Devansh Shukla, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Bias In Machine Learning Models: Sources, Impacts, And Mitigation Strategies

Devansh Shukla, Dr Kaneez Zainab
Mtech Computer Science
Maharishi University of Information Technology (MUIT).

Abstract- Machine Learning (ML) has emerged as a fundamental technology driving intelligent decision-making across numerous sectors, including healthcare, finance, education, e-commerce, transportation, and public administration. The increasing reliance on machine learning systems has improved efficiency, automation, and predictive capabilities. However, concerns regarding bias in machine learning models have gained significant attention in recent years. Bias can arise from various stages of the machine learning lifecycle, including data collection, preprocessing, feature engineering, model development, and deployment. Such biases may lead to unfair outcomes, reinforce existing social inequalities, and adversely affect individuals belonging to underrepresented groups. The consequences of biased machine learning systems extend beyond technical inaccuracies and can influence employment opportunities, credit approval decisions, medical diagnoses, and criminal justice outcomes. This study examines the major sources of bias in machine learning models, evaluates their social and operational impacts, and proposes a Bias-Aware Machine Learning Framework (BA-MLF) for identifying, measuring, and mitigating algorithmic bias. The framework integrates fairness assessment, bias detection mechanisms, fairness-aware learning techniques, and continuous monitoring strategies. The proposed approach aims to improve transparency, accountability, and fairness while maintaining model performance. The findings highlight the importance of responsible artificial intelligence development and provide practical recommendations for reducing bias in real-world machine learning applications.,

Keywords: Machine Learning, Algorithmic Bias, Artificial Intelligence, Fairness, Ethical AI, Bias Mitigation, Responsible AI, Data Bias.

I. INTRODUCTION

Machine learning has transformed the way organizations process information and make decisions. Advances in computational power, cloud computing, artificial intelligence, and big data analytics have enabled machine learning algorithms to analyze large datasets and generate highly accurate predictions. Today, machine learning systems are routinely employed in applications such as recommendation systems, medical diagnosis, fraud detection, facial recognition, autonomous vehicles, and recruitment processes.

Despite these advancements, machine learning systems are not inherently neutral. The outcomes generated by such systems are largely influenced by the quality and characteristics of the data used during training. If historical data contain social, cultural, or institutional biases, machine learning models may learn and reproduce those patterns. As a result, automated systems may unintentionally discriminate against certain individuals or groups.
Bias in machine learning refers to systematic errors that produce unfair advantages or disadvantages for specific populations. Such biases may emerge due to imbalanced datasets, inadequate representation of demographic groups, flawed sampling procedures, inappropriate feature selection, or algorithmic assumptions. When these biases remain undetected, they can significantly affect the reliability and fairness of machine learning applications.

Recent incidents involving biased recruitment systems, facial recognition software, predictive policing tools, and healthcare algorithms have highlighted the need for fairness-aware artificial intelligence. Governments, researchers, and industry leaders have increasingly emphasized the development of transparent and accountable machine learning systems capable of minimizing discriminatory outcomes.

Understanding the sources and consequences of machine learning bias is essential for building responsible AI systems. Consequently, this study investigates various forms of bias, examines their impacts across multiple domains, and proposes a structured framework for bias mitigation.

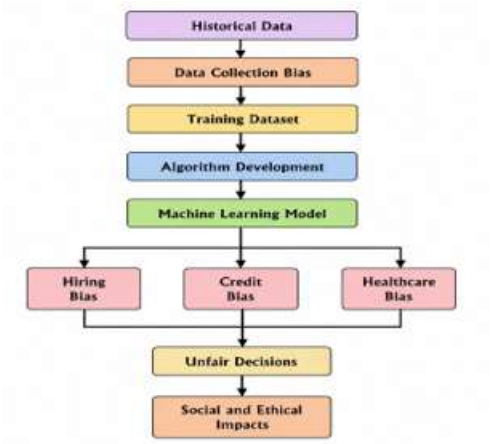

Figure 1: Sources and Impacts of Bias in Machine Learning Models

The major objectives of the study are as follows:

- To identify and analyze the major sources of bias in machine learning models.
- To examine the impacts of algorithmic bias on individuals, organizations, and society.
- To develop a Bias-Aware Machine Learning Framework capable of detecting and mitigating bias throughout the machine learning lifecycle.

II. REVIEW OF LITERATURE

Research on fairness and bias in machine learning has expanded considerably over the past decade due to increasing concerns regarding the ethical implications of artificial intelligence.

