Int. j. adv. multidisc. res. stud. 2025; 5(6):445-453

Received: 22-09-2025
Accepted: 02-11-2025

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

Analyzing Fairness of Classification Machine Learning Model with Structured Dataset

[1] Ahmed Rashed, [2] Abdelkrim Kallich, [3] Mohamed Eltayeb

[1, 2] Department of Physics, Shippensburg University of Pennsylvania, Franklin Science Center, 1871 Old Main Drive, Pennsylvania, 17257, USA

[3] Islamic University of Madinah, Medina, Al Jamiah, Madinah 42351, Saudi Arabia

[3] University of Khartoum, Khartoum, Al-Nil Avenue, Khartoum 11115, Sudan

DOI: https://doi.org/10.62225/2583049X.2025.5.6.5222

Corresponding Author: **Ahmed Rashed**

Abstract

Machine learning (ML) algorithms have become integral to decision-making in various domains, including healthcare, finance, education, and law enforcement. However, concerns about fairness and bias in these systems pose significant ethical and social challenges.

To evaluate and mitigate biases, three prominent fairness libraries-Fairlearn by Microsoft, AIF360 by IBM, and the What-If-Tool by Google were employed. These libraries provide robust frameworks for analyzing fairness, offering tools to evaluate metrics, visualize results, and implement bias mitigation strategies.

The study aims to evaluate and mitigate biases in a structured dataset using classification models. The main aim of the paper is to present a comparative study for the performance of the mitigation algorithms in two fair-ness libraries by applying them individually one at a time in one of the three stages of the machine learning lifecycle (pre-processing, in-processing, or post-processing), and applying the algorithms in a sequential order in different stages at the time. The findings demonstrate that some sequential order applications enhance the mitigation algorithms performance by reducing bias and maintaining the model performance.

A publicly available dataset from Kaggle was selected for analysis, offering a realistic scenario for evaluating fairness in machine learning workflows.

Keywords: Machine Learning Fairness, Bias Analysis

1. Introduction

Machine learning algorithms are widely used in various domains, including entertain-ment, shopping, healthcare, finance, education, law enforcement, and high-stakes areas like loans [1] and hiring decisions [2, 3]. They provide advantages such as tireless per-formance and the ability to process numerous factors [4, 5]. However, algorithms can also exhibit biases, leading to unfair outcomes [6, 7]. Bias in machine learning can lead to discriminatory outcomes, especially when decisions directly affect individuals or com-munities. Addressing these issues is essential to ensure that machine learning systems operate ethically and equitably. Fairness in decision-making requires the absence of prej-udice or favoritism based on inherent or acquired characteristics, and biased algorithms fail this standard by skewing decisions toward certain groups.

The concept of ”fairness” in algorithmic systems is heavily influenced by the so-ciotechnical context. Various types of fairness-related harms have been identified:

1. **Allocation Harm**: Unfair distribution of opportunities, resources, or information, such as an algorithm selecting men more often than women for job opportunities [8].
2. **Quality-of-Service Harm**: Disproportionate failures affecting certain groups, e.g., facial recognition misclassifying Black women more often than White men [9], or speech recognition underperforming for users with speech disabilities [10].
3. **Stereotyping Harm**: Reinforcement of societal stereotypes, such as image searches for ”CEO” predominantly showing photos of White men [8].
4. **Denigration Harm**: Offensive or derogatory outputs from systems, like misclas-sifying people as gorillas or chatbots using slurs [8].

5. **Representation Harm**: Over- or under-representation of certain groups, e.g., racial bias in welfare fraud investigations or neglect of elderly populations in public-space monitoring [8].
6. **Procedural Harm**: Decision-making practices violating social norms, such as pe-nalizing job applicants for extensive experience or failing to provide transparency, justification, or appeals for algorithmic decisions [11].

These harms often overlap and are not exhaustive, emphasizing the need for careful consideration of fairness from the development stage of algorithmic systems.

Integrating machine learning fairness techniques into a research paper as an indus-try application is essential for advancing the adoption of ethical AI practices. Fairness libraries provide a range of tools to assess and mitigate biases in machine learning mod-els, addressing the growing need for equity as industries increasingly rely on AI-driven decision-making systems. These tools are particularly beneficial in sectors like finance, banking, and healthcare, where fairness is critical. By enabling intuitive and interactive exploration of model behavior, fairness tools empower stakeholders to effectively evalu-ate and address fairness trade-offs. Showcasing the practical applications of these tools bridges the gap between academic innovation and industrial implementation, fostering the development of transparent and equitable AI systems. This paper presents use cases employing three widely trusted fairness libraries—Fairlearn by Microsoft, AIF360 by IBM, and the What-If Tool by Google—to assess and mitigate bias in machine learning models before deployment. This work aims to encourage and guide industry profession-als in incorporating these libraries into their workflows, promoting fairness across diverse applications.

