Turkish Journal of Computer and Mathematics Education (TURCOMAT) ISSN: 3048-4855

Vol. 09 No. 2 (2018):753-787
DOI:https://doi.org/10.61841/turcomat.v10i1.14805753

Bias and Fairness in Machine Learning: A Systematic Review of Mitigation Techniques

Nischal Ravichandran, Anil Chowdary Inaganti, Senthil Kumar Sundaramurthy, Rajendra Muppalaneni,

1. Senior Identity Access Management Engineer, nischalravichandran@gmail.com
2. Workday Techno Functional Lead, anilchowdaryinaganti@gmail.com
3. AI/ML Architect, Cloud &Technical Leader, sundaramurthysenthilkumar2@gmail.com
4. Lead Software Developer, muppalanenirajendra@gmail.com

Abstract

Bias and fairness in machine learning (ML) algorithms are critical concerns that impact decision-making processes across various domains, including healthcare, finance, and criminal justice. This systematic review explores the state-of-the-art mitigation techniques employed to address bias and ensure fairness in ML systems. The review identifies and categorizes methods into pre-processing, in-processing, and post-processing strategies, while analyzing their effectiveness and limitations. Key findings indicate that although significant progress has been made, challenges remain in balancing fairness with other performance metrics such as accuracy and efficiency. The review highlights the need for more standardized benchmarks and improved algorithms that provide equitable outcomes without compromising system performance. We provide insights into future directions for enhancing fairness across machine learning models.

Keywords: Bias, Fairness, Machine Learning, Mitigation Techniques, Systematic Review, Algorithmic Bias, Fairness Metrics

Introduction

- **Problem Statement**:
  Machine learning algorithms are increasingly being used to make decisions in sensitive areas, yet they can inadvertently perpetuate or even exacerbate biases. These biases are often reflected in underrepresented groups and lead to unfair outcomes, presenting a significant challenge to fairness in AI systems.
- **Importance of the Topic**:
  Bias in machine learning models raises ethical concerns and may reinforce systemic inequalities. Ensuring fairness is essential for ML models to be widely accepted and implemented in critical decision-making applications.

- **Scope and Objectives**:
  This paper aims to provide a comprehensive review of current bias mitigation techniques in ML, focusing on the categorization of approaches, their effectiveness, challenges, and gaps in the current research. The objective is to synthesize existing findings and propose future directions for bias mitigation.

Literature Review

- **Overview of Bias in Machine Learning**:
  Bias arises when a model's predictions favor certain groups over others, often due to biased data, flawed algorithms, or unrepresentative training data. Examples include racial or gender bias in hiring algorithms and predictive policing models.
- **Previous Mitigation Techniques**:
  Past research has proposed several strategies to address bias, including:
  - **Pre-processing**: Altering the training dataset to ensure fair representation of different groups.
  - **In-processing**: Modifying the learning algorithm to promote fairness during training.

Post-processing: Adjusting the model's predictions after the learning phase to ensure equitable outcomes.

Diagrams and Frameworks

Bias Mitigation Approaches Framework

The Bias Mitigation Approaches Framework explains the various methods for reducing bias in machine learning models, categorized into three main strategies. These strategies aim to make ML systems fairer and less likely to favor certain groups over others based on factors like race, gender, or socioeconomic status. The three categories are:

1. **Pre-processing**:
   These techniques are applied before the model is trained. The goal is to modify the data so that it is more balanced and less likely to perpetuate biases. Methods in this category include:

- **Re-sampling**: Adjusting the dataset by over-sampling underrepresented groups or under-sampling overrepresented groups.
- **Re-weighting**: Assigning different weights to examples from different groups to ensure fairer representation during training.
- **Data Transformation**: Modifying the features in the data to remove or reduce any biases that could influence the model.

