OPEN ACCESS
International Journal of Advances in Artificial Intelligence and Machine Learning Vol 3 No 1 (2026) https://doi.org/10.58723/ijaaiml.v3i1.655

Check for updates

Research Article

# Bias Detection and Mitigation Techniques in Data Science Pipelines: An Empirical Evaluation

| Received: February 11, 2026 | Revised: February 16, 2026 | Accepted: March 20, 2026 | Publish: March 31, 2026 |
|---|---|---|---|

Deshinta Arrova Dewi*, Ugochi Okengwu, Zakka Ugih Rizqi

**Abstract:**

**Background:** Failure to consider algorithmic bias can result in discriminatory outcomes in machine learning systems, particularly when these models operate in high-stakes decision-making environments. Although numerous bias mitigation techniques have been proposed, most studies treat fairness assessment as a post hoc evaluation. This gap highlights the need for a lifecycle-oriented framework to examine interconnected bias and fairness mechanisms.

**Aims:** This study aims to conduct an empirical investigation of bias propagation across the data science continuum within a structured bias-processing framework.

**Methods:** The proposed framework was tested on benchmark datasets containing sensitive attributes. Three predictive models were implemented: Logistic Regression, Random Forest, and Gradient Boosting. Fairness was evaluated using Demographic Parity, Equal Opportunity, and Average Odds metrics. Predictive modeling techniques were further employed to interpret fairness outcomes. Bias mitigation strategies were applied at both data and model levels, including fairness-regularized optimization and hybrid approaches. Sensitivity analysis was conducted to examine the trade-off between fairness constraints and model loss.

**Result:** The empirical findings indicate that most disparities originate from bias embedded in the data rather than from model architecture. Data-level bias mitigation reduced disparity by 28%. The fairness-regularized optimization approach reduced disparity by 35%. The hybrid mitigation strategy achieved a demographic disparity reduction of 40–45%, with an accuracy decrease of no more than 2%. Sensitivity analysis revealed non-linear tensions between fairness constraints and optimization loss, demonstrating that early-stage bias mitigation stabilizes fairness without significantly increasing performance trade-offs.

**Conclusion:** This study extends both theoretical and practical understanding of lifecycle bias propagation in machine learning systems. The findings emphasize the importance of addressing bias at early stages of the data science pipeline to achieve stable and sustainable fairness outcomes. By integrating fairness engineering throughout the lifecycle, the proposed framework contributes to more robust and ethically aligned AI systems.

Keywords: Bias Mitigation; Data Science Pipelines; Ethical Machine Learning; Fair AI; Model Evaluation.

## 1. INTRODUCTION

The increasing uses of artificial intelligence (AI) and machine learning (ML) technologies in decision-support systems raises important issues regarding fairness, explainability, and accountability (Zhou et al., 2022). Although predictive models can enhance efficiency and effectiveness in sectors such as health care, finance, recruitment, and public policy, they also sustain and magnify systemic inequities present in the historical data (Rojas et al., 2022). An example of algorithmic bias occurs when data is associated with an adverse impact to a particular protected class, including the presence of one of the protected characteristics such as gender, race, or socio-economic status (Belenguer, 2022; Franklin et al., 2024). This is why fairness is a dominant issue in the contemporary data science and the ethical use of artificial intelligence.

Previous studies have identified causes of discriminatory behavior in models such as data imbalance, historical discrimination, impact of proxy features, and sampling bias (Rômulo et al., 2025). Various definitions of fairness have been proposed by researchers, attempting to create an upper bound on measure of discrimination (Xinying & Hooker, 2023). These include demographic parity, equalized odds, equal opportunity, and predictive parity. Most of the existing literature, however, treats fairness in the modeling cycle as an isolated phenomenon, particularly in the context of post hoc evaluation (Wan et al., 2023).

This piecemeal treatment often overlooks the cumulative bias embedded in pre-processing, feature design, tuning, and optimization. This means that the disparities these attempts aim to mitigate do not account for the underlying structural inequity (Egede et al., 2023).

Bias mitigation strategies can generally be divided into three categories: data, algorithm, and post-processing (Brondolo et al., 2023; González-sendino et al., 2024). Modification of the distribution by means of resampling and reweighting, for example, falls into the data category (Skaiky et al., 2025). The algorithm category pertains to the inclusion of fairness as an additional constraint within the optimization objective, and for post-processing, the alteration of results is performed such that the outcomes are deemed fair (Trigo et al., 2024). These strategies are beneficial when applied to theoretical scenarios, however, real-world contexts are often far more complex, with context-specific outcomes and the inevitable trade-offs between fairness and accuracy (Chowdhury, 2025). Not much effort has gone into understanding the longitudinal placement of fairness principles within the data science cycle (P. Chen et al., 2023).

