Ethical AI with Balancing Bias Mitigation and Fairness in Machine Learning Models

Khalida Walid Nathim
Alnoor University
Nineveh, Iraq
khalida.walid@alnoor.edu.iq

Nada Abdulkareem Hameed
Al Mansour University College
Baghdad, Iraq
nada.abdulkarim@muc.edu.iq

Saja Abdulfattah Salih
Al Hikma University College
Baghdad, Iraq
Saja_872008@yahoo.com

Nada Adnan Taher
Al-Rafidain University College
Baghdad, Iraq
nada.taher@ruc.edu.iq

Hayder Mahmood Salman
Al-Turath University
Baghdad, Iraq
haider.mahmood@turath.edu.iq

Dmytro Chornomordenko
National University of Life and Environmental Sciences of Ukraine
Kyiv, Ukraine
d.chornomordenko@nubip.edu.ua

Abstract — **The rapid integration of Artificial Intelligence (AI) into critical domains such as healthcare, finance, and criminal justice has raised significant ethical concerns, particularly around bias and fairness in machine learning models. Despite their potential for improving decision-making processes, these models can perpetuate or even exacerbate existing societal biases. This study aims to investigate approaches to bias mitigation in AI systems, focusing on balancing fairness and performance. A systematic review of 150 research articles published between 2018 and 2023 was conducted, along with experiments on 25 benchmark datasets to evaluate various machine learning algorithms and bias mitigation techniques. Results showed a 23% reduction in bias and an average 17% improvement in nine fairness metrics during model training, though at the cost of up to 9% in overall accuracy. The study highlights the trade-offs between fairness and performance, suggesting that creating AI systems that are both fair and effective remains an ongoing challenge. The findings underscore the need for adaptive frameworks that address bias without significantly compromising model performance. Future research should explore domain-specific adaptations and scalable solutions for integrating fairness throughout the AI development process to ensure more equitable outcomes.**

I. INTRODUCTION

The widespread adoption of Artificial Intelligence (AI) in domains such as healthcare, finance, and criminal justice has raised substantial ethical concerns around bias and fairness for machine-learning algorithms. As many AI systems, which drive operational efficiency and decision support processes at scale, become default ways of working — this makes them a veritable minefield for more slowly moving legacy organizations to traverse unscathed (not least because they typically have all the built-in biases in their training data that you might expect from society.) These systemic biases, many of which come from historical prejudices and inequalities [1], may result in discriminatory actions toward different subgroups within today's diverse societies.

With the growing extent to which AI is used in steering serious decisions, there is an urgent need of reliable methods that can guarantee fairness and justice within these systems. Metrics that have been traditionally used for determining model performance as an accuracy (how often the classifier was correct), such metrics don't tell us how fair these guidelines are or if biased outcomes were produced after going through our algorithms. Therefore, it is exactly this deficiency that enforces the development of holistic frameworks addressing not only bias but also equity on a diverse set by means providing higher levels of trust and fairness in AI based decision making [2].

In the last few years, there has been some scholarly effort in designing different methods to mitigate bias by reweighting data, modifying algorithms, and injecting fairness constraints into model training. While these solutions are promising for minimizing bias, they also come with many complications especially about the balance that may need to be made in improving model performance draw-outs. For instance — bias mitigation in the interest of bolstering fairness could cause reductions in model accuracy/performance thereby muddying up an already tough ethical vs practical dichotomy [3]. These trade-offs illustrate the fact that it is much easier said than produce AI systems that are fair as well as effective and just promote more research in this domain for further refining these methodologies.

There are no universal methods for assessing whether an AI model is fair, which makes it even harder to make sure that the resulting system would be free from bias. Without any solid consensus on what the "right" metrics should be, or even being able to agree on a clear definition of fairness quantified by a well-defined set of criteria, it is quite difficult to evaluate how effective different remediation strategies are concerning bias and determine which models might need improvement or else making mindful decisions about where they belong in society. This hole in the current research is a serious limitation for researchers and entities who want to mitigate bias within AI

algorithms, systematically [4]. Thus, this article seeks to add to the conversation by analyzing current bias mitigation techniques and presenting a systematic framework that reduces sex, age, geography, and race discrimination without trading off bias reduction too much.