2. **In-processing**:
   These approaches modify the training process itself to ensure fairness during the learning phase. By adjusting the optimization process, these techniques aim to balance performance with fairness. Methods include:
   - **Adversarial Debiasing**: Using adversarial networks to ensure the model does not learn to rely on sensitive features (such as race or gender) during training.
   - **Fairness-Constrained Optimization**: Adding fairness constraints into the objective function, ensuring that the model not only minimizes error but also satisfies fairness criteria.
3. **Post-processing**:
   These methods are applied after the model has been trained and its predictions are made. They aim to adjust the outcomes to make them fairer without altering the underlying model. Techniques include:
   - **Equalized Odds Correction**: Adjusting the decision thresholds for different groups to ensure equal false positive and false negative rates across those groups.
   - **Re-ranking**: Changing the order of predictions to promote fairness, such as re-ordering candidates in a hiring system to ensure diverse representation.

The diagram likely visualizes these three categories and their respective methods, helping to provide a clear understanding of how bias mitigation strategies are applied at different stages of the machine learning process.

Overview of Pre-processing, In-processing, and Post-processing Techniques

Category	Techniques	Description
Pre-processing	Re-sampling	Adjusting the dataset by over-sampling underrepresented groups or under-sampling overrepresented groups.
	Re-weighting	Assigning different weights to examples from different groups to ensure fairer representation during training.
	Data Transformation	Modifying features in the data to reduce bias and ensure fairness across sensitive attributes.

Category	Techniques	Description
In-processing	Adversarial Debiasing	Using adversarial networks to prevent the model from relying on sensitive attributes like race or gender.
	Fairness-Constrained Optimization	Incorporating fairness constraints into the model's optimization process to ensure fairness during training.
Post-processing	Equalized Odds Correction	Adjusting decisions are thresholds for different groups to ensure equal false positive and false negative rates.
	Re-ranking	Re-ordering predictions or results to ensure equitable representation, especially in ranking-based tasks like hiring.

This table summarizes various techniques used to mitigate bias in machine learning models, categorized into three stages: Pre-processing, In-processing, and post-processing. Each stage involves different methods aimed at addressing bias, ensuring fairness, and improving the performance of machine learning models, particularly in sensitive applications such as hiring, healthcare, and criminal justice.

Pre-processing Techniques:

These techniques are applied before training the model and primarily focus on preparing the dataset to be more balanced and less biased.

1. **Re-sampling**:
   This technique addresses imbalances in the dataset by either over-sampling underrepresented groups or under-sampling overrepresented groups. The goal is to create a more representative and balanced dataset that does not favor any particular group, which helps reduce bias in the model's predictions.

2. **Re-weighting**:
   In this method, different weights are assigned to examples from different groups. For instance, if one group is underrepresented, their examples can be given a higher weight during training. This ensures that the model considers these examples more heavily, promoting fairness across different groups.

3. **Data Transformation**:
   This technique involves altering the features in the dataset to reduce or eliminate any biases associated with sensitive attributes, such as race, gender, or socioeconomic status. For example, sensitive features might be removed, or modified versions of features can be used to prevent the model from using these attributes to make decisions.

In-processing Techniques:

In-processing techniques focus on modifying the model's training process itself to ensure fairness during learning, integrating fairness considerations directly into the model optimization.

1. **Adversarial Debiasing**:

   This technique uses adversarial networks to train the model in such a way that it does not learn to rely on sensitive attributes (e.g., gender or ethnicity) during the prediction process. An adversarial debiasing model typically includes a second "adversary" network that attempts to predict the sensitive attribute, while the primary model is trained to prevent the adversary from succeeding. The goal is to make the model more robust to bias.

2. **Fairness-Constrained Optimization**:

   In fairness-constrained optimization, fairness constraints are added to the model's objective function during training. These constraints ensure that the model not only minimizes prediction error (e.g., mean squared error) but also satisfies fairness criteria, such as equal treatment of different demographic groups. This approach ensures fairness is embedded within the learning process.

Post-processing Techniques:

Post-processing techniques are applied after the model has been trained, adjusting the model's predictions to improve fairness without altering the model itself.

1. **Equalized Odds Correction**

   Equalized odds are a fairness metric that ensures both false positive rates and false negative rates are equal across different groups. In post-processing, decision thresholds can be adjusted to make sure that the model's predictions do not disproportionately affect one group over another in terms of errors (false positives or false negatives). This method aims to equalize the outcomes across sensitive groups.