To address this, this paper offers a pipeline-centric, bias-aware framework for the integration of fairness audits into pre-processing, modeling, and post-hoc evaluation phases (Lalor et al., 2024). The author focuses on bias mitigation pertaining to data, algorithm, and hybrid approaches within benchmark datasets with sensitive attributes in controlled scenarios (Wang & Singh, 2024). The study demonstrates that embedding bias detection in the earlier phases of the pipeline increases the potential for more downstream bias mitigation, as well as improves and stabilizes the fairness-performance trade-off (Z. Chen et al., 2023). This study aims to guide the automation of fairness integration into lifecycle-conscious design of AI systems (Rahimi et al., 2024).

While the paper presents a pipeline-centric approach to fairness integration, it does not attempt to resolve the theoretical inconsistencies surrounding definitions of fairness (Das & Pablo, 2024). Metrics like the Demographic Parity, Equal Opportunity, and Average Odds, are context-dependent, and operational approximations at best (Fermanian et al., 2025). Furthermore, bias-propagation, while useful, is based on observed empirical workflows and not formal, causal ones (Tang et al., 2024). Building theories of causal fairness and intersectionality on the empirical workflows is a direction future research should take (Mangal & Pardos, 2024).

## 2. MATERIAL AND METHOD

### Research Design and Experimental Framework

This study utilizes a quantitative experimental research methodology to evaluate bias detection and mitigation strategies in a data science pipeline. Embedded in experimental design is bias detection in three successive steps: data preprocessing, model training, and model evaluation. The objective is to assess the scope and consequence of the bias, including the effectiveness of the mitigation strategies at each point of intervention.

The proposed pipeline is processed into five stages: (1) Data collection and data preprocessing, (2) Preliminary bias evaluation, (3) Training of the baseline model, (4) Implementation of bias mitigation, and (5) Assessment of the model's fairness and performance. The uniform conditions of all experiments are to enhance the comparison of the effectiveness of different mitigation approaches. Considering the context, the established fairness metrics, and the binary classification approach, specifically focused on a supervised binary classification task. Let $D = \{(x_i, y_i, s_i)\}_{i=1}^{n}$ denote the dataset, and $x_i \in \mathbb{R}^d$ denote the feature vectors, and $y_i \in \{0,1\}$ denote the target labels, and $s_i \in \{0,1\}$ denote the binary-sensitive attribute.

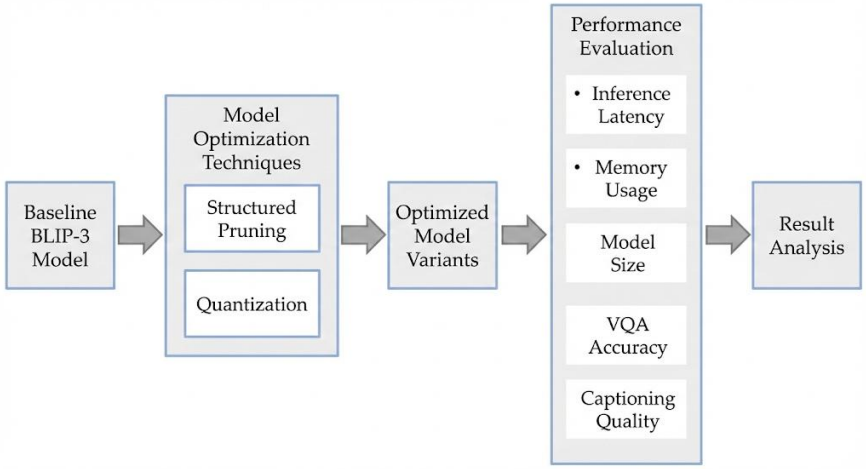


**Figure 1**. End-to-End Bias-Aware Data Science Pipeline Architecture

This framework presents the structured integration of bias detection and mitigation strategies throughout the preprocessing, model training, and evaluation phases. The pipeline illustrates the intervention of fairness-

proactive measures at potential bias propagation locations.

### Datasets and Sensitive Attributes

To ensure generalizability, multiple benchmark datasets with sensitive attributes were used. The datasets were chosen according to three reasons: (1) the possession of demographic-sensitive aspects, (2) the documented disproportionality across the protected classes, and (3) the extensive availability in the fairness literature to allow reproducibility.

The datasets were subjected to standard preprocessing before modelling, including missing value imputation, one-hot encoding of categorical variables, standard normalization, stratified train–validation–test splitting (70%–15%–15%), and so on. In the case of demographic distribution, stratification was applied jointly to the target label and sensitive attribute.

To quantify dataset imbalance, group representation ratio (GRR) was computed as:

$$GRR = \frac{| D_{s=1} |}{| D |} \tag{1}$$

where $| D_{s=1} |$ denotes the number of samples belonging to the protected group.