In our study, we conducted a comparative evaluation of two strategies for mitigating bias in machine learning models. We examined the application of individual mitiga-tion algorithms in isolation and compared it to the sequential application of multiple algorithms across different stages of the ML lifecycle: pre-processing, in-processing, and post-processing. The sequential approach is designed to harness the unique strengths of each stage-specific algorithm, providing a more comprehensive solution to addressing bias. To conduct this study, publicly available datasets [12] from Kaggle was selected. Kaggle datasets provide diverse and realistic scenarios for analyzing machine learning models, making them ideal for this type of research. The dataset was preprocessed and used to develop classification models, a common task in machine learning that involves predicting discrete labels based on input features. Classification problems are partic-ularly relevant for fairness studies because biased predictions can disproportionately impact specific groups.

To evaluate and mitigate potential biases, three state-of-the-art fairness libraries were employed: Fairlearn by Microsoft, AIF360 by IBM, and the What-If Tool by Google. These libraries provide comprehensive toolsets for fairness analysis, including metrics to assess fairness, visualizations to interpret model behavior, and algorithms to mitigate bias. By leveraging these libraries, this research systematically evaluates the fairness of classification models and explores techniques to reduce bias in their predictions. The same fairness analysis was done for unstructured datasets with computer vision and natural language processing models in our paper [44].

While previous studies have primarily addressed fairness interventions at isolated stages of the ML lifecycle, our research advances the field of AI by introducing a se-quential approach that integrates fairness interventions across all three stages of the ML lifecycle. This approach provides a comprehensive framework for enhancing bias mitiga-tion, ensuring a more holistic treatment of biases, reducing the propagation of fairness issues during model development, and minimizing residual disparities that may remain when only single-stage interventions are employed [45, 46]. By applying this lifecycle-based framework to real-world datasets, our study offers robust empirical evidence of its effectiveness. This work bridges the gap between theoretical fairness concepts and their practical application, promoting wider adoption of lifecycle-based fairness methodologies [47]. The objectives of this research are:

1. The central objective is to demonstrate how a sequential application of fairness mitigation algorithms across the ML lifecycle stages, pre-processing, in-processing, and post-processing, leads to superior mitigation of biases compared to applying these methods in isolation.
2. To assess the extent of bias present in machine learning models trained on the selected structured dataset.
3. To compare the performance and effectiveness of the three fairness libraries in identifying and mitigating bias.
4. To provide actionable insights into the application of fairness tools in real-world machine learning workflows.

The remainder of this paper is organized as follows: Section 2 provides a review of related work on machine learning fairness. Section 3 describes the methodology, including the dataset, preprocessing steps, and model development process. Section 4 discusses the implementation of fairness analyses using the selected libraries and their capabilities. Section 5 presents a comparative analysis and results of the libraries. Finally, Section 6 concludes with a summary of findings.

2. Review of Related Work

Bias in machine learning (ML) models has been a growing area of concern, particularly as these models increasingly impact critical societal domains such as healthcare, hiring, and criminal justice. Numerous studies have explored the origins, manifestations, and mitigation strategies of bias, providing a comprehensive foundation for understanding and addressing this pervasive issue.

One key area of research focuses on identifying and characterizing biases in machine learning models. Ref. [13] provide a broad taxonomy of biases, categorizing them into historical, representation, and measurement biases. Historical bias originates from inequities in the data itself, even before ML techniques are applied. Representation bias emerges when certain groups are under- or over-represented in the training data, leading to skewed model predictions [14]. Measurement bias arises when the features or labels used for training do not accurately reflect the target variable due to flawed measurement processes.

Another stream of work has delved into bias detection methods. Techniques such as disparate impact analysis [15] and fairness metrics like demographic parity, equal op-portunity, and disparate mistreatment [16] have become standard tools. For structured datasets, researchers often

focus on quantifying group fairness and individual fairness. Group fairness ensures equitable treatment across predefined demographic groups, while individual fairness emphasizes treating similar individuals similarly [17]. Ref. [18] dis-cusses how these fairness definitions often conflict, necessitating trade-offs tailored to specific applications.

The literature also emphasizes the technical challenges of mitigating bias. One pop-ular approach involves pre-processing techniques to address biases within the dataset itself. For example, Ref. [19] proposes re-weighting data samples or modifying labels to ensure fairness before model training. In-processing methods, such as adversarial debias-ing [20], introduce fairness constraints directly into the training process. Post-processing techniques adjust the model outputs to achieve fairness metrics, such as the re-ranking methods proposed by [16]. However, Ref. [21] highlights the inherent trade-offs between fairness metrics, illustrating that achieving fairness often requires sacrificing accuracy.

Bias analysis in structured datasets, specifically, has garnered attention due to the widespread use of tabular data in decision-making systems. Structured datasets often carry latent biases stemming from historical inequities in human decision-making or sys-temic discrimination. The COMPAS dataset, used in criminal justice, exemplifies these challenges, with studies showing racial disparities in predictive outcomes [22]. Research on structured datasets also highlights the role of feature selection and data preprocess-ing in amplifying or mitigating biases. Ref. [23] examines how feature correlation with sensitive attributes impacts fairness, proposing strategies for disentangling these rela-tionships.