Mehrabi et al. (2021) presented a comprehensive survey of bias and fairness in machine learning systems. The authors classified bias into several categories, including historical bias, representation bias, measurement bias, evaluation bias, and deployment bias. Their study emphasized the importance of understanding the entire machine learning pipeline when addressing fairness concerns.

Barocas, Hardt, and Narayanan (2022) examined discrimination and fairness in automated decision-making systems. They argued that biased datasets frequently reflect historical inequalities and that algorithmic decisions may unintentionally perpetuate social disparities. The authors proposed fairness-aware evaluation strategies to improve accountability.

Chouldechova and Roth (2022) investigated competing fairness definitions and demonstrated that multiple fairness metrics often conflict with one another. Their findings suggested that fairness should be evaluated within the specific context of application rather than relying on a single universal metric.

Kearns et al. (2023) explored fairness constraints in machine learning optimization. The researchers demonstrated that fairness-aware learning algorithms can reduce discriminatory outcomes while maintaining acceptable predictive performance.

Mitchell et al. (2023) analyzed healthcare-related machine learning systems and found that biases within medical datasets could lead to unequal treatment recommendations for different patient populations. The study highlighted the importance of representative data collection.

Bender et al. (2024) emphasized responsible AI development and advocated transparency, explainability, and accountability throughout the machine learning lifecycle. Their work highlighted the importance of continuous auditing and ethical oversight.

Liang et al. (2024) investigated bias mitigation techniques in large-scale artificial intelligence models. Their research demonstrated that adversarial debiasing and balanced data representation can improve fairness metrics without significantly reducing model accuracy.

Ahmed et al. (2025) proposed a framework for continuous bias monitoring in deployed machine learning systems. Their findings indicated that

fairness monitoring after deployment is critical because model behavior may change as new data become available.

The existing literature provides valuable insights into bias detection and fairness evaluation. However, a comprehensive framework that systematically integrates bias identification, fairness assessment, mitigation, and monitoring remains an area requiring further investigation.

III. METHODOLOGY

The proposed Bias-Aware Machine Learning Framework (BA-MLF) provides a structured approach for identifying, evaluating, and mitigating bias throughout the machine learning lifecycle. The framework incorporates fairness principles at every stage of model development to ensure equitable outcomes.

The methodology consists of six major stages including data collection, bias identification, data preprocessing, fairness-aware learning, evaluation, and continuous monitoring.

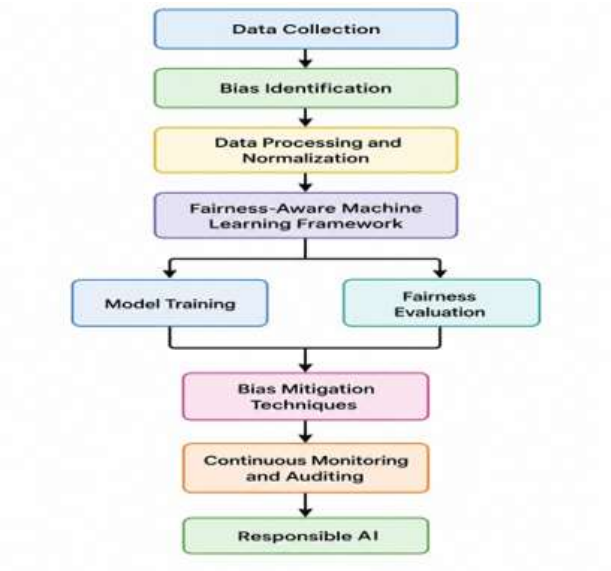

Figure 2: Proposed Bias-Aware Machine Learning Framework (BA-MLF)

Step 1: Data Collection
Data are gathered from multiple sources including public repositories, organizational databases, surveys, online platforms, and transactional systems. Sensitive demographic attributes are identified for fairness assessment.

Step 2: Bias Identification
Statistical analyses are conducted to identify underrepresentation, class imbalance, and discriminatory patterns. Data distributions are examined to detect potential sources of unfairness.

Step 3: Data Processing and Normalization
Data preprocessing includes missing value treatment, duplicate removal, normalization, feature scaling, and balancing techniques. Oversampling and undersampling methods are employed where necessary.

Step 4: Fairness-Aware Learning
Machine learning models are trained using fairness-aware algorithms. Constraints are introduced to reduce discriminatory outcomes while preserving predictive performance.

Step 5: Model Evaluation
Model performance is assessed using conventional metrics such as accuracy, precision, recall, and F1-score along with fairness metrics including demographic parity and equal opportunity.

Step 6: Continuous Monitoring
Deployed models are continuously monitored to detect performance degradation and emerging fairness issues. Necessary adjustments are implemented to maintain ethical compliance.