2. **Re-ranking**

Re-ranking techniques are commonly used in ranking-based tasks, such as job candidate selection or loan approval, where the model outputs a ranked list. In this case, the results can be reordered to ensure that diverse or underrepresented groups are equally represented or given fairer consideration. Re-ranking ensures that the fairness of the final output is improved, especially when ranking is based on multiple factors, such as qualifications or creditworthiness.

Fairness Metrics Comparison

Fairness Metric	Description	Applicability	Advantages	Disadvantages
Demographic Parity	Ensuring that the decision rate (e.g., loan approval, hiring) is the same across groups (e.g., gender, race).	Suitable for applications where equal representation is desired across groups.	Easy to understand and apply; guarantees group parity.	May lead to suboptimal decisions for the groups with lower representation in the dataset.

Fairness Metric	Description	Applicability	Advantages	Disadvantages
Equalized Odds	Ensures that both false positive rates and false negative rates are equal across groups.	Suitable for classification tasks where accurate decision-making is critical.	Ensures fairness in error rates, providing balanced outcomes.	Might not be achievable in highly imbalanced datasets.
Equal Opportunity	A variant of Equalized Odds, focusing on equalizing **false negative rates** across groups.	Applicable in sensitive areas like criminal justice (e.g., risk assessment).	Focus on reducing discriminatory outcomes in critical decisions.	May not always lead to overall fairness in all types of decision-making.
Predictive Parity	Ensures that the predictive value (e.g., precision) is equal across groups.	Suitable for applications where accurate prediction is critical, such as medical diagnoses.	Ensures that predictive performance is fair across groups.	May conflict with other fairness goals (e.g., equalized odds).
Calibration Within Groups	Ensure that predicted probabilities are accurate for all groups, i.e., equal probability of an event for a given score.	Relevant for predictive models that provide probability outputs (e.g., risk prediction models).	Helps ensure that predictions reflect the true probabilities for all groups.	May not be directly applicable to all machine learning models.
Fairness Through Awareness	Focuses on ensuring that sensitive attributes are explicitly accounted for in decision-making.	Applicable when fairness depends on understanding and incorporating specific sensitive attributes.	Focuses on explicit handling of sensitive attributes.	Difficult to implement in practice, especially with complex models.
Disparate Impact	Measures whether decisions disproportionately affect a protected group compared to others, often using statistical thresholds.	Suitable for legal applications, especially in employment and lending practices.	Clear legal and ethical guidelines for evaluating fairness.	May overlook other important fairness aspects, such as the quality of decision outcomes.

Detailed Explanation of Fairness Metrics:

1. **Demographic Parity**:
    - **Description**: This metric ensures that the proportion of positive decisions (e.g., approvals) is the same for different groups (e.g., male vs. female, white vs. non-white). If one group receives significantly fewer positive outcomes, the model is considered biased.
    - **Applicability**: Often used in hiring, lending, or admission processes where equal representation is desirable.
    - **Advantages**: It's easy to implement and measure. It ensures fairness in terms of group-level outcomes.
    - **Disadvantages**: While it ensures equal representation, it can sometimes lead to suboptimal outcomes for groups with lower representation in the data. For example, a less qualified group might still receive the same proportion of positive decisions.
2. **Equalized Odds**:

- **Description**: This metric requires that both **false positive rates (FPR)** and **false negative rates (FNR)** are equal across different groups. This means that the model should treat both groups with equal accuracy and error rates, regardless of their demographic composition.
- **Applicability**: Ideal for critical tasks like criminal justice risk assessment or medical diagnoses, where fair error rates are necessary.
- **Advantages**: It ensures fairness in terms of both error types, which can prevent discrimination against specific groups.
- **Disadvantages**: In datasets with significant imbalance, achieving equalized odds may be challenging because the group with fewer instances may have a high error rate due to insufficient training data.