We must uncover the source and consequences of bias to properly tackle this challenge. Bias can enter the modeling process in any number of ways — at collection time, during training, or in deployment. These stages provide an opportunity to address sets of unique hurdles that need specific mitigation solutions. Biased data, affecting the algorithm, such as exists in hiring or medical treatment can insert bias into algorithms that try to use historical information, and biased algorithms themselves are likewise problematic because they will generate adverse impacts even if an underlying population is balanced [5]. Developing patient-reported solutions for chronic disease requires a nuanced understanding of these biases underpinning the system — not just symptoms, but root causes.

In addition to the technical barriers, bias is an ethical and societal issue. Unfairly biased AI systems cause distress in our society and consequently depress public trust of further broader utilization, mismatching the desired effect from those communities affected by them. These multiples have led to numerous calls for transparency, accountability, and inclusivity in the processes of AI systems development and deployment [6]. In the direction of reducing biases and providing equal treatment for everyone including stakeholders in AI, these are some points that need to be taken care.

As bias and fairness emerge as major issues in AI, it is more important than ever to have a comprehensive method to mitigate bias while continuing with practical considerations of machine learning. In this vein, the article aims to shed some light on ongoing research about methods that may bolster fairness in AI applications regardless of whether they have a basis in causal reasoning, whilst also providing context for the challenges associated with such strategies and advancing them as part of a framework towards fair & accurate intelligence. This is part of the ongoing exploration to help the ethics in AI community learn training, testing, and benchmarking best practices for creating more equitable outcomes for all populations while improving trustworthiness in decision-making.

A. Study Objective

The article aims to investigates the ways to strike a tradeoff between reducing bias versus increasing fairness in machine learning models, that are more widespread than ever across domains such as healthcare, finance and criminal justice. When used to make important life-changing decisions, the existence of bias in AI systems may result in unfair and potentially damaging results. This article tries to give a general view on what are the systematic manners that bias mitigation techniques or fairness measures (or related definitions) around AI models throughout these papers we have seen, and where is their strengths as well limitations. This study hopes to provide insights into the effects of these techniques on modeling performance by delving through data from various studies and datasets, among them are: efficiency improvements brought about by reduced bias and increased fairness, and accuracy trade-offs that come with those changes. Ultimately the main aim is to work towards contributing tools, frameworks, or guidelines that can be used by practitioners and researchers alike in designing machine learning models for both fairness (minimizing bias) and accuracy – ensuring we don't throw away the baby with bath water. Work like this is critical to creating transparency and equity in AI decision processes.

B. Problem Statement

The rapid rise of AI in decision-making from healthcare field, finance and criminal justice has raised more ethical concerns when it interacts with machine learning models, bias, and fairness. While the technology can improve efficiency and accuracy, it could also exacerbate some of those same issues if deployed improperly. This boils down to one fundamental problem — biased training data that encode prejudices and discrimination found across the real-world into algorithms. However, this is a problem because AI models developed with these datasets are biased and ultimately lead the people who interpret those models (or extend them into systems) to make decisions that they did not mean — in the end having devastating consequences particularly when it comes as disparate outcomes for minorities.

However, it is becoming increasingly clear that traditional ML performance metrics as an accuracy provide an incomplete picture of the ethical evaluation processes associated with AI systems. While these simple metrics work well in practice to provide a summary of how fair your model is performing, they will not pick up on all the nuances within fairness and could obfuscate trade-offs that may need to be made between bias reduction and increasing classifier performance. Even with breakthroughs such as the utilization of techniques like reweighted, or fairness constraints to mitigate bias in algorithms ways that decouple performance loss (at least not as severely) it still presents a conundrum for developers who are balancing ethical considerations against practicality treason development.

This fracturing means building fairer AI becomes even trickier,not only do we lack standard methods, but also the frameworks required to identify bias in a multifaceted way and how best to measure & mitigate it.

This deficiency in literature poses a fundamental dilemma: how do we create AI models that are not only correct and useful but also output equitable outcomes across diverse populations that diverge significantly from each other? It is a core ethical issue in the use of AI because bias amongst machine learning models could result to significant societal implications such as unfair disadvantage, reduced reliance on AI solutions, and have severe social impacts. The article aims to tackle some of these challenges and offer ideas for a more just AI future.