Recent work has explored interpretability and its role in bias analysis. Ref. [24] in-troduced LIME (Local Interpretable Model-agnostic Explanations) to help stakeholders understand model predictions, aiding the detection of biased decision-making patterns. In Ref. [25] further developed SHAP (SHapley Additive exPlanations), which provides consistent and locally accurate feature importance values. These tools have been instru-mental in identifying bias within structured datasets, as they enable granular analyses of how individual features contribute to unfair predictions.

Additionally, researchers are increasingly incorporating intersectionality into bias studies. Ref. [26] emphasized the importance of evaluating models across multiple demographic axes, demonstrating how performance disparities can compound for inter-sectional groups, such as Black women in facial recognition systems. For structured datasets, studies by [27] propose fairness-enhancing interventions that consider multiple subgroups simultaneously, avoiding the pitfalls of single-axis fairness analysis.

The literature on bias in machine learning models spans a wide range of topics, from foundational definitions and detection methods to mitigation strategies and inter-pretability tools. While significant progress has been made, challenges remain in apply-ing these techniques to structured datasets, particularly in balancing fairness with other competing objectives such as accuracy and interpretability. This review underscores the importance of continued research into holistic and context-sensitive approaches for analyzing and mitigating bias in machine learning models.

3. Dataset and Model Details

We used two datasets in this work to generalize our conclusion as much as possible. The first data is a loan dataset [12]. The dataset is designed to automate the real-time loan eligibility process based on customer details provided during the online application form submission. These details include gender, marital status, education, number of dependents, income, loan amount, credit history, and other relevant factors. The primary objective is to determine eligibility for granting a home loan (Yes/No) by predicting loan eligibility based on the provided information. Gender is considered a sensitive feature in the analysis. The dataset comprises 614 samples with a total of 11 features.

The second dataset is a job application one [12]. The "Employability Classification of Over 70,000 Job Applicants" dataset provides detailed information about job applicants and their employability scores, aiming to assist organizations in evaluating candidates for various employment opportunities. Leveraging machine learning techniques, it offers insights into factors influencing employability, enhancing the efficiency of the hiring pro-cess. The dataset, compiled from job portals, career fairs, and online applications, spans diverse industries, roles, and qualifications, ensuring broad applicability. It includes features such as age, education level, gender, professional experience, coding expertise, previous salary, and computer skills, with "Employed" serving as the target variable in-dicating whether the applicant was hired. This structured dataset represents applicants through rows, with attributes organized into columns, making it a valuable resource for analyzing employability trends.

To process the datasets, categorical variables were encoded using techniques such as one-hot encoding, and continuous variables were normalized to ensure compatibility with the machine learning models. The dataset was split into training and testing subsets using an 80-20 split to evaluate model performance.

For the model, the pipeline employed several machine learning algorithms, including Logistic Regression, Decision Trees, CatBoost, and Gradient Boosting. CatBoost [48] emerged as the most effective algorithm for the classification task. Hyperparameter tuning was performed using grid search to optimize the model's performance.

The study evaluated the model using standard classification metrics such as accuracy, precision, recall, and F1-score. In addition, fairness metrics such as demographic parity and disparate impact were calculated to analyze potential biases. Results highlighted disparities in prediction accuracy across demographic groups, with notable differences between different subgroups in the sensitive features. These findings underscored the im-portance of incorporating fairness evaluations into traditional performance assessments for machine learning models.

4. Implementation of Fairness Analyses

In recent years, the increasing reliance on machine learning (ML) in various sectors has led to growing concerns over fairness and bias in classification models. As these models can significantly influence decision-making processes in critical areas such as healthcare, finance, and criminal justice [28], ensuring their fairness has become imperative. Bias in ML models can manifest due to various factors, including skewed training data, model selection, and

underlying societal biases, ultimately leading to discriminatory outcomes against marginalized groups [29].

To address these challenges, several libraries and tools have been developed to assist practitioners in analyzing and mitigating bias in their models. Among them, Fairlearn, AIF360, and What-If Tool stand out as comprehensive resources that offer unique func-tionalities for fairness evaluation and enhancement.

- **Fairlearn** [30]: Developed by Microsoft, this toolkit helps data scientists assess and improve the fairness of their AI models by providing a suite of metrics to evaluate fairness and algorithms for mitigating unfairness. Fairlearn emphasizes the importance of both social and technical dimensions of fairness in AI systems. It facilitates the understanding of how different aspects of a model contribute to disparities among groups.

In this analysis, we will use the fairness metric Demographic Parity Difference (DPD) that measures the disparity between the selection rates of two or more groups. It's calculated by finding the difference between the largest and smallest group-level selection rate. A lower value indicates less disparity.

Different mitigation algorithms will be used to mitigate the bias in the model. There are algorithms work on different ML life cycle stages such as pre-processing, in-processing, and post-processing levels. The best algorithms that will work good for our use cases are Exponentiation Gradient and Threshold Optimizer.