3. **Equal Opportunity**:
   - **Description**: A variant of equalized odds, this metric only focuses on ensuring **equal false negative rates** across groups. This is particularly important when avoiding the wrongful denial of services or opportunities (e.g., parole, job applications).
   - **Applicability**: Useful in applications such as criminal justice, where reducing false negatives (i.e., unfairly denying someone a favorable decision) is a priority.
   - **Advantages**: This is crucial when it is important to minimize wrongful denials or under-predictions of potential positive outcomes.
   - **Disadvantages**: Focusing only on false negatives could lead to disparities in other error types (e.g., false positives), which might not be acceptable in certain applications.
4. **Predictive Parity**:
   - **Description**: This metric ensures that the **predictive accuracy (precision)** is the same across groups. For example, if the model predicts that a person will repay a loan, predictive parity ensures that the likelihood of success is equally accurate for all groups.
   - **Applicability**: Critical in medical diagnostics, credit scoring, or any application that requires accurate predictions across all groups.
   - **Advantages**: Ensures that the predictions made by the model are just as reliable for all groups, making it more fair in terms of accuracy.
   - **Disadvantages**: Predictive parity may conflict with other fairness metrics, such as equalized odds, as balancing prediction accuracy across all groups may not always align with minimizing errors.
5. **Calibration Within Groups**:

- **Description**: This metric ensures that predicted probabilities are accurate within each group. For instance, if the model predicts a 70% chance of an event occurring, this should be equally valid for each group.
- **Applicability**: Ideal for predictive models that output probabilities (e.g., loan approval probabilities or medical risk assessments).
- **Advantages**: Ensures that the predicted probabilities truly reflect the likelihood of an event, improving the fairness of decision-making.
- **Disadvantages**: It is not applicable to all types of models, especially non-probabilistic ones like certain classification models.

6. **Fairness Through Awareness**:
   - **Description**: This metric ensures that sensitive attributes (e.g., race, gender) are explicitly accounted for during decision-making. The idea is that certain groups should not be treated unfairly because their sensitive attributes are part of the decision process.
   - **Applicability**: Important when the decision-making process is inherently dependent on sensitive attributes, like in affirmative action policies or when designing fairness-aware algorithms.
   - **Advantages**: It explicitly takes sensitive attributes into account, preventing the model from indirectly discriminating against protected groups.
   - **Disadvantages**: Implementing this metric can be challenging because it requires complex understanding and control over how sensitive attributes influence the model's behavior.
7. **Disparate Impact**:
   - **Description**: This metric assesses whether decisions disproportionately affect certain protected groups (e.g., minorities, women) compared to others. It is often used in legal settings, particularly when there is a need to ensure compliance with anti-discrimination laws.
   - **Applicability**: Common in employment and lending decisions where there are strict legal frameworks for fairness.
   - **Advantages**: It provides a clear legal framework for evaluating fairness, making it easier for organizations to ensure compliance with laws and regulations.

Disadvantages: Disparate impact focuses solely on statistical fairness and might overlook other fair aspects, such as the quality or accuracy of decisions.