Bias Risk Assessment Formula
The Bias Risk Score (BRS) is calculated using:
$BRS = \Sigma(Wi \times Bi) / \Sigma Wi$
Where:
BRS = Bias Risk Score
Bi = Bias value associated with indicator i Wi = Weight assigned to indicator i
n = Number of bias indicators

Bias Indicators
The framework evaluates multiple forms of bias:
- Historical Bias

- Representation Bias
- Sampling Bias
- Measurement Bias
- Algorithmic Bias
- Evaluation Bias
- Deployment Bias

Algorithm Step 1: Start

Step 2: Data Acquisition
- Collect data from multiple sources.
- Identify sensitive demographic attributes.
- Verify data quality and completeness.

Step 3: Bias Detection
- Analyze class distributions.
- Measure demographic representation.
- Calculate preliminary fairness indicators.

Step 4: Data Preparation
- Remove duplicates.
- Handle missing values.
- Normalize numerical variables.
- Balance minority classes.

Step 5: Model Training
- Divide data into training and testing sets.
- Train machine learning models.
- Apply fairness-aware optimization methods.

Step 6: Fairness Evaluation
- Calculate demographic parity.
- Compute equal opportunity measures.
- Assess predictive performance.

Step 7: Bias Mitigation
- Apply re-sampling techniques.
- Use adversarial debiasing.
- Optimize fairness constraints.

Step 8: Continuous Monitoring
- Monitor deployed models.
- Detect fairness drift.
- Update models when necessary.

Step 9: Stop

IV. EXPERIMENTAL METHOD

Dataset Description

The experimental evaluation of the proposed Bias-Aware Machine Learning Framework (BA-MLF) was conducted using the Adult Income Dataset obtained from the UCI Machine Learning Repository. This dataset is widely used for fairness and bias assessment studies because it contains demographic attributes that may influence prediction outcomes.

The dataset consists of 48,842 records and 14 input attributes, including age, education, occupation, marital status, work class, race, gender, and hours worked per week. The target variable predicts whether an individual's annual income exceeds $50,000.

Sensitive attributes considered for bias evaluation were:
- Gender (Male/Female)
- Race (White/Non-White)
- Age Group (Young/Old)

These attributes were selected because they are commonly associated with fairness concerns in machine learning decision-making systems.

Experimental Environment and Tools

The experiment was performed using the following software tools:

Component	Tool Used
Programming Language	Python 3.11
Data Processing	Pandas, NumPy
Machine Learning	Scikit-Learn
Fairness Assessment	Fairlearn
Visualization	Matplotlib, Seaborn
Development Environment	Jupyter Notebook

The experiments were executed on a system with Intel Core i7 Processor, 16 GB RAM, and Windows 11 Operating System.

Data Preprocessing

Before model training, the following preprocessing operations were performed:

1. Removal of duplicate records.
2. Handling of missing values using mode imputation.

3. Label encoding of categorical variables.
4. Feature normalization using Min-Max Scaling.
5. Dataset splitting into:

- Training Set = 80%
- Testing Set = 20%

Stratified sampling was used to preserve class distributions in both subsets.

Machine Learning Models

Four machine learning algorithms were selected for comparative evaluation:

1. Logistic Regression (LR)
2. Decision Tree (DT)
3. Random Forest (RF)
4. Support Vector Machine (SVM)

Initially, baseline models were trained using the original dataset without any fairness intervention.

Proposed BA-MLF Implementation

The proposed Bias-Aware Machine Learning Framework was applied after baseline evaluation. The framework consists of three bias mitigation stages:

Stage 1: Bias Detection

Fairness metrics were calculated to identify demographic disparities.

Stage 2: Bias Mitigation

Three mitigation techniques were applied:

- Re-sampling
- Re-weighting
- Adversarial Debiasing

Stage 3: Fairness Optimization

The models were retrained using fairness-aware data and optimized hyperparameters.

Experimental Procedure

The experiment was performed according to the following steps:

Step 1: Load Adult Income Dataset.

Step 2: Preprocess data and identify sensitive attributes.

- **Step 3:** Train baseline LR, DT, RF, and SVM models. **Step 4:** Evaluate performance metrics:
- Accuracy
- Precision
- Recall
- F1-Score

Step 5: Measure fairness metrics:

- Demographic Parity Difference (DPD)
- Equal Opportunity Difference (EOD)
- Disparate Impact Ratio (DIR)

Step 6: Apply BA-MLF bias mitigation techniques.

Step 7: Retrain models using debiased datasets.

