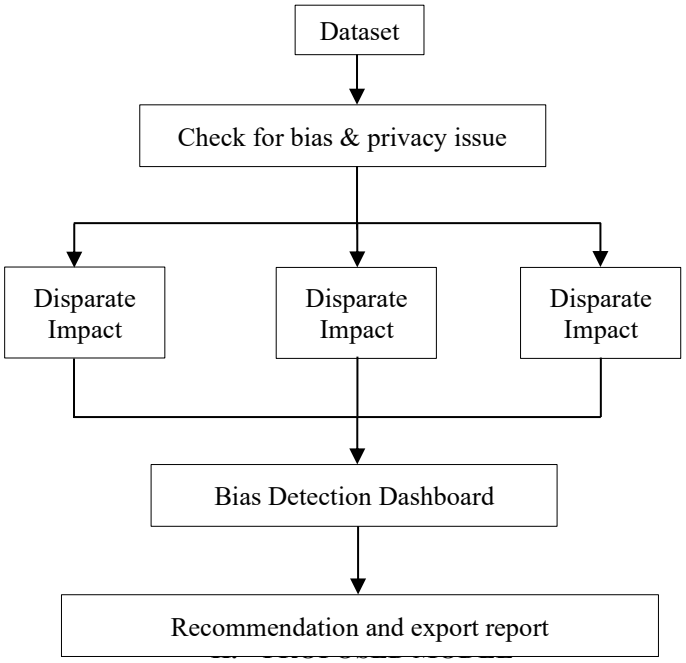


Fig 1: Structure ensuring foundation for AI systems.

II. LITERATURE SURVEY

Barocas, S., Hardt, M., & Narayanan, A. (2019). [1] Fairness and Machine Learning This foundational textbook lays out the theoretical groundwork for fairness in machine learning. It introduces concepts like statistical parity, equal opportunity, and

individual fairness. However, it is primarily theoretical and lacks a practical tool that practitioners can implement.

IBM AI Fairness 360 (AIF360) IBM's AIF360 toolkit [3] is one of the most comprehensive fairness libraries available. It includes bias detection, mitigation, and model evaluation techniques. Our solution is unique in that it simplifies these operations using an interactive interface, providing instant visual insights and exportable reports.

Raji, I. D., & Buolamwini, J. (2019). Actionable Auditing: [2] Investigating the Impact of Publicly Naming Biased Performance ResultsThis study highlights how public transparency and documentation can pressure organizations into ethical AI practices.

[6] Veale, M., & Binns, R. (2017). Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive DataThis paper argues for fairness approaches that work even when sensitive features like race or gender are unavailable. While our tool focuses on datasets where such attributes are identifiable, future versions can build upon these ideas by detecting proxy variables or applying fairness heuristics in the absence of direct labels.

[2] Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy This paper examines how different philosophical concepts of fairness apply to machine learning. It argues for context-specific fairness interventions, suggesting that developers should align fairness goals with social objectives.

[4] Sandvig, C., Hamilton, K., et al. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination in 'Black Box' Systems. The authors recommend pre-deployment audits, which our tool supports by focusing on dataset inspection before modeling begins. It aligns well with ethical frameworks for accountability.