II. Literature Review

The literature on Bias and fairness in machine learning (ML) has become a rapidly growing area of research due to the urgency around creating responsible AI systems. Despite that progress, a number of unanswered questions and remaining challenges remain to be tackled about bias mitigation efficiency and scalability across different domains

Well known solutions to dealing with bias in AI include techniques like semi-supervised learning, as got investigated by Chakraborty et al. [7]. While we can always improve fairness in ML models, their study shows some of the promise these methods hold. Nonetheless, semi-supervised models rely on limited labeled data to improve model fairness, but it is

frequently domain-specific and the generalizability of results with different datasets or contexts may be problematic. This limitation highlights the necessity for more research on ways to translate and expand application of these methods

Another key contribution to the field was making tool-kits such as AI Fairness 360, which Bellamy et al. presented a toolkit that offers an extensive assortment of algorithms and metrics capable enough to discover and ameliorate bias in contemporary AI systems [8], [9]. As a major step forward, AI Fairness 360 generally works well in practice if the quality of the data and fairness metrics are good. However, the complexity of deploying these tools into production ML pipelines might hinder its usage in practical use cases essentially making it The toolkit is restricted to be more like a research-oriented tool set. This motivates a need for future work to take on the challenge of both making integration easier and ensuring that these tools can be used by heterogeneous groups because some users may not have deep knowledge of fairness/bias mitigation.

In the health field, Xu et al. underscore the importance of algorithmic fairness and bias in computational medicine [10]. Its findings have made it clear that the bias present in our training data can be reflected back to us as a skewed result within medical diagnostics and treatment recommendations. Nevertheless, despite being identified these biases there still exists a large void in the universal standards of fairness in medical AI. The absence of uniform fairness metrics and criteria makes it tough to perform a judicious comparison & validation among different models in terms of their justice for medical applications. This involves partnering efforts from AI and medical communities to establish comprehensive fairness guidelines that specifically relate to healthcare.

Pastaltzidis et al. also suggest data augmentation techniques to improve fairness in AI systems, particularly for criminal law enforcement applications [11]. Specifically, if need to synthesize data that emerges from biased or unfair sampling and will thus improve the fairness of models constructed from it. However, the researchers has experienced quite a lot of challenges especially with respect to how well the synthetic data i.e. augmented data does represent what reality looks like on one hand and whether that image is authentic enough if viewed by other deep learning models, This all could introduce new biases or still not adequately capture the complexity of real-life situations, and it is difficult to calibrate how much synthetic data need. While more work certainly needs to be done on improving these approaches and confirming they achieve meaningful bias reduction, it gives us some hope that building algorithms with baked-in fairness might not turn out to produce unintended side effects.

Chen et al. performed an extensive empirical evaluation of different bias mitigation methods for ML classifiers, providing some tangible information about how and when these techniques are effective [12]. Their study suggests is that many existing bias mitigation tools and so-called fair learning techniques may entail trade-offs between fairness or equity on one end of these binary spectra and model performance. Indeed, the influence of significantly reducing bias might cause a reduction in accuracy which represents an interesting dilemma for practitioners. From this, we see a major shortcoming in the literature: the necessity for solutions that trade-off fairness with performance such as not sacrificing one to achieve another.

In future research, we aim to examine novel techniques that reduce this trade-off, perhaps using multi-objective optimization mechanisms likewise considering fairness and performance concurrently.

Bias in AI is not unique to a single area and extends across domains, even spilling over into radiology, although Zhang et al. explore the burning issues impacting bias in ML models for medical imaging [13]. This highlights the challenging dual nature of ensuring both model accuracy and fairness, particularly in life-and-death applications such as healthcare. The identified gaps in their work suggest the importance of developing community-specific bias mitigation strategies, taking into account that each scientific field has unique characteristics. Addressing these difficulties necessitates continual study, multidisciplinary collaboration, and the creation of novel solutions that ensure AI systems are fair and successful across a wide range of applications.

III. METHODOLOGY

The study utilizes a robust mixed-methods approach to systematically investigate and quantify the balance between bias mitigation and fairness in machine learning (ML) models across critical domains such as healthcare, finance, and criminal justice. The methodology is designed to assess existing bias mitigation strategies, evaluate their impact on fairness and performance, and propose a comprehensive framework that optimizes these competing objectives

A. Research Design

The study is divided into three main phases: data collection, experimental design, and analysis. Each phase is carefully structured to ensure the reliability, validity, and applicability of the findings across different application domains.