Model Performance Metrics with Bias Mitigation

Metric	Description	Purpose	When to Use	Effect of Bias Mitigation
Accuracy	The proportion of correct predictions made by the model.	Measures the overall performance of the model.	General performance metric for classification models.	It can be misleading if the dataset is imbalanced; bias mitigation might adjust this metric to avoid favoring dominant classes.
Precision	The proportion of true positive predictions among all positive predictions.	Measures the model's accuracy in predicting positive outcomes.	Use when false positives have high costs (e.g., medical diagnostics).	Bias mitigation could lead to more balanced precision across different groups, avoiding disparities.
Recall	The proportion of true positive predictions among all actual positive cases.	Measures the model's ability to identify all relevant instances.	Use when false negatives have high costs (e.g., crime prediction).	Bias mitigation could improve recall for underrepresented groups by minimizing false negatives.
F1-Score	The harmonic means of precision and recall.	Balances precision and recall ensure both are considered in performance.	Useful when both false positives and false negatives are critical.	Bias mitigation might adjust the F1 score by balancing precision and recalling across groups.
Area Under the Curve (AUC)	The area under the Receiver Operating Characteristic (ROC) curve, which plots true positive rate vs. false positive rate.	Measures the model's ability to distinguish between classes.	Ideal when comparing the trade-off between true positives and false positives.	Bias mitigation can lead to a more balanced AUC, ensuring the model performs fairly across different groups.
Equal Opportunity Difference	The difference in true positive rates between groups (e.g., men vs. women).	Assesses fairness in terms of opportunity.	When fairness in opportunity is prioritized over other metrics.	Bias mitigation aims to equal true positive rates, reducing this difference between groups.
Demographic Parity Difference	The difference in the proportion of positive predictions between groups (e.g., racial or gender groups).	Measures fairness in terms of equal representation.	When ensuring group-level parity is a priority.	Bias mitigation reduces this difference, aiming for equal representation in the outcomes.
Disparate Impact	The ratio of the positive prediction rates between two groups (e.g., minority vs. majority groups).	Measures fairness by checking whether a protected group is disadvantaged.	Used in regulated sectors like employment or lending.	Bias mitigation reduces disparate impact, ensuring a similar positive prediction rate for different groups.
Fairness Accuracy	Accuracy measured across different groups in the dataset, ensuring fair performance across all groups.	Measures fairness while accounting for different groups' accuracy.	Use when it is critical that the model is accurate for all groups.	Bias mitigation helps in achieving fair accuracy across groups, ensuring no group is unfairly penalized.

Detailed Explanation of Model Performance Metrics with Bias Mitigation:

1. **Accuracy**:
    - **Description**: Accuracy is the ratio of correct predictions (both true positives and true negatives) to the total number of instances. While a useful overall performance metric, it

can be misleading when the dataset is imbalanced (e.g., a majority class with very few minority class samples).

- **Purpose**: General performance metric to evaluate classification models.
- **Effect of Bias Mitigation**: In the presence of bias, accuracy may favor the majority of class. Bias mitigation techniques can adjust model outputs to reduce this bias, aiming for a more balanced outcome across groups, potentially lowering accuracy if the minority groups' accuracy improves at the expense of the majority.

2. **Precision**:
   - **Description**: Precision is the proportion of true positive predictions among all positive predictions. It is important when the cost of false positives is high, such as in medical diagnoses (e.g., predicting a disease when the patient does not have it).
   - **Purpose**: Measures how well the model avoids false positives.
   - **Effect of Bias Mitigation**: Bias mitigation techniques can be used to ensure precision is equally high across groups, preventing a situation where one group is unfairly penalized with more false positives.
3. **Recall**:
   - **Description**: Recall is the proportion of true positive predictions among all actual positive cases. It measures the model's ability to identify all relevant instances in the dataset.
   - **Purpose**: Helps to identify how many positive cases are captured by the model.
   - **Effect of Bias Mitigation**: Bias mitigation can improve recall for underrepresented groups by reducing false negatives (e.g., not identifying people who should be flagged as positive).
4. **F1-Score**:
   - **Description**: The F1-score is the harmonic means of precision and recall, offering a balanced measure of model performance when both false positives and false negatives are important.
   - **Purpose**: Useful when both precision and recall are critical and need to be balanced, such as in fraud detection.
   - **Effect of Bias Mitigation**: By improving both precision and recall across different groups, bias mitigation can improve the overall F1 score for each group, reducing disparities between them.
5. **Area Under the Curve (AUC)**:

- **Description**: AUC refers to the area under the ROC curve, which plots the true positive rate against the false positive rate. This metric is useful for understanding how well the model distinguishes between classes, especially in imbalanced datasets.
- **Purpose**: Evaluates how well the model discriminates between classes, independent of the decision threshold.
- **Effect of Bias Mitigation**: AUC is sensitive to the balance between classes. Bias mitigation can lead to a more balanced AUC, ensuring that the model distinguishes well between classes for all groups, not just the majority class.