III. PROPOSED MODEL

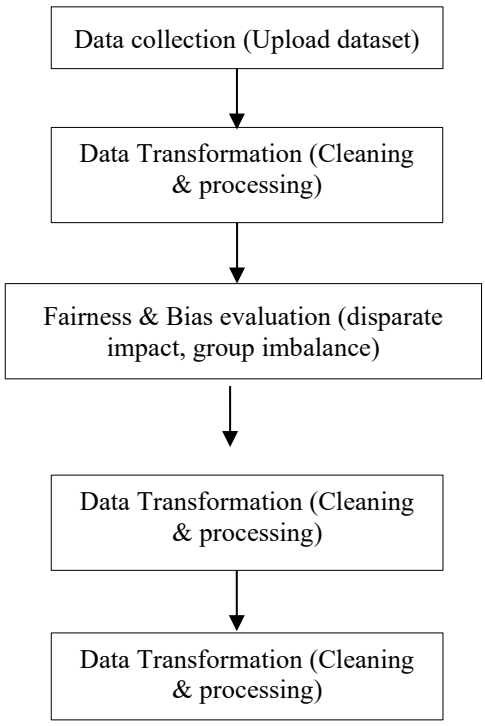


Figure 2: Proposed work

The proposed model aims to embed fairness, privacy, and ethical checks into the early stages of AI model development. It begins with data collection, where the user uploads a dataset that typically includes sensitive features like gender, age, or income. Once the data is loaded, a transformation process cleans the dataset by removing missing values, correcting inconsistent formats, and detecting outliers. This prepares the data for a robust fairness evaluation phase.

In the evaluation step, fairness metrics such as Disparate Impact and demographic parity are calculated to identify if certain groups are being unfairly treated. The model also performs visual analysis through charts and distributions to expose imbalances in data representation. Simultaneously, it performs privacy risk detection by flagging personally identifiable information (PII) like email IDs and phone numbers that could lead to data protection issues.

After these checks, the dashboard provides corrective suggestions and recommendations—such as data rebalancing or anonymization—along with a downloadable summary report. This modular and step-wise design ensures developers are informed of ethical gaps early, allowing for corrections before any machine learning model is trained. By offering a visual and metric-based overview, the system supports data scientists in building inclusive, legal, and socially responsible AI models. t also promotes transparency by enabling users to visualize how decisions are influenced by data quality and distribution. Additionally, the framework supports iterative updates, allowing developers to re-upload improved datasets and verify progress after each round of correction.

This model ensures that AI developers can make informed decisions regarding dataset fairness and legal compliance before proceeding to model training and deployment.

IV. METHODOLOGY

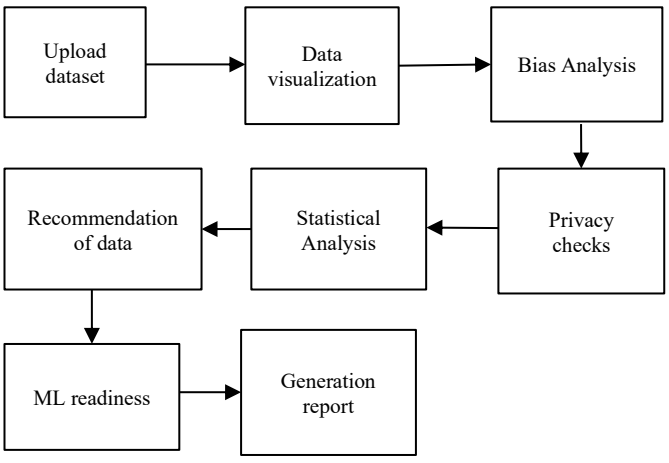


Figure 3: Design and Approach

The project plan is depicted in Figure 3, where the problem is

A. Problem Statement:

Artificial Intelligence systems are increasingly used in decision-making processes across sectors like recruitment, banking, and healthcare. However, when trained on biased or unbalanced datasets, these systems can perpetuate existing social inequalities, resulting in discriminatory outcomes.

Moreover, datasets often include sensitive personal information, leading to potential privacy violations. This project proposes a pre-model intervention framework that detects these issues early—

before any AI model is trained—ensuring responsible, fair, and ethical AI development.

B. Dataset Used

To demonstrate the capabilities of the proposed framework, publicly available HR datasets were used. These datasets contain applicant features such as age, gender, education level, total experience, and selection status.

These variables are ideal for assessing fairness, especially for evaluating whether certain groups—such as women or specific age brackets—are underrepresented or unfairly treated during the selection process.

C. Tools and Technologies

The system is built using:

- Python: programming language for backend logic
- Pandas: for dataset handling and manipulation
- Seaborn & Matplotlib: for data visualization
- Fairlearn: to compute fairness metrics
- Streamlit: for designing the dashboard interface

D. Step-by-Step Workflow Design

1. Data Upload & Extraction

The user uploads a CSV or Excel dataset through the dashboard interface.

2. Data Preprocessing

The system automatically cleans the dataset by handling missing values, eliminating irrelevant fields, and standardizing data formats. Sensitive features such as gender, race, and age are detected and flagged for fairness evaluation.