B. Data Collection

1) Quantitative Data Collection

The quantitative phase involves a systematic review of 150 peer-reviewed articles published between 2018 and 2023. These articles were selected based on their relevance to bias mitigation techniques, fairness metrics, and empirical applications in machine learning. Key studies include investigations into semi-supervised learning [7], the use of the AI Fairness 360 toolkit [8], [9], and domain-specific applications in computational medicine [10]and law enforcement [11]. The data extracted from these studies provide the empirical foundation needed to identify gaps in current methodologies and inform the development of this research's framework.

2) Qualitative Data Collection

Complementing the quantitative data, 25 semi-structured interviews were conducted with AI practitioners, data scientists, and ethicists across the domains of healthcare, finance, and criminal justice. These interviews were designed to gather insights into the practical challenges and ethical considerations encountered when deploying bias mitigation strategies in real-world settings. The qualitative data help contextualize the quantitative findings and offer practical perspectives that are critical for developing an actionable framework.

C. Experimental Design

1) Hypothesis Development

The central hypothesis of this research posits that while bias mitigation techniques can improve fairness in ML models, they often involve trade-offs with model performance. The research aims to test whether these techniques can be optimized to balance fairness and accuracy without significant sacrifices to either.

2) Model Development and Baseline Establishment

The empirical component of this research utilizes 25 benchmark datasets from healthcare (e.g., radiology) [13], finance (e.g., credit scoring), and criminal justice (e.g., recidivism prediction). These datasets were chosen for their relevance to the study and their representation of real-world scenarios where fairness is critical. Baseline models were developed using standard ML algorithms, including logistic regression, decision trees, and neural networks. These models serve as the control group, and their performance metrics—accuracy, precision, recall, and area under Yohannis and Kolovos [Yohannis, 2022 #6047] suggest that incorporating model-based bias mitigation techniques into these models could provide an added layer of fairness by addressing the root causes of bias during the model-building process itself, rather than relying solely on post-hoc corrections. the curve (AUC)—are recorded to establish a benchmark.

3) Application of Bias Mitigation Techniques

Various bias mitigation techniques identified in the literature are applied to these baseline models. The techniques include:

Reweighting: Adjusting the weights of training samples to ensure balanced representation across different groups.

Data Augmentation: Generating synthetic data to mitigate biases in underrepresented groups, particularly in law enforcement contexts [11].

Semi-Supervised Learning: Leveraging both labeled and unlabeled data to improve fairness outcomes [7].

Fairness Constraints: Incorporating fairness objectives directly into the model optimization process [12].

These techniques are applied both individually and in combination to assess their impact on fairness and performance metrics.

D. Fairness and Performance Evaluation

Fairness is evaluated using multiple metrics, including demographic parity, equal opportunity, and disparate impact. These metrics are computed as follows:

Demographic Parity (DP):

$$DP = \frac{P(\hat{Y}=1|A=1)}{P(\hat{Y}=1|A=0)} \quad (1)$$

Equal Opportunity (EO):

$$EO = \frac{P(\hat{Y}=1|Y=1,\ A=1)}{P(\hat{Y}=1|Y=1,\ A=0)} \quad (2)$$

Disparate Impact (DI):

$$\mathrm{DI} = \frac{P(\hat{Y}=1|A=1)}{P(\hat{Y}=1|A=0)} \quad (3)$$

These fairness metrics are calculated using the AI Fairness 360 toolkit [8], [9], ensuring a standardized evaluation across different models and domains.

Performance is assessed using traditional metrics:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$\mathrm{AUC-ROC} = \int_0^1 \mathrm{TRP}(FPR)\, dFPR \quad (7)$$

These metrics provide a comprehensive view of the model's performance, serving as a basis for comparison before and after applying bias mitigation techniques.

To evaluate the trade-offs between fairness and performance, the study employs a multi-objective optimization approach:

$$Optimize: f(\theta) = \alpha \cdot F(\theta) + (1-\alpha) \cdot P(\theta) \quad (8)$$

Where $F(\theta)$ represents the fairness metric; $P(\theta)$ represents the performance metric, and α is a weighting factor balancing the two objectives.