The Exponentiated Gradient (EG) is designed to reduce unfairness in machine learning models by framing the problem as a constrained optimization task. The algorithm seeks to minimize loss (maximize predictive accuracy) while satisfying constraints related to fairness. Exponentiated Gradient works by iteratively finding a weighted combination of models (or classifiers) that achieves the best trade-off between accuracy and fairness constraints. This combination forms a probabilis-tic ensemble, where models are assigned weights using an exponentiated updates scheme. The ensemble is then used for predictions. The fairness constraints are typically defined in terms of statistical fairness metrics, such as demographic par-ity, equalized odds, or disparate impact. These constraints are enforced within a specified tolerance level.

The Exponentiated Gradient method addresses the following optimization problem [37]:

$$\min_{\theta \in \Theta} \mathbb{E}_{h \sim \theta}[L(\hat{h})]$$

$$\text{subject to: } \mathbb{E}_{\hat{h} \sim \theta}[g_i(\hat{h})] \leq \epsilon, \quad \forall i \in \{1, \ldots, m\},$$

Where Θ is the set of all possible classifiers, $L(h\hat{})$ is the loss function measuring predictive performance (e.g., log-loss or mean squared error), $g_i(h\hat{})$ are the fair-ness constraint functions, quantifying the extent to which fairness conditions (e.g., equal opportunity) are violated for the i-th constraint, ϵ is the allowed constraint violation margin, and $h\hat{}$ is a hypothesis (or classifier).

The Threshold Optimizer in the Fairlearn library is a fairness mitigation algorithm designed to adjust decision thresholds of a pre-trained model to satisfy fairness con-straints. Instead of retraining the model, it modifies how the model's predictions are converted into decisions, making it efficient and easy to integrate into existing workflows.

Threshold Optimizer takes the predicted scores from a model and applies group-specific thresholds to ensure that fairness constraints are met. This approach is particularly useful for binary classification tasks, where decisions are based on whether a prediction exceeds a threshold. The algorithm assigns different thresh-olds for different demographic groups, balancing fairness and predictive perfor-mance.

Fairness constraints, such as demographic parity or equalized odds, are specified, and the algorithm ensures that the decision-making process respects these con-straints. The optimization minimizes a loss function (e.g., error rate) while satis-fying fairness conditions.

Threshold Optimizer solves the following constrained optimization problem for a binary classification setting [38, 39]:

$$\min_{T} \mathbb{E}[L(Y, \hat{Y_T})],$$

$$\text{subject to: } g_i(T) \leq \epsilon, \quad \forall i \in \{1, \ldots, m\},$$

Where $T = \{T_g\}_{g \in G}$ is the set of thresholds, one for each demographic group $g \in G$, $L(Y, \hat{Y_T})$ is the loss function, comparing true labels Y and predicted labels $\hat{Y_T}$ derived using thresholds T, $g_i(T)$ are fairness constraint functions that quantify fairness violations (e.g., difference in true positive rates between groups), and ϵ is the tolerance for constraint violations. By adjusting thresholds instead of retrain-ing, Threshold Optimizer provides a straightforward and computationally efficient method to ensure fairness in decision-making.

- **AIF360** [31]: This comprehensive library offers a wide range of metrics for as-sessing fairness and techniques for mitigating bias across the entire AI application lifecycle. Developed by IBM, AIF360 includes methods that can be integrated into different stages of the machine learning pipeline to facilitate fairness-aware modeling. It includes metrics for evaluating fairness across different societal de-mographics and offers re-parameterization strategies to improve model robustness.

  Two successful mitigation algorithms will provide very good results in reducing the bias in the ML model with maintaining the metric chosen to measure the perfor-mance which is Average Odds Difference (AOD). These algorithms are Reweighing and Equalized Odds.

  Reweighing is a preprocessing technique that assigns weights to the data instances to reduce biases associated with sensitive attributes, such as race, gender, or age. The reweighing process ensures that different groups are treated fairly in terms of representation when training a machine learning model [40].

  The main idea of Reweighing is to balance the dataset so that the proportion of favorable and unfavorable outcomes is equal across different demographic groups defined by the sensitive attribute. The algorithm does this by computing instance weights based on the joint distribution of sensitive attributes and class labels. These weights are then applied to the training dataset, allowing the model to learn a more unbiased representation.

Reweighing can mitigate fairness issues such as disparate impact or statistical parity, depending on the fairness metric being addressed. This method does not modify the feature values or the labels but adjusts their importance in the training process.

Equalized Odds is a fairness mitigation method that ensures a model's predictions satisfy the fairness criterion known as equalized odds. This criterion requires that the prediction outcomes be independent of sensitive attributes, conditional on the true outcome. In other words, the model should have the same true positive rate (TPR) and false positive rate (FPR) across all groups defined by the sensitive attribute.

The Equalized Odds approach modifies the predictions of a classifier to ensure that the TPR and FPR are approximately equal across groups [41, 42, 43]. It achieves this by adjusting decision thresholds for different groups or directly post-processing the predictions. This technique is a post-processing method, meaning it does not alter the original model but modifies its outputs to improve fairness. The algorithm is useful in applications where fairness is critical and ensures that predictions are fair across groups while maintaining as much accuracy as possible.