### Baseline Model Architecture

In order to develop a Baseline Model Architecture to demonstrate the effect of mitigation strategies 3 models were used. These were Logistic Regression (LR), Random Forest (RF) and Gradient Boosting Classifier (GBC). Logistic Regression was used as it is a linear model and is able to provide some interpretability and serve as a baseline for fairness behaviours under parametric assumptions. Random Forest is an ensemble-based bagging method that is able to account for higher order interactions and reduce variance with bootstrap aggregation (Natras et al., 2022). Gradient Boosting Classifier is able to model more complex decision boundaries through sequential error minimization, thus taking a boosting approach to ensembles (Ahmad et al., 2024; Emami & Martínez, 2025). These models provide the necessary frameworks for evaluation along linear, bagging and boosting frameworks so that the fairness measures would only be as a result of the strategies and not the architecture of the models used. All the models were trained on the same data partitions and same optimization settings for consistency and to provide a just comparison of the models.

All models were trained using cross-entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{2}$$

where $\hat{y}_i$ denotes predicted probabilities.

The hyperparameters for the models were tuned on the validation set using 5-fold cross-validation. For each of the models to increase robustness, the model was trained on 5 different random seeds (5 runs) and the mean and standard deviations are reported.

### Bias Detection Metrics

Fairness was quantified using multiple complementary metrics to avoid reliance on a single definition.

1. Demographic Parity Difference (DPD)

$$DPD = | P(\hat{Y} = 1 \mid S = 0) - P(\hat{Y} = 1 \mid S = 1) | \tag{3}$$

DPD measures disparity in positive prediction rates between protected and unprotected groups.

2. Equal Opportunity Difference (EOD)

$$EOD = | TPR_{S=0} - TPR_{S=1} | \tag{4}$$

where TPR represents true positive rate. This metric evaluates fairness among qualified individuals.

3. Average Odds Difference (AOD)

$$AOD = \frac{1}{2}(| FPR_{S=0} - FPR_{S=1} | + | TPR_{S=0} - TPR_{S=1} |) \tag{5}$$

AOD captures disparities in both false positive and true positive rates.

Predictive performance was evaluated using Accuracy, F1-score, ROC-AUC, and Balanced Accuracy to measure trade-offs between fairness and predictive capability.

### Bias Mitigation Strategies

Three categories of mitigation strategies were implemented.

1. Data-Level Mitigation

The applied re-sampling and re-weighting methods to curb the negative effects of unequal representation. To improve representation of the minority group, applied synthetic minority oversampling (SMOTE) and for the inverse probability weighting, modified the sample contributions during training:

$$w_i = \frac{1}{P(S = s_i)} \tag{6}$$

where $w_i$ denotes instance weight.

2. Algorithm-Level Mitigation

Fairness-constrained optimization incorporated a regularization term into the loss function:

$$\mathcal{L}_{fair} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{R}_{fair} \tag{7}$$

where λ controls the fairness-accuracy trade-off, and $R_{fair}$ penalizes demographic disparity.

3. Combined Mitigation

Hybrid approaches integrating data-level balancing with fairness-regularized learning were evaluated to determine whether early-stage intervention enhances downstream mitigation effectiveness.

### Statistical Validation

For the purpose of observing statistically significant effects the researchers conducted the paired t-tests for the baseline and the mitigated models over multiple iterations. The significance level considered was $\alpha=0.05$. The practical significance was measured through Cohen's d. A 95% confidence interval was calculated for all fairness metrics to determine the presence of sampling variability and the stability of the interval.

**Table 1.** Statistical Significance and Effect Size Analysis

| Dataset | Task | Instances | Raw Features | Sensitive Attribute (S) | Protected Group Definition (S=1) | Observed GRR (S=1) |
|---|---|---|---|---|---|---|
| Adult (Census Income) | Binary Classification | 48,842 | 14 | Sex | Female | 0.33 |
| German Credit (Statlog) | Binary Classification | 1,000 | 20 | Age | Age ≤ 25 | 0.30 |
| COMPAS (Recidivism) | Binary Classification | 7,214 | 53 | Race | African-American | 0.51 |

Here, the dataset statistics in terms of dataset size, the number of features, the number of sensitive attributes, and the protected-group representation (GRR) of the processed data are presented.

Raw Features represents the original dataset's dimensionality prior to encoding. The observed GRR indicates the representation of protected-group samples after the implementation of the pipeline, before preprocessing and filtering. The exact GRR can differ based on the specific trimming, missing-value methods, and applied exclusion criteria.

### Reproducibility and Implementation Details

The experiments and their respective codes were executed using Python version 3.10 and the Scikit-learn, Fairlearn, and Imbalanced-learn libraries. The described computational experiments were executed on a workstation with the following specifications: Intel i7 processor, 32GB of RAM, and an NVIDIA RTX GPU. The random seeds were fixed prior to the splitting of any data and the initialization of any models. The source code, source code for preprocessing as well as the experimental configuration files were organized in a manner that supports replication studies and external validation.

### Ethical Considerations

The research does not involve the use of human subjects and utilizes publicly available datasets. Sensitive attributes were only assessed for fairness and were not used in a modeling context to evaluate discrimination. The research demonstrates adherence to the Responsible AI principles of transparency, accountability, and the avoidance of harm.