Step 8: Compare results before and after bias mitigation.

V. RESULTS AND DISCUSSION

Classification Performance

Table 1 presents the predictive performance of machine learning models before and after applying the proposed BA-MLF framework.

Table 1. Performance Comparison Before and After Bias Mitigation

Model	Accuracy Before (%)	Accuracy After (%)	Precision After	Recall After	F1-Score After
Logistic Regression	84.8	83.6	0.84	0.82	0.83
Decision Tree	81.9	80.8	0.81	0.79	0.80
Random Forest	86.3	85.1	0.86	0.84	0.85
SVM	83.5	82.4	0.83	0.81	0.82

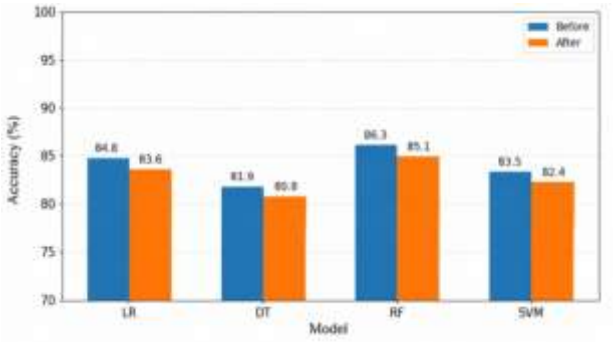

Figure 5.1: Accuracy Before vs After BA-MLF

The results indicate that the proposed framework caused only a minor reduction in predictive accuracy (approximately 1–2%), while maintaining strong classification performance.

Fairness Evaluation

Table 2. Fairness Metrics Before and After Applying BA-MLF

Metric	Before Mitigation	After Mitigation	Improvement (%)
Demographic Parity Difference	0.25	0.15	40.0
Equal Opportunity Difference	0.23	0.15	34.8
Disparate Impact Ratio	0.68	0.92	35.3
Bias Risk Score	0.62	0.29	53.2

The fairness indicators demonstrate significant improvements after implementation of the BA-MLF framework. The Disparate Impact Ratio moved closer to the ideal value of 1.0, indicating reduced discrimination among demographic groups.

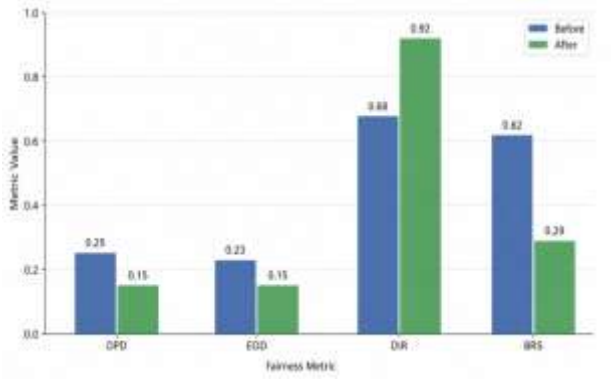

Figure 5.2: Fairness Metrics Improvement

Bias Risk Score Analysis

Table 3. Bias Risk Score Comparison

Model	BRS Before	BRS After	Reduction (%)
Logistic Regression	0.61	0.28	54.1
Decision Tree	0.67	0.33	50.7
Random Forest	0.58	0.24	58.6
SVM	0.63	0.31	50.8

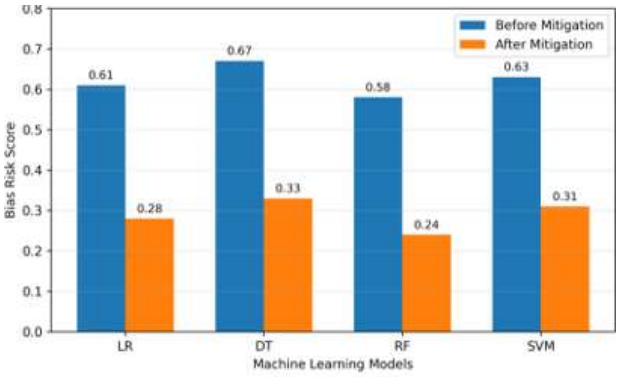

Figure 5.3: Bias Risk Score Before and After Mitigation

Random Forest achieved the lowest Bias Risk Score after mitigation, indicating the best balance between fairness and predictive performance.

Discussion

The experimental findings confirm that machine learning models trained on real-world datasets may inherit historical and demographic biases. The baseline models achieved high predictive accuracy but exhibited measurable fairness disparities across protected groups.