3. Visualization & Group Balance Check

Graphs such as histograms and box plots help users assess the distribution of values and detect any imbalance among different groups. This step is critical to understanding whether certain communities are underrepresented.

4. Fairness Metric Evaluation

Metrics like Disparate Impact are calculated to determine if the dataset favors one group over another. A score below 0.8, for example, may indicate unfair treatment. This allows developers to quantify the degree of bias present.

5. Privacy Risk Detection

Personally Identifiable Information (PII) such as names, email IDs, or contact numbers is automatically identified. These fields are flagged so that they can be removed or masked, thus ensuring the dataset adheres to privacy laws like GDPR.

6. Final Report Generation

A final downloadable report is generated which summarizes the findings—highlighting bias scores, flagged columns, imbalance trends, and mitigation suggestions. This makes it easier for teams to collaborate, document compliance, and iterate based on feedback.

This detailed methodology ensures a structured, scalable, and user-friendly process to examine fairness and privacy before model training even begins. It provides AI practitioners with a practical path to building more responsible and equitable AI systems. This methodology encourages transparency by involving users in every stage of dataset evaluation.

Each module in the dashboard is designed for real-time interaction, enabling users to test different datasets or apply mitigation strategies iteratively. One unique strength of the approach lies in its visual feedback every fairness score, imbalance, or privacy flag is shown with intuitive graphs and indicators.

This supports not only technical analysis but also better communication with non-technical stakeholders such as HR professionals or compliance teams. Moreover, the dashboard is modular, allowing new fairness metrics or PII checks to be added in future versions. It also aligns with ethical AI principles by promoting pre-emptive action over reactive fixes. With minimal coding skills required, the tool becomes accessible to researchers, analysts, and developers alike—bridging the gap between theory and practice in AI fairness.

All processes—from data upload to report generation—are automated to enable even non-technical users to assess data quality. Each module operates independently, allowing for iterative testing and improvement without restarting the workflow.

V. RESULTS

This photo displays the welcome screen of the Bias Detection Dashboard. On the left-hand side, we see a sidebar navigation panel that allows users to move through different stages of analysis. The dashboard is user-friendly and customizable—users can set a theme or appearance mode as per their preference. The first action is to upload a dataset, which is the entry point for initiating bias analysis.

In this section, users can upload the dataset for analysis. Once uploaded, the system generates an immediate preview of the data. As seen in the screenshot, the dataset includes attributes like email, phone number, age, gender, and more. The interface also displays:

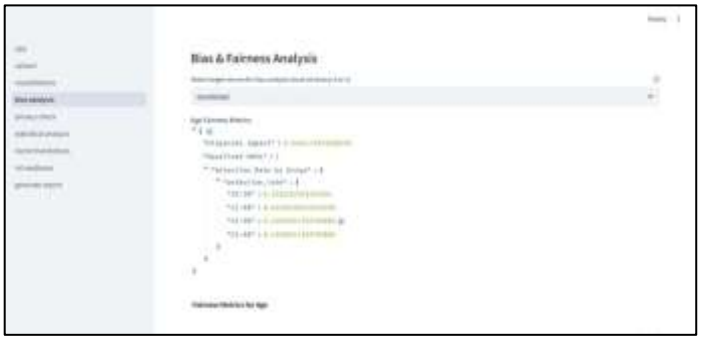
- The number of rows and columns
- A count of sensitive columns (such as gender or age)
- Which specific columns are classified as sensitive

This information is critical before running any bias or privacy analysis.

This screen allows users to visually explore the dataset using graphical tools. The visualization includes:

- Distributions
- Relationships between variables
- Box plots
- Data flow structures

In the current view, we see a histogram of the "Daily Rate" feature, which helps understand the spread and frequency of this variable across the dataset.



Users can select a binary target column (e.g., 0 or 1 for "shortlisted" or "not shortlisted") to perform fairness analysis. In the screenshot, fairness metrics related to age are being displayed. This step ensures that we identify any unethical or unintended bias present in the decision-making process.