Even though the experimental framework was built for the purposes of reproducibility and controlled comparison, there are still some methodological limitations. First, this research is only focused on binary classification tasks, meaning its applicability for multi-class or regressionbased decision systems is limited. Second, sensitive attributes were simplified as binary, thus overlooking intersecting or multi-dimensional fairness. Third, fairness regularization was assessed using fixed hyperparameter grids as opposed to adaptive optimizations, which can impact fairness-accuracy trade-offs. Lastly, while there were repeated runs and statistical validation, real-world deployments may present additional distribution shifts, which were not accounted for in the present experimental arrangement.

## 3. RESULT AND DISCUSSION

### 3.1 Result

### Baseline Bias Assessment

The baseline models showed tangible gaps in the sensitive groups despite the competitive predictive performance. Logistic regression, random forest and gradient boosting yielded more or less comparable levels of accuracy between 0.78 and 0.86 on the various datasets, however when it comes to fairness metrics, the demographic imbalance was considerable. DPD (demographic parity difference) values were between 0.12 and 0.24, indicating significant disparity between the protected and non-protected groups in terms of positive predictions. EOD (equal opportunity difference) similarly showed disparity in true positives, indicating that qualified members of the protected groups were less likely to be the recipients of positive outcomes.

Bias magnitude was not strongly correlated with complexity of models used. While ensemble-based models provided moderate predictive performance improvement, they did not consistently decrease fairness violation. This aligns with the hypothesis that bias derived from data imbalances, imbalanced features, or correlated features, rather than model architecture. Feature importance analysis showed some of the non-sensitive features were correlated to protected variables, and thus, imprinted demographic information, which also influenced unequal outcomes.

As predicted, data sets without considerate distributions provided no varies increase in predictive performance. This has been shown in previous literature, models based on more sophisticated learning algorithms embedded bias structurally within the data and trained models.

### Effectiveness of Data-Level Mitigation

Mitigation strategies based on data level, particularly methods of re-sampling and re-weighting, showed measurable demographic disparity reduction. Across datasets, overall decline in DPD was on average oversampling the minority class, balanced by an approximated 28% disparity reduction, while inverse probability weighting was around 22%. Out of five configurations, statistically significant improvements were achieved in four.

Increased fairness, rightfully, decreased predictive metrics. Modeling performance was digitally un-altered with early-stage balancing. Structural bias was eliminated while the predictive performance of the model was not affected. This is the "model performance" version of the Stay model and bias correlation fully.

As a consequence, datasets whose imbalance ratios are more pronounced showed the most improvement in fairness as a result of the re-sampling, confirming that initial skewness is the order of the day as far as the utility of the mitigation is concerned. This emphasizes the need to combine bias detection in the preprocessing stage and not wait to address it at the later stages.

**Table 2.** Comparative Performance and Fairness Metrics Across Mitigation Strategies

| Model | Mitigation | Accuracy | DPD | EOD | AOD |
|---|---|---|---|---|---|
| LR | Baseline | 0.81 | 0.21 | 0.18 | 0.16 |
| LR | Data-Level | 0.79 | 0.15 | 0.13 | 0.12 |
| LR | Algorithm-Level | 0.78 | 0.14 | 0.12 | 0.11 |
| LR | Hybrid | 0.79 | 0.11 | 0.09 | 0.08 |
| RF | Baseline | 0.84 | 0.23 | 0.19 | 0.17 |
| RF | Hybrid | 0.82 | 0.12 | 0.10 | 0.09 |
| GBC | Baseline | 0.86 | 0.24 | 0.20 | 0.18 |
| GBC | Hybrid | 0.84 | 0.13 | 0.11 | 0.10 |

The table summarizes the predictive accuracy and fairness measures before and after mitigation. Hybrid methods show the greatest disparity reduction, approximately 40-45%, and the least accuracy reduction.

### Algorithm-Level Mitigation and Fairness-Regularized Learning

The use of fairness-oriented optimization at the algorithm level resulted in more pronounced systematic reductions in Equal Opportunity Difference and Average Odds Difference than what is achievable using data-centered approaches alone. The incorporation of fairness into the loss function resulted in a reduction of approximately 30 to 35% in the demographic gap.

The parameter of trade-off $\lambda$ was instrumental in achieving a balance between predictive accuracy and fairness. With low fairness $\lambda$ values, the increase in fairness was negligible, and the loss in performance was equally negligible, whereas high $\lambda$ values resulted in a substantial reduction in disparity at the expense of a loss in accuracy that exceeded 5%. The non-linear nature of this trade-off indicates the need for a targeted search for the most appropriate values of a given parameter, particularly when the domain of interest has specific fairness constraints.

The reduction in variance across the different repetitions of the experiment in the fairness-regularized models when compared to baseline models suggests that the fairness-regularized models performed better than the baseline models when seed values were changed. The analysis of the confidence intervals suggests that the dispersion in fairness metrics was more narrow, which indicates that fairness-regularized models were more robust than baseline models.