- **What-If-Tool** [32]: Created by Google, this interactive visualization tool allows users to explore and analyze machine learning models without requiring any cod-ing. It supports performance testing in hypothetical scenarios, facilitating the understanding and explanation of model behavior. It enables users to observe model performance across different demographics and explore various "what-if" scenarios. It supports users in understanding how changes in input features affect model predictions, thus empowering them to conduct deeper bias analysis.

Together, these tools are essential for researchers aiming to understand and mitigate bias in machine learning classification models, equipping them with the methodologies to ensure equitable AI systems and informed decision-making processes.

5. Results and Discussions

In this section, we present and analyze the results obtained from each fairness library. The primary objective of our analysis is to conduct a comparative study on the effective-ness of applying individual mitigation algorithms versus applying multiple algorithms sequentially across the three stages of the ML lifecycle. The results demonstrate that, in some cases, sequential applications yield better outcomes compared to individual ap-plications, while in others, they perform worse. These outcomes are evaluated based on their ability to improve fairness metrics while maintaining or enhancing performance metrics. The details for each library are provided below. The codes and details can be found in [33]. The sensitive feature in the loan and job application datasets that we measured the bias in is gender, and we used the CatBoost algorithm [48] in the classifi-cation model. The model of the loan dataset predicts the eligibility of the applicant for a loan, and the model of the job applicant dataset predicts the eligibility of employment. For both models, we use the accuracy as the model performance metric. We investigate the model if it is biased against any group (male/female) in gender.

For Fairlearn, we use demographic parity difference as a fairness metric to evaluate the bias in the machine learning model. To address and mitigate the detected biases, we applied mitigation algorithms at three stages of the machine learning pipeline: prepro-cessing, in-processing, and postprocessing. Preprocessing techniques involved modifying the training data to reduce bias before feeding it into the model. For example, sensitive features in a dataset may be correlated with non-sensitive features. The Correlation Re-mover addresses this by eliminating these correlations, while preserving as much of the original data as possible, as evaluated by the least-squares error. In-processing methods integrated fairness constraints into the model training process, with approaches such as exponentiated gradient ensuring that the model learned fairer decision boundaries. Postprocessing focused on adjusting the predictions after model training, ensuring that the final outputs adhered to fairness criteria without retraining the model such as thresh-old optimizer which is built to satisfy the specified fairness criteria exactly and with no remaining disparity [34, 35, 36].

Beyond these individual techniques, we also explored the use of combined mitigation approaches, where two algorithms were applied in series to enhance fairness outcomes. For instance, preprocessing adjustments were complemented by postprocessing tweaks, leading to improved alignment with both accuracy and fairness objectives. This com-bined approach aimed to leverage the strengths of each mitigation stage to produce more equitable and reliable model predictions.

The evaluation process identify the best algorithm/s as the one that achieve a dual objective: minimizing bias metrics like demographic parity difference while maintaining performance metrics such as accuracy, precision, and recall. Our findings highlight the importance of an integrated approach to fairness, where multiple strategies are utilized in conjunction to address complex biases inherent in structured datasets.

Fig 1 presents the results of fairlearn applied to the two datasets. For the loan dataset, the baseline model, prior to implementing any mitigation algorithm, achieved a performance metric (Accuracy) of 0.7818 and a fairness metric (Demographic Parity Difference) of 0.0672. For the job application dataset, the baseline model, prior to imple-menting any mitigation algorithm, achieved a performance metric (Accuracy) of 0.8378 and a fairness metric (Demographic Parity Difference) of 0.1046. Ideally, the accuracy would be 1.00, and the demographic parity difference would be 0.00. To address bias, we applied mitigation algorithms both individually and sequentially, using two different algorithms at distinct stages of the machine learning lifecycle. For the loan dataset, the best result is obtained by the exponential gradient algorithm, which maintained an accuracy of 0.7638 and reduced the demographic parity difference to 3.17%. For the job applicants dataset, the best result is obtained by the sequential algorithms Correlation Remover + Threshold Optimizer, which maintained an accuracy of 0.8369 and reduced the demographic parity difference to 0.38%.

Notably, overall performance of some individual application of the mitigation algo-rithms is better than the sequential algorithms as we see from Fig 1 plots. Both exponentiated gradient and threshold optimizer reduced the

bias in the models and maintained their performance. On the other hand, there is no sequential algorithms hold a good performance across the two datasets.

For AIF360, we employ both individual mitigation algorithms as well as two distinct mitigation algorithms within each stage of the machine learning pipeline to comprehen-sively address bias. We use the fairness metrics Statistical Parity Difference, Average Odds Difference, Equal Opportunity Difference, Theil Index, and Generalized Entropy Index. We employ the mitigation algorithms Reweighing, Disparate Impact Remover, Adversarial Debiasing, and Calibrated EqOdds.

In the preprocessing stage, Reweighing assigns weights to instances based on their representation in different demographic groups. This approach ensured a balanced dis-tribution of data, directly addressing biases embedded in the training dataset.