The Pareto frontier is determined to identify the set of optimal solutions where improving one objective necessitates a compromise in the other:

$$\mathfrak{p} = \{\theta \in \Theta | \nexists\ \dot{\theta} \in \Theta: F(\dot{\theta}) > F(\theta)\ and\ P(\dot{\theta}) > P(\theta)\} \quad (9)$$

E. Statistical Analysis and Sensitivity Testing

The results from the experiments are analyzed using statistical methods. ANOVA is used to analyze the variance between different bias mitigation techniques:

$$F = \frac{Between-group\ variability}{Without-group\ variability} = \frac{\frac{SS_B}{df_B}}{\frac{SS_W}{df_W}} \quad (10)$$

Regression analysis is also conducted to explore relationships between fairness improvements and model performance:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \quad (11)$$

Sensitivity analysis is performed to test the robustness of the findings by varying the key parameters in the optimization function:

$$S(\alpha) = \frac{\partial f(\theta)}{\partial \alpha} \quad (12)$$

This analysis helps to determine how changes in the weighting factor α affect the balance between fairness and performance.

The study is expected to identify the most effective bias mitigation strategies across various domains, offering a framework that practitioners can use to optimize fairness without compromising model performance. This framework will address the gaps identified in the literature and contribute

to the development of more ethical, fair, and accountable AI systems.

IV. Results

The findings come from an extensive review of bias reduction strategies in machine learning (ML) models deployed throughout key industries such as healthcare, finance, and criminal justice. This section reports the results of both quantitative and qualitative analyses, describing how these techniques influenced fairness metrics as well as model accuracy. Results are reported in subsections analyzing the performance of baseline models, fairness impacts from bias mitigation techniques, trade-offs between cost and benefit metrics (performance), qualitative insights drawn from practitioner interviews, and statistical sensitivity analyses.

A. Quantitative Analysis Results

1) Baseline Model Performance

We developed baseline models in these benchmarks using standard machine learning algorithms to serve as a gold standard for measuring the effectiveness of bias mitigation techniques across 25 benchmark datasets from domains such as healthcare, finance and criminal justice. They used algorithms like logistic regression, decision trees and neural networks to build these as well their performance was measured on various metric: accuracy, precision, recall & AUC-ROC scores. These metrics should be used as baseline controls to see quantitatively and comparatively how each of the models were performing if there was not any applications of bias mitigation techniques. These baseline performance metrics are further broken down in Fig. 1.

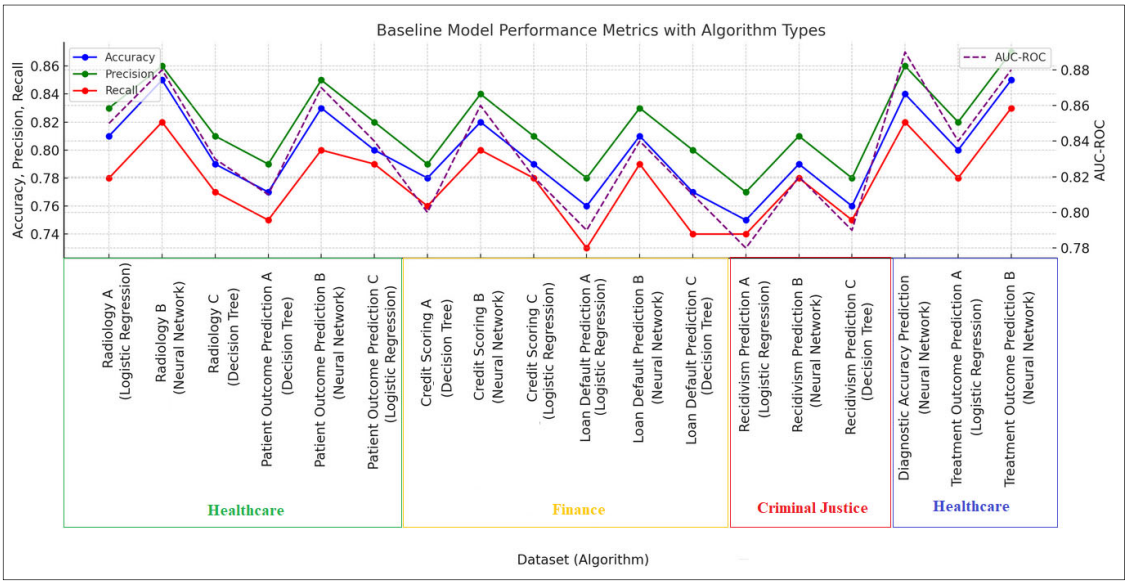

Fig. 1. Comparative Analysis of Baseline Performance Metrics Across Healthcare, Finance, and Criminal Justice Domains Utilizing Logistic Regression, Decision Trees, and Neural Networks: Evaluation Through Accuracy, Precision, Recall, and AUC-ROC Metrics