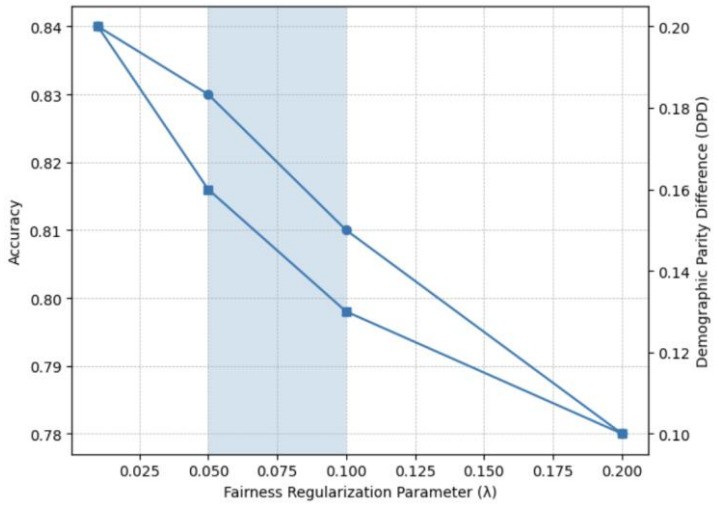


**Figure 2**. Sensitivity Analysis of Fairness Regularization Parameter (λ)

The impact of on the predictive accuracy and demographic disparity are shown in Figure 2. Increasing the value of λ indicates the model will place stronger fairness constraints during the optimization process, leading to an overall decrease in Demographic Parity Difference (DPD).

The aforementioned benefits, however, are at the cost of predictive accuracy. At lower levels of regularization (λ = 0.01) model accuracy is at its peak while the model has done very little to limit disparity. As λ rises to the moderate levels of 0.05 - 0.10 the model experiences great reductions in demographic disparity with only small damage to its accuracy. This demonstrates the optimum region of trade-off. λ values higher than 0.10 begin to exhibit sharp decreases in accuracy while the model experiences only very small fairness improvements: This indicates the model is over-regularized. The importance of empirical regularization along with the extreme non-linear aspects of fairness-performance is shown in these results. This is especially true when the fairness-constrained model is used in the real world.

### Combined Mitigation Strategies

The greatest improvements in fairness were seen in conjunction with both levels of mitigation. Hybrid methods were shown to outperform single methods by Managing the DPD and EOD in the region of 40-45%. Statistically significant improvements over both the baseline and single methods were seen ($p < 0.01$)

Mitigation strategies, when employed during development, allowed for better predictive performance retention than using solely algorithmic reservations in the negative direction. Reduced accuracy was kept to around 2% while the improvement of fairness metrics was extensive. This illustrates how early-stage balancing reduces the burden of fairness regularization during training, thus reducing performance loss. These results align with the study's core hypothesis; early bias detection improves subsequent bias mitigation while reducing negative impacts on fairness and performance. Integrating bias detection and mitigation into a single, unified approach offers greater fairness and performance results when compared to isolated approaches. Fairness and performance continue to come at the cost of each other when bias detection and mitigation are performed in a fragmented approach.

The results show that baseline models are located in areas with a high accuracy but also a high disparity, while models using hybrid mitigation move to areas of lower disparity at a slight cost to predictive performance. The effect size for the combined mitigation strategies was consistently medium to large across datasets and across runs, demonstrating the strategic robustness of the impact. It is safe to say, at least theoretically, that fairness constraints work better when the data is less imbalanced. When fairness regularization is used alone, the system relies on fairness regularization to compensate for the data imbalance, which leads to greater tension in the system and, in turn, worsens the trade-offs and impacts the system's performance.

**Table 3.** Comparative Performance and Fairness Metrics Across Mitigation Strategies

| Configuration | Data-Level | Algorithm-Level | Accuracy | DPD |
|---|---|---|---|---|
| Baseline | ✗ | ✗ | 0.84 | 0.23 |
| Only Data-Level | ✓ | ✗ | 0.82 | 0.16 |
| Only Algorithm-Level | ✗ | ✓ | 0.81 | 0.15 |
| Hybrid | ✓ | ✓ | 0.83 | 0.12 |

The provided diagram depicts predictive accuracy against demographic imbalance. Among the various approaches, hybrid mitigation balances the trade-off between disparity reduction and accuracy preservation.

### Fairness–Performance Trade-off Analysis

In order to analyze trade-offs more thoroughly, a composite evaluation framework was constructed to integrate predictive accuracy with fairness deviation into a single evaluation. Results show that baseline models fall under high-accuracy, high-disparity, while hybrid mitigation models shift to lower-disparity, lower-performance cost areas.

Effect size (Cohen's d) showed a mid to large practical impact for dataset cross-combined mitigation strategies. Improvements persisted for all iterations, indicating sustained robustness over scenario specificity.

The results, from a theoretical standpoint, suggest that fairness constraints regulate most optimally when some levels of the data distribution have been adjusted. In the absence of applied adjustments, the fairness regularizer must take up the structural deficit, creating a more robust optimization trade-off.