6. **Equal Opportunity Difference**:
   - **Description**: This metric calculates the difference in true positive rates (TPRs) between different groups. A smaller difference means the model is treating all groups more equally in terms of identifying positive cases.
   - **Purpose**: Ensures fairness in terms of opportunity to be positively predicted by the model.
   - **Effect of Bias Mitigation**: Bias mitigation techniques aim to reduce disparities in true positive rates, promoting equal opportunity across all groups.
7. **Demographic Parity Difference**:
   - **Description**: This metric measures the difference in the proportion of positive predictions between different groups. A smaller difference means that the decision rate is more equal across groups.
   - **Purpose**: Measures fairness in terms of equal representation in positive outcomes.
   - **Effect of Bias Mitigation**: Bias mitigation aims to equalize the positive prediction rate across groups, achieving demographic parity.
8. **Disparate Impact**:
   - **Description**: Disparate impact compares the rate of positive predictions between two groups, typically a minority and majority group. A ratio closer to 1 indicates fairness.
   - **Purpose**: Measures the potential discrimination against a protected group (e.g., gender, race).
   - **Effect of Bias Mitigation**: Bias mitigation techniques reduce disparate impact by ensuring that the positive prediction rate for different groups is more aligned, reducing the likelihood that a minority group is unfairly disadvantaged.
9. **Fairness Accuracy**:
   - **Description**: This metric measures the accuracy of a model across different groups, ensuring that no group suffers from lower accuracy.

- **Purpose**: Ensures fairness in terms of model performance for each group.

Effect of Bias Mitigation: Bias mitigation ensures that the model performs equitably across all groups, maintaining high accuracy for both the majority and minority groups, thereby enhancing fairness.

Conclusion

In this systematic review, we have explored the various bias and fairness challenges present in machine learning models and the techniques used to mitigate them. Bias in machine learning can lead to unfair outcomes that disproportionately affect certain groups, raising ethical concerns, especially in sensitive applications like healthcare, hiring, and law enforcement. Our review examined three major categories of bias mitigation techniques: pre-processing, in-processing, and post-processing, each offering unique strategies for addressing these issues at different stages of the machine learning pipeline.

We also discussed various fairness metrics that evaluate the effectiveness of these mitigation strategies, highlighting the trade-offs between different fairness objectives and model performance. The comparison of these metrics provides a comprehensive understanding of how fairness can be quantified and balanced with traditional performance measures such as accuracy, precision, and recall.

Although progress has been made in developing bias mitigation techniques, challenges remain in achieving fairness without sacrificing model performance. Future research should focus on developing more sophisticated methods that dynamically balance fairness and performance, along with exploring new fairness metrics tailored to specific domains and real-world applications.

Ultimately, the adoption of fair and unbiased machine learning systems will require collaboration across fields, including machine learning, ethics, and law, to ensure that these technologies benefit all individuals equitably. The continuous evaluation of fairness in machine learning models is essential for building trust and ensuring that AI systems serve the broader good of society.

References

1. Gholami, A., Yao, Z., Mahoney, M. W., & Keutzer, K. (2018). A Survey on Deep Learning Hardware: Challenges and Trends. arXiv preprint arXiv:1805.10399, 1(1), 1–21.
2. Dosovitskiy, A., & Brox, T. (2018). Generating Videos with Scene Dynamics. International Journal of Computer Vision, 126(10), 1073–1088.
3. Dalal, A., Abdul, S., Kothamali, P. R., & Mahjabeen, F. (2015). Cybersecurity Challenges for the Internet of Things: Securing IoT in the US, Canada, and EU.International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence,6(1), 53-64.
4. Dalal, A., Abdul, S., Kothamali, P. R., & Mahjabeen, F. (2017). Integrating Blockchain with ERP Systems: Revolutionizing Data Security and Process Transparency in SAP.Revista de Inteligencia Artificial en Medicina,8(1), 66-77.
5. Dalal, A., Abdul, S., Mahjabeen, F., & Kothamali, P. R. (2018). Advanced Governance, Risk, and Compliance Strategies for SAP and ERP Systems in the US and Europe: Leveraging Automation and

Analytics. International Journal of Advanced Engineering Technologies and Innovations, 1(2), 30-43. https://ijaeti.com/index.php/Journal/article/view/577