In the post-processing stage, Equalized Odds imposes constraints during model train-ing to ensure that predictive outcomes were not disproportionately distributed across sensitive attributes such as race or gender. By enforcing parity in true positive and false positive rates, Equalized Odds can enhance fairness without significantly compromising model performance.

By looking at the two plots in Fig 2, one can recognize that the mitigation algo-rithms are not effective in reducing bias in the models as much as the ones of the Fairlearn library as shown in Fig 1. Fig 2 shows a pattern for the impact of the mitigation algorithm applications to the model performance both in individual and sequential order. The best algorithms in maintaining the model performance are Reweighing, Disparate Impact Remover, and Calibrated EqOdds in individual order, as well as Reweighing + Calibrated EqOdds in sequential order where the accuracy of the original model is 78.18% for the loan dataset and 83.78% for the job applicant dataset and after applying these mitigation algorithms the accuracy did not change much. Regarding the bias in the models, both individual and sequential algorithms do not show much impact in re-ducing bias. On the contrary, some individual and sequential algorithms have increased the bias in the models.

Our findings underscore the importance of selecting appropriate mitigation strategies tailored to specific stages of the machine learning pipeline. By leveraging the effective mitigation algorithms, we demonstrated that it is possible to achieve a balanced trade-off between fairness and accuracy, highlighting the potential of integrated approaches to bias mitigation in structured datasets.

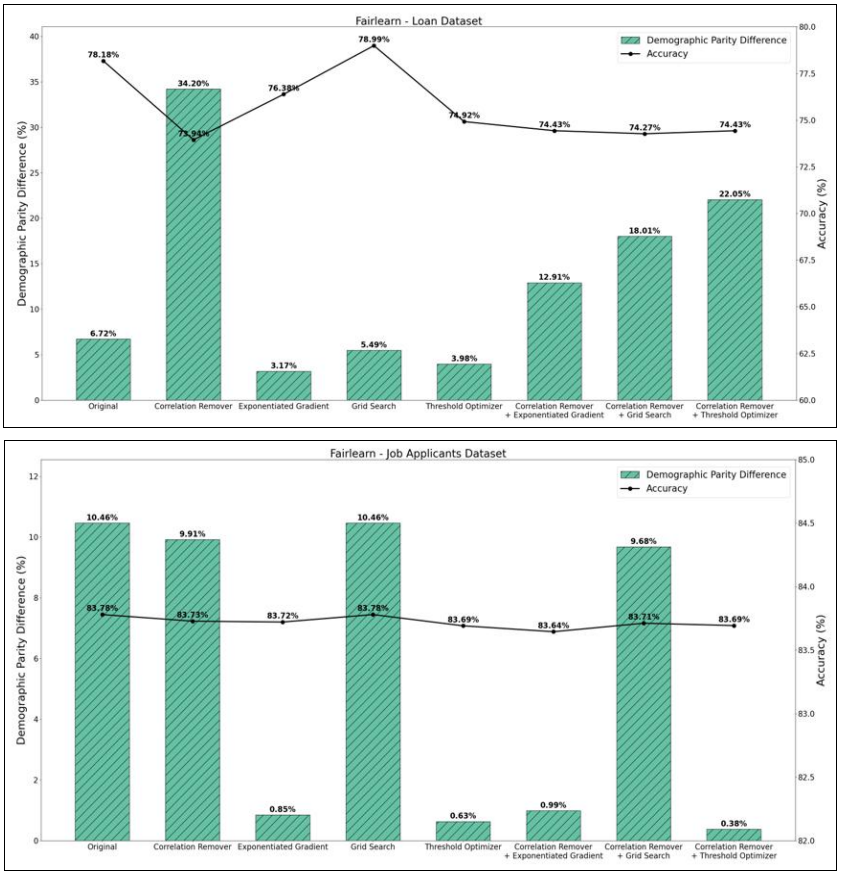

Fig 1: The results of applying Fairlearn to loan and job application datasets with classification model after applying the mitigation algorithms one at a time and in sequential order

In the What-If Tool, there is no mitigation algorithms to reduce the model bias. Therefore, we do not discuss this library here in the paper for comparing between mit-igation algorithms. Instead we demonstrate that the bias may change by adjusting the threshold for the labeled class that impacts both the model's performance and its bias metrics. The optimal thresholds identified for this model were 0.2 and 0.4. At these thresholds, the model performance metric (Accuracy) increased from 0.33 to an average of 0.62, representing a 29% improvement, while the bias metric (Demographic Parity Difference) decreased from 0.19 to 0.01, reflecting an 18% reduction. These results are summarized in Table 3.