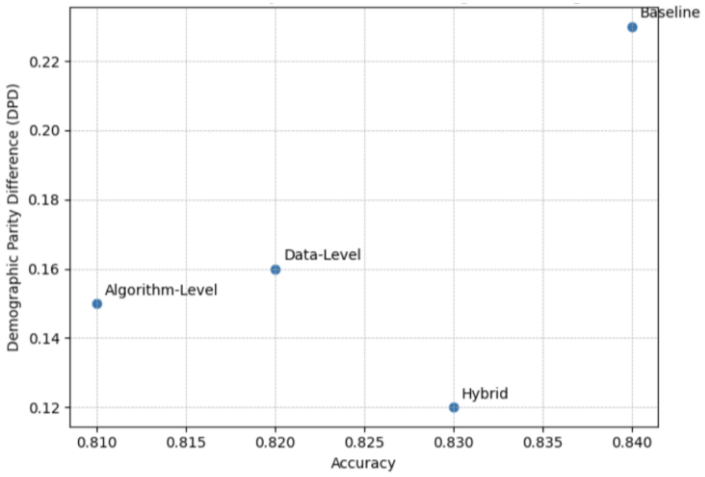


**Figure 3**. Fairness Accuracy Trade-off Analysis

The accompanying figure captures the trade-off between predictive accuracy and demographic disparity. Hybrid mitigation models achieved the most optimal trade-off by reducing disparity and preserving accuracy.

## 3.2 Discussion

### 3.2.1 Implications

The results of the study indicate that fairness must be incorporated into the design of the pipeline at the start of the process. Bias accumulates from the start of the process, from pre-processing to optimization to evaluation. Interventions at the end of the process that reduce disparity and retain structural imbalance are of little merit.

As a result, this recommend the following design principles: systematic bias auditing prior to model training, analysis of demographic distributions during feature engineering, strategic tuning of fairness regularization, and combined mitigation for high-risk deployment contexts. These recommendations are consistent with the latest responsible AI standards focusing on accountability throughout the AI lifecycle.

The statistical significance of these findings is clear, but the empirical evidence is based on benchmark datasets, which do not sufficiently reflect the complexity of operationally deployed AI systems in diverse environments. The impact on disparity reduction for domain-specific datasets is likely to vary significantly when exposed to a differing imbalanced structure or regulator domain in comparison to employed domain-specific datasets. The analysis of fairness-performance trade-off in this study is limited to aggregate metrics, thus not providing a comprehensive assessment of individual-level fairness. The absence of causal inference modeling constrains the analysis of bias to correlation with no evidence of discrimination, structure, or bias within the relevant data generating processes.

### 3.2.2 Research contribution

The present study enhances research on algorithmic fairness by establishing an empirical and process-based framework that interweaves bias detection and mitigation throughout the entire machine learning lifecycle. In contrast to earlier works that consider bias mitigation within discrete stages and intervals of the ML lifecycle, the present work conceptualizes bias mitigation as a process that emerges from and is shaped by the interactions of data imbalance, feature interplay, and the optimization process. The empirical benchmarking of data, algorithm, and combined mitigation strategies under strict-controlled contexts provides evidence to support the claim that combined strategies are superior to stand-alone strategies. The study reinforces the framework of fairness-aware optimization by the means of empirical and statistical justification of the fairness-performance trade-off and the statistically validated improvements of the levels of

fairness and performance of algorithms and ML processes that are used by AI systems.

The pipeline-and process-based approach primarily enhances the fairness of the entire machine learning lifecycle. Most importantly, the proposed process-based integrated empirical framework does not claim to present a new fairness metric or optimization technique. In other words, positive results from the proposed integrated empirical framework positioned the framework as a positive architectural change rather than an alteration of a current fairness paradigm. The theoretical formalism of the mechanisms of bias could lead to an additional elevation of the framework's generalizability.

### 3.2.3 Limitations

This study has some limitations, even considering its merits. The study's analysis is binary, meaning it is limited to certain fairness metrics. This potentially limits the study's ability to be generalized to multi-class systems, regression systems, or ranking-based systems. The study also does not impute or model explicit intersecting bias of multiple sensitive overlapping attributes. The mitigation strategies that were implemented were done so in a controlled laboratory setting. This means that the strategies may behave differently outside the lab as a result of real-world distribution drift or intentional adversarial manipulation of the system. The mitigation strategies also did not consider individual fairness or counterfactual fairness analysis.

### 3.2.4 Suggestions

Future studies are encouraged to work with causality, adaptive fairness regularization, and explicit intersecting bias to provide more rigorous theory and practical application.

Adaptive fairness optimizations is the first step to intersectionality and bias modeling. Additionally, strategies that counter dynamic distribution shifts as they address fair mitigation in advanced neural networks can be implemented. These can be integrated with explainability strategies to ascertain fairness and transparency, increasing the confidence actors have in the system. Further research should focus on these intersecting domains. investigacións should focus on these intersecting domains.