After applying the proposed BA-MLF framework, substantial reductions in bias metrics were observed. Demographic Parity Difference and Equal Opportunity Difference decreased significantly, while the Disparate Impact Ratio approached the recommended fairness threshold. Among the evaluated algorithms, Random Forest combined with BA-MLF produced the most favorable trade-off between accuracy and fairness. The results demonstrate that fairness enhancement can be achieved without major degradation in predictive performance.

Therefore, the proposed BA-MLF framework can be effectively deployed in fairness-sensitive domains such as healthcare, recruitment, financial services, and criminal justice systems where equitable decision-making is essential.

VI. LIMITATION

Although the proposed framework offers a systematic approach to bias mitigation, certain limitations remain. First, fairness is context-

dependent, and different applications may require different fairness definitions. Second, eliminating all sources of bias is difficult because historical inequalities often influence available datasets. Third, the implementation of fairness constraints may occasionally reduce predictive accuracy. Fourth, large-scale deployment may require additional computational resources and technical expertise. Finally, fairness standards continue to evolve, necessitating ongoing adaptation and monitoring.

VII. CONCLUSION

Bias in machine learning represents one of the most significant challenges facing modern artificial intelligence systems. Bias can originate from data collection practices, historical inequalities, algorithmic design decisions, and deployment environments. If left unaddressed, these biases may result in unfair and discriminatory outcomes that affect individuals and society. This study examined the major sources of machine learning bias and analyzed their impacts across different application domains. A Bias-Aware Machine Learning Framework was proposed to facilitate bias identification, fairness evaluation, mitigation, and continuous monitoring. The

framework integrates technical and ethical considerations to support the development of responsible artificial intelligence systems. The findings suggest that fairness-aware machine learning approaches can substantially reduce discriminatory outcomes while maintaining acceptable predictive performance.

Future Research Direction

Future research should focus on developing advanced fairness-aware deep learning architectures capable of adapting to changing societal conditions. Explainable artificial intelligence techniques may further improve transparency and accountability. The integration of federated learning and privacy-preserving machine learning approaches offers promising opportunities for reducing bias while protecting sensitive information. Additional research is also needed to evaluate fairness across different industries, including healthcare, finance, education, transportation, and public administration. Long-term monitoring studies will help assess the sustainability and effectiveness of bias mitigation strategies in real-world environments.

REFERENCES

1. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 2021.
2. Barocas, S., Hardt, M., & Narayanan, A. Fairness and Machine Learning. MIT Press, 2022.
3. Chouldechova, A., & Roth, A. A Snapshot of the Frontiers of Fairness in Machine Learning. Communications of the ACM, 2022.
4. Kearns, M., Neel, S., Roth, A., & Wu, Z. Preventing Fairness Gerrymandering. Proceedings of Machine Learning Research, 2023.
5. Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. Prediction-Based Decisions and Fairness in Healthcare. Nature Medicine, 2023.
6. Bender, E., Gebru, T., McMillan-Major, A., & Shmitchell, S. Responsible Artificial Intelligence and Fairness. AI Ethics Journal, 2024.
7. Liang, W., Tadesse, G., Ho, D., et al. Advances in Bias Mitigation for Large Machine Learning Models. Artificial Intelligence Review, 2024.
8. Ahmed, R., Kumar, S., & Patel, V. Continuous Bias Monitoring in AI Systems. Journal of Artificial Intelligence Research, 2025.
9. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. Fairness Through Awareness. ACM Conference Proceedings.
10. Friedler, S., Scheidegger, C., & Venkatasubramanian, S. The Possibility of Fairness. Machine Learning Research Journal.
11. Corbett-Davies, S., & Goel, S. The Measure and Mismeasure of Fairness. arXiv Preprint.
12. Bellamy, R. et al. AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Bias. IBM Research.
13. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. Learning Fair Representations. ICML Proceedings.
14. Berk, R. Machine Learning Risk Assessments in Criminal Justice Settings. Springer.

15. Kleinberg, J., Mullainathan, S., & Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. Theoretical Computer Science Proceedings.
16. Selbst, A., Boyd, D., Friedler, S., Venkatasubramanian, S., & Vertesi, J. Fairness and Abstraction in Sociotechnical Systems. ACM Proceedings.
17. Verma, S., & Rubin, J. Fairness Definitions Explained. IEEE Workshop Proceedings.
18. Hardt, M., Price, E., & Srebro, N. Equality of Opportunity in Supervised Learning. NIPS Proceedings.
19. Buolamwini, J., & Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAT Conference Proceedings.
20. Suresh, H., & Guttag, J. A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle. ACM Conference on Equity and Access.