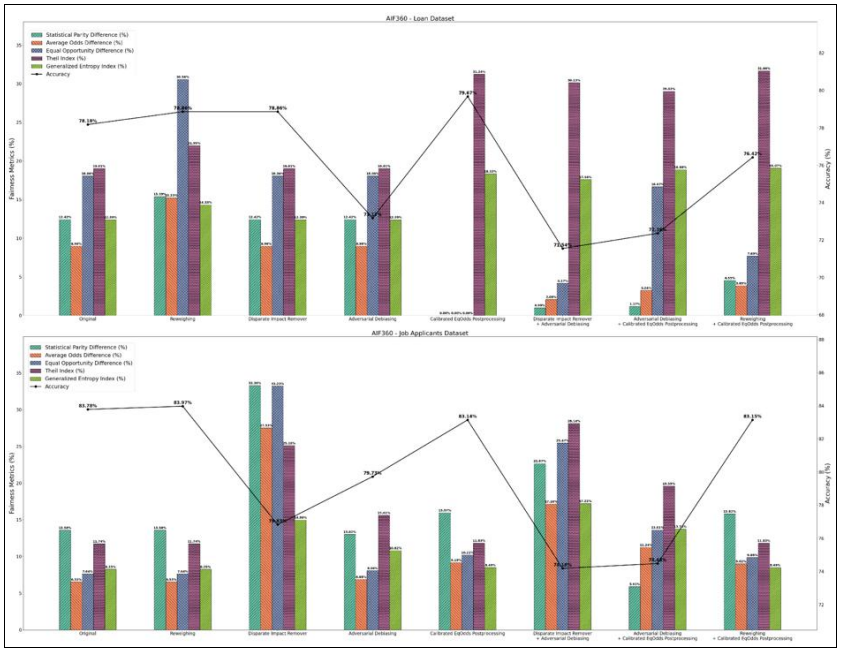

Fig 2: The results of applying AIF360 to loan and job application datasets with classification model after applying the mitigation algorithms one at a time and in sequential order

6. Conclusion

This work examined the fairness of machine learning models with classification tasks using structured datasets, focusing on how biased predictions can reinforce systemic inequalities. Kaggle datasets were analyzed to provide the model fairness, utilizing two fairness libraries Fairlearn (Microsoft) and AIF360 (IBM) to evaluate and mitigate bias. We discussed a comparative study of applying the mitigation algorithms of these libraries individually one at a time in one of the ML stages including pre-processing, in-processing, and post-processing versus applying the mitigation algorithms in a sequential order at more than one stages at the same time.

For the Fairlearn library, we observed that applying the mitigation algorithms both:

Table 1: The results by applying What-If-Tool library to the classification model

Thresholds	Demographic Parity Male (DPM)	Demographic Parity Female (DPF)	Performance Change Using Average Accuracy Score	Fairness Improvement Using Demographic Parity Diff = DPM - DPF
Baseline 0	0.303	0.112	0.33	0.19
0.2	0.566	0.551	0.63 30% better	0.01 18% better
0.4	0.628	0.611	0.61 28% better	0.01 18% better
0.6	0.655	0.627	0.55 22% better	0.02 17% better
0.8	0.716	0.686	0.50 17% better	0.03 16% better
0.9	0.702	0.689	0.45 12% better	0.01 18% better
1.0	0	0	0 33% worse	0 19% better

individually and in a sequential order have a good power in reducing the bias in the model. Some individual applications such as exponentiated gradient and threshold op-timizer showed better performance in reducing bias and maintaining the model per-formance over the sequential application of the algorithms. However, no sequential algorithm consistently maintained high performance across both datasets.

In contrast, the AIF360 library's mitigation algorithms showed less effectiveness in reducing bias while maintaining model performance. Sequential algorithms in AIF360 had minimal impact on bias reduction and, in some cases, increased bias in the models. Overall, the study demonstrated that Fairlearn algorithms outperform those in AIF360 in balancing fairness and model performance.

7. Acknowledgement

We would like to thank Yale Center for Research Computing for supporting A.K. through the CAREERS project administered by the PSU Institute for Computational and Data Sciences (ICDS) under the National Science Foundation Award with No. 2018873.

8. References

1. Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, Amrit P Mathur. Multi-objective evolutionary