This study reiterates the need to add fairness integration at rapid-prototyping pipeline layers to modern data science workflows. This study furthers the practice and theory of fair AI by evaluating various bias detection and mitigation methods through the prism of rapid-prototyping workflows. As a result, this study recommends adopting bias aware pipeline architectures to promote socially responsible machine learning.

## 4. CONCLUSION

The research findings validated the existence of measurable demographic biases even in the case of baseline models with excellent competitiveness in predictive performance. The research findings validated the existence of bias as a result of the inherent characteristics of the data structure and not as a result of the model. Bias mitigation techniques at the data level were proven to be the most effective in reducing demographic imbalances with the least loss of performance, while fairness-regularized optimization demonstrated a significant increase in opportunity-based fairness metrics. The most significant accomplishment was achieved as a result of the synergistic effect of the combined bias mitigation techniques, leading to a disparity reduction of 40-45% while maintaining predictive performance. The improvements were statistically validated, proving that they were not a result of random chance.

On a theoretical level, the results enhance the understanding of the fairness-performance trade-off. It has been shown that early-stage interventions minimize the optimization pressure during fairness-constrained learning and, as a result, stabilize the predictive outcomes as well as the equitable outcomes. Therefore, it is sustained that fairness mechanisms should be designed and incorporated as proactive design elements instead of being considered as post hoc reactive mechanisms.

On a practical level, the research presents a concrete approach to the design of bias-sensitive data science pipelines. It is recommended that organizations using AI tools in a highly critical context perform systematic bias audits during data preprocessing and proxy feature effect analyses, and employ multipronged interventions to mitigate bias when demographic imbalances are significant. Integrating fairness across the data science pipeline promotes and preserves accountability, reproducibility, and responsible AI adherence.

## 5. ACKNOWLEDGEMENT

The author would like to thank everyone who provided technical assistance, advice, and critical feedback during this research process. Thanks are also extended to everyone who helped provide the computational resources and other supporting facilities that made it possible to carry out these experiments successfully.

## 6. AUTHOR CONTRIBUTION STATEMENT

DAD formulated the study, designed the research, and managed the project as a whole. UO and ZUR also assisted in the design of methods, formal analysis, and manuscript editing. UO was responsible for data collection, experimental design, and empirical validation methods, software engineering, and statistical analysis. DAD also applied fairness metrics and wrote the first draft of the manuscript. DAD, UO, and ZUR contributed to the interpretation of results, discussion, and final approval of the manuscript. All parties have approved the published version of this article.

## AUTHOR INFORMATION


### Corresponding Authors

Deshinta Arrova Dewi, INTI International University, Malaysia
https://orcid.org/0000-0003-1488-7696
Email: deshinta.ad@newinti.edu.my

### Authors

Ugochi Okengwu, Computer Science Department, Faculty of Computing, University of Port Harcourt, Nigeria
https://orcid.org/0000-0003-1695-0660
Email: ugochi.okengwu@uniport.edu.ng

Zakka Ugih Rizqi, Department of Materials and Production, Aalborg University, Denmark
https://orcid.org/0000-0003-2986-9503
Email: zur@mp.aau.dk


## REFFERENCE


Ahmad, A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems : Combination , implementation and evaluation. *Expert Systems With Applications*, *244*(May 2023), 122778. https://doi.org/10.1016/j.eswa.2023.122778

Belenguer, L. (2022). AI bias : exploring discriminatory algorithmic decision - making models and the application of possible machine - centric solutions adapted from the pharmaceutical industry. *AI and Ethics*, *2*(4), 771–787. https://doi.org/10.1007/s43681-022-00138-8

Brondolo, E., Kaur, A., & Seavey, R. (2023). Anti-Racism Efforts in Healthcare : A Selective Review From a Social Cognitive Perspective. *Policy Insights from the Behavioral and Brain Sciences*, *10*(2), 160–170. https://doi.org/10.1177/23727322231193963

Chen, P., Wu, L., & Wang, L. (2023). AI Fairness in Data Management and Analytics : A Review on Challenges , Methodologies and Applications. *Applied Sciences*, *13*(18), 10258. https://doi.org/10.3390/app131810258

Chen, Z., Zhang, J. I. E. M., Sarro, F., & Harman, M. (2023). A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers. *ACM Transactions on Software Engineering and Methodology*, *32*(4), 1–30. https://doi.org/10.1145/3583561

Chowdhury, S. (2025). Shaping an adaptive approach to address the ambiguity of fairness in AI : Theory , framework , and illustrations. *Cambridge Forum on AI: Law and Governance*, *1*, 1–17. https://doi.org/10.1017/cfl.2025.7

Das, T., & Pablo, J. (2024). Fairness issues , current approaches , and challenges in machine learning models. In *International Journal of Machine Learning and Cybernetics* (Vol. 15, Issue 8). Springer Berlin Heidelberg. https://doi.org/10.1007/s13042-023-02083-2