This module is dedicated to privacy analysis. The system performs a scan to detect Personally Identifiable Information (PII).
The screenshot shows that columns like email and phone number are flagged as PII. It also educates users on:

- What constitutes PII
- Why it's important to detect and remove it

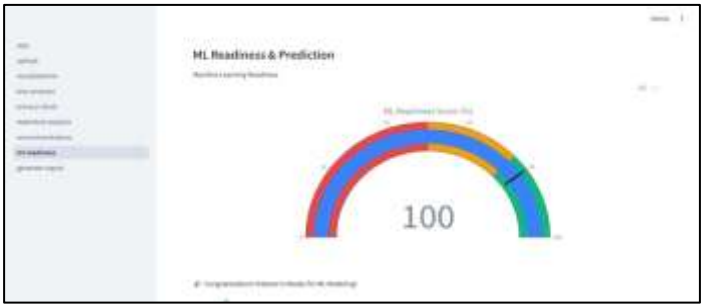
This step ensures compliance with privacy regulations such as GDPR or India's Digital Personal Data Protection Act.

This section displays a correlation heatmap, which is a powerful statistical tool to understand how different variables are related. In the screenshot, correlations between features like Daily Rate, Monthly Income, Years at Company, and Shortlisted status are visualized. The varying intensities of the heatmap cells indicate the strength and direction of the relationships. This step provides data scientists with insight into key influencing features.

This slide introduces the bias mitigation module. The dashboard provides:

- Recommendations based on the analysis
- Options to apply advanced bias mitigation techniques
- The ability to download the mitigated dataset

This ensures that users not only detect bias but can also take corrective action and use a cleaned dataset for ethical modeling.



The system calculates a readiness score (in percentage) out of 100. This score reflects:

- Completeness of the dataset
- Absence of bias and PII
- Data quality

This helps users know whether the dataset is suitable for model training and prediction tasks.

The final slide enables users to generate and download a PDF report of the entire analysis. Additionally, it provides the option to download the dataset of accepted candidates, if applicable.
This ensures documentation, transparency, and audit-readiness for internal reviews or regulatory submissions.

VI. CONCLUSION

The Development of a system that can reliably and efficiently Artificial Intelligence (AI) has become a powerful part of modern life. It is no longer limited to science labs or high-tech companies. Today, AI is present in everyday tools like voice assistants, recommendation systems, medical devices, and even in agriculture. This report has taken a deep look into the many aspects of AI—how it works, the challenges of making it understandable, and most importantly, the need to develop it responsibly. Throughout this report, we discussed the need for making AI models easier to understand. Techniques like model-specific interpretability, post-hoc explanations, and tools like LIME or SHAP allow developers to break down how AI models work. These techniques help users, stakeholders, and regulators see that decisions are not random or biased.
Being able to explain AI decisions is not just a technical requirement—it is a moral and social one. If people cannot trust how a decision was made, they are less likely to accept or support AI technologies. Therefore, making AI models interpretable increases public confidence and ensures that AI is used in a just and fair manner.

FUTURE SCOPE

As Artificial Intelligence (AI) continues to grow and integrate into more aspects of human life, the future scope of work in this field becomes broader, deeper, and more impactful. The focus is no longer just on making AI systems more accurate and powerful, but also on making them understandable, fair, ethical, and beneficial to all sections of society. This shift in priorities opens up a wide range of future possibilities and challenges that researchers, developers, policymakers, and society at large must collectively address.

REFERENCES

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). *Deep learning with differential privacy.* Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS), 308–318.

[2] Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities.* Retrieved from https://fairmlbook.org

[3] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy.
Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT), 149-159.

[4] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214-226.

[5] European Commission. (2021). *Proposal for a regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act).* Retrieved from https://digital-strategy.ec.europa.eu

[6] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 259–268.

[7] Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems (TOIS), 14(3), 330–347.

[8] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

[9] Google AI. (2018). Responsible AI Practices: Fairness, Interpretability & Privacy. Retrieved from https://ai.google/responsibility/

[10] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems (NeurIPS), 29, 3315-3323. 7.

[11] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? Proceedings of the 2019 CHI Conference.

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

[13] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229.
https://doi.org/10.1145/3287560.3287596

[14] Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES), 429–435. https://doi.org/10.1145/3306618.3314244