algorithms for the risk-return trade-off in bank loan man-agement. Int. Trans. Oper. Res. 2002; 9(5):583-597.
2. Miranda Bogen, Aaron Rieke. HelpWanted: An Examination of Hiring Algorithms, Equity and Bias. Technical Report. Upturn, 2018.
3. Lee Cohen, Zachary C Lipton, Yishay Mansour. Efficient candidate screening under multiple tests and implications for fairness, 2019. arXiv:cs.LG/1905.11361
4. Shai Danziger, Jonathan Levav, Liora Avnaim-Pesso. Extraneous factors in judicial decisions. Proc. Nat. Acad. Sci. 2011; 108(17):6889-6892.
5. Anne O'Keeffe, Michael McCarthy. The Routledge Handbook of Corpus Linguistics. Routledge, 2010.
6. Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's bi-ased against blacks, 2019. https://www.propublica.org/article/machine-biasrisk-assessments-in-criminal-sentencing.
7. Cathy O'Neil. Weapons of Math Destruction: How Big Data Increases Inequal-ity and Threatens Democracy. Crown Publishing Group, New York, NY, 2016.
8. Madaio MA, Stark L, Wortman Vaughan J, Wallach H. Co-Designing Check-lists to Understand Organizational Challenges and Opportunities around Fairness in AI. Chi 2020, 2020, 1-14.
9. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and trans-parency, 2018, 77-91.
10. Guo A, Kamar E, Vaughan JW, Wallach HM, Morris MR. Toward fairness in AI for people with disabilities: A research roadmap. CoRR, abs/1907.02227, 2019. URL http://arxiv.org/abs/1907.02227
11. Rudin C, Wang C, Coker B. The age of secrecy and unfairness in recidivism prediction, 2018, 1-46. URL http://arxiv.org/abs/1811.00731
12. Datasets used in the study: https://www.kaggle.com/datasets/shaijudatascience/ loan-prediction-practice-av-competition, https://www.kaggle.com/datasets/ ayushtankha/70k-job-applicants-data-human-resource
13. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys. 2021; 54(6):1-35.
14. Suresh H, Guttag JV. A Framework for Understanding Unintended Consequences of Machine Learning. Communications of the ACM. 2021; 64(8):62-71.
15. Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and Removing Disparate Impact. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2015.
16. Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS), 2016.
17. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through Awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012.
18. Binns R. Fairness in Machine Learning: Lessons from Political Philoso-phy. Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT), 2018.
19. Kamiran F, Calders T. Data Preprocessing Techniques for Classification without Discrimination. Knowledge and Information Systems. 2012; 33(1):1-33.
20. Zhang BH, Lemoine B, Mitchell M. Mitigating Unwanted Biases with Adversarial Learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2018.
21. Chouldechova A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data. 2017; 5(2):153-163.
22. Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias. ProPublica, 2016.
23. Xu D, Yuan S, Zhang L, Wu X. FairGAN: Fairness-aware Genera-tive Adversarial Networks. Proceedings of the 2020 International Joint Conference on Artificial Intelligence (IJCAI), 2020.
24. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016.
25. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Pre-dictions. Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), 2017.
26. Buolamwini J, Gebru T. Gender Shades: Intersectional Accuracy Dis-parities in Commercial Gender Classification. Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT), 2018.
27. Kearns M, Neel S, Roth A, Wu ZS. Preventing Fairness Ger-rymandering: Auditing and Learning for Subgroup Fairness. Proceedings of the 35th International Conference on Machine Learning (ICML), 2018.
28. Barocas S, Hardt M, Narayanan A. Fairness and Accountability in Machine Learning. Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 2019.
29. Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias: There's Software Used Across the Country to Predict Future Criminals and it's Biased Against Blacks. ProPublica, 2016.
30. Fairlearn by Microsoft, n.d. Retrieved from: https://fairlearn.org/
31. AIF360 by IBM, n.d. Retrieved from: https://aif360.mybluemix.net/
32. What-If Tool by Google, n.d. Retrieved from: https://github.com/google/tf-what-if
33. Github of the project: https://github.com/mohammad2012191/Fairness-in-Machine- Learning-Identifying-and-Mitigation-of-Bias/
34. Moritz Hardt, Eric Price, Nati Srebro. Equality of opportunity in supervised learning. In NeurIPS, 2016, 3315-3323. URL: https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html
35. Hilde Weerts, Lamb`er Royakkers, Mykola Pechenizkiy. Does the end justify the means? On the moral justification of fairness-aware machine learning, 2022. arXiv preprint arXiv:2202.08536
36. Brent Mittelstadt, Sandra Wachter, Chris Russell. The

unfairness of fair ma-chine learning: Levelling down and strict egalitarianism by default, 2023. arXiv preprint arXiv:2302.02404
37. Alekh Agarwal, Alina Beygelzimer, Miroslav Dud´ık, John Langford, Hanna Wallach. A Reductions Approach to Fair Classification, 2018. arXiv:1803.02453
38. Hilde Weerts, Lamb`er Royakkers, Mykola Pechenizkiy. Does the end justify the means? on the moral justification of fairness-aware machine learning, 2022. arXiv preprint arXiv:2202.08536
39. Brent Mittelstadt, Sandra Wachter, Chris Russell. The unfairness of fair ma-chine learning: levelling down and strict egalitarianism by default, 2023. arXiv preprint arXiv:2302.02404
40. Kamiran F, Calders T. Data Preprocessing Techniques for Classification with-out Discrimination. Knowledge and Information Systems, 2012.
41. Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ. On Fairness and Calibration. Conference on Neural Information Processing Systems, 2017.
42. Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learn-ing. Conference on Neural Information Processing Systems, 2016.
43. Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ. On Fairness and Calibration. Conference on Neural Information Processing Systems, 2017.
44. Ahmed Rashed, Abdelkrim Kallich, Mohamed Eltayeb. Analyzing Fairness of Computer Vision and Natural Language Processing Models. arXiv:2412.09900 [cs.LG].
45. Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Em-beddings, arXiv:1607.06520
46. Denis Rystsov. CASPaxos: Replicated State Machines without logs. arXiv:1802.07000
47. Greenlees JPC. The Balmer spectrum of rational equivariant cohomology theo-ries. arXiv:1706.07868
48. Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Doro-gush, Andrey Gulin. CatBoost: Unbiased boosting with categorical features. arXiv:1706.09516v5 [cs.LG]