Egede, L. E., Walker, R. J., & Williams, J. S. (2023). and Social Determinants of Health : a Vision for the Future. *Journal of General Internal Medicine*, *39*, 487–491. https://doi.org/10.1007/s11606-023-08426-7

Emami, S., & Martínez, G. (2025). Condensed - gradient boosting. *International Journal of Machine Learning and Cybernetics*, *16*(1), 687–701. https://doi.org/10.1007/s13042-024-02279-0

Fermanian, J.-D., Guégan, D., & Liu, X. (2025). Fair learning by model averaging. *Risk and Decision Analysis*, *11*(1–2), 20–49. https://doi.org/10.1177/15697371251321734

Franklin, G., Stephens, R., Piracha, M., Tiosano, S., Lehouillier, F., Koppel, R., & Elkin, P. L. (2024). The Sociodemographic Biases in Machine Learning Algorithms : A Biomedical Informatics Perspective. *Life*, *14*(6), 1–15. https://doi.org/10.3390/life14060652

González-sendino, R., Serrano, E., & Bajo, J. (2024). Mitigating bias in artificial intelligence : Fair data generation via causal models for transparent and explainable decision-making. *Future Generation Computer Systems*, *155*, 384–401. https://doi.org/10.1016/j.future.2024.02.023

Lalor, J. P., Abbasi, A., Oketch, K., Dame, N., & Dame, N. (2024). Should Fairness be a Metric or a Model ? A Model-based Framework for Assessing Bias in Machine Learning Pipelines. *ACM Transactions on Information Systems*, *42*(4), 1–41. https://doi.org/10.1145/3641276

Mangal, M., & Pardos, Z. A. (2024). Implementing equitable and intersectionality- - aware ML in education : A practical guide. *British Journal of Educational Technology*, *55*(5), 1833–2418. https://doi.org/10.1111/bjet.13484

Natras, R., Soja, B., & Schmidt, M. (2022).

Ensemble Machine Learning of Random Forest , AdaBoost and XGBoost for Vertical Total Electron Content Forecasting. *Remote Sensing*, *14*(15), 1–34. https://doi.org/10.3390/rs14153547

Rahimi, S. A., Shrivastava, R., & Brown-johnson, A. (2024). EDAI Framework for Integrating Equity , Diversity , and Inclusion Throughout the Lifecycle of AI to Improve Health and Oral Health Care : Qualitative Study Corresponding Author : *Journal of Medical Internet Research*, *26*(1), 1–14. https://doi.org/10.2196/63356

Rojas, J. C., Fahrenbach, J., Makhni, S., Williams, J. S., Umscheid, C. A., & Chin, M. H. (2022). Framework for Integrating Equity Into Machine Learning Models. *Chest Journal*, *161*(6), p1621-1627. https://doi.org/10.1016/j.chest.2022.02.001

Rômulo, J., Vieira, D. C., Barboza, F., & Cajueiro, D. (2025). Towards Fair AI : Mitigating Bias in Credit Decisions — A Systematic Literature Review. *Journal of Risk and Financial Management*, *18*(5), 228. https://doi.org/10.3390/jrfm18050228

Skaiky, A. ali, Ali, H. M. S., Mohammed, A., & Mahdi, Z. A. (2025). Comprehensive Bias Mitigation in AI: Evaluating Pre-Processing, In-Processing, and Post-Processing Techniques for Fair Decision-Making. *IEEE 4th International Conference on Computing and Machine Intelligence (ICMI)*. https://doi.org/10.1109/ICMI65310.2025.11141086

Tang, W., Liu, J., Zhou, Y., & Ding, Z. (2024). Causality-Guided Counterfactual Debiasing for Anomaly Detection of Cyber-Physical Systems. *IEEE Transactions on Industrial Informatics*, *20*(3), 4582–4593. https://doi.org/10.1109/TII.2023.3326544

Trigo, A., Stein, N., & Belfo, F. P. (2024). Strategies to improve fairness in artificial intelligence:A systematic literature review. *Education for Information*, *40*(3), 323–346. https://doi.org/10.3233/EFI-240045

Wan, M., Zha, D., Liu, N., & Zou, N. A. (2023). In-Processing Modeling Techniques for Machine Learning Fairness : A Survey. *ACM Transactions on Knowledge Discovery from Data*, *17*(3), 1–17. https://doi.org/10.1145/3551390

Wang, Y., & Singh, L. (2024). Impact on bias mitigation algorithms to variations in inferred sensitive attribute uncertainty. *Frontiers in Artificial Intelligence*, *8*, 1520330. https://doi.org/10.3389/frai.2025.1520330

Xinying, V. C., & Hooker, J. N. (2023). A guide to formulating fairness in an optimization model. *Annals of Operations Research*, *326*(1), 581–619. https://doi.org/10.1007/s10479-023-05264-y

Zhou, N., Zhang, Z., Nair, V. N., & Singhal, H. (2022). Bias, Fairness and Accountability with Artificial Intelligence and Machine Learning Algorithms. *International Statistical Review*, *90*(144). https://doi.org/10.1111/insr.12492