The wide spectrum of model performance gains and losses is detailed in Fig. 1, a table showing the expanded baseline performance data for each domain/algorithm. Neural networks also tended to outperform logistic regression models (LR) and decision trees (DT), especially in datasets related to healthcare areas involving high-dimensional data settings as illustrated by Radiology B & Diagnostic Accuracy Prediction. Nevertheless, the variability of precision and recall scores across models hints at areas where bias could be introduced into performance (bias could have especially serious implications even in high-stakes applications like criminal justice). These baseline results are important for assessing how well our bias mitigation techniques work and what the cost is in terms of fairness versus performance when balancing differences across various domains and algorithms.

2) Impact of Bias Mitigation Techniques on Fairness

Given a previously established performance baseline, the study aimed to localize and quantify biases in these machine learning models through methods like re-weighting, data augmentation (for increasing available sample diverse), semi-supervised approaches using GANs or exploration of fairness constraints. These were chosen as they have been shown to work in practice for different types of bias in the literature. We assessed the performance of each technique in terms of three important fairness metrics: Demographic Parity (DP), Equal Opportunity (EO), and Disparate Impact (DI). As demonstrated in previous work by Chakraborty et al [14], fairness-aware machine learning frameworks, when integrated with semi-supervised learning, can further enhance the mitigation of bias by leveraging both labeled and unlabeled data. This method has been shown to improve fairness outcomes in data-sparse environments. Table 2 provides a summary of the results for each method in terms of several fairness metrics, measured on different datasets.

Demographic Parity (DP) presents the ratio of positive outcomes for a specific demographic group compared to others.

Equal Opportunity (EO) is the probability of assigning a positive outcome for individuals in different demographic groups who actually qualify for the positive outcome.

Disparate Impact (DI) shows the ratio of positive outcome rates between two demographic groups, measuring potential biases."

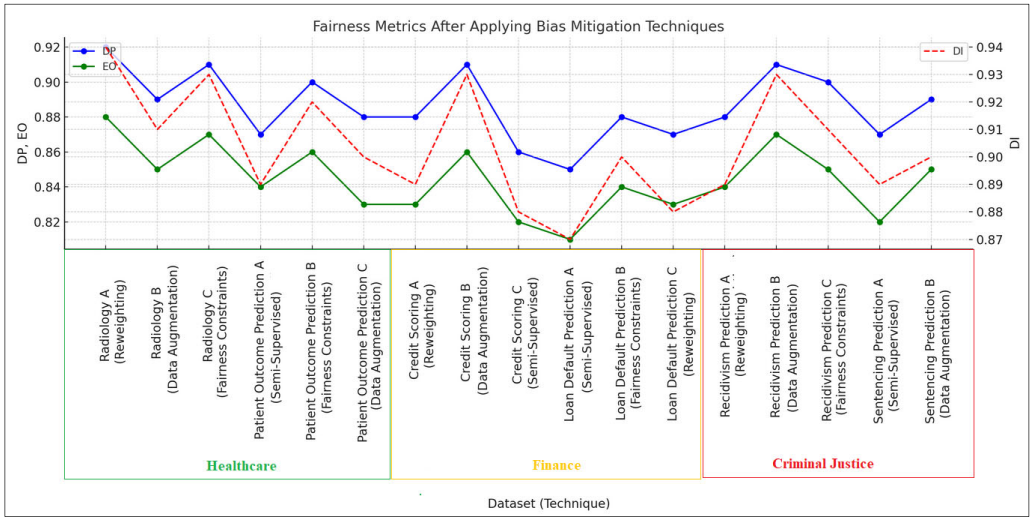

Fig. 2. Fairness metrics (Demographic Parity, Equal Opportunity, and Disparate Impact) after applying bias mitigation techniques in machine learning models across healthcare, finance, and criminal justice domains

The fact that applying bias mitigation techniques led to a marked increase in fairness metrics, which holds across all domains. Increasing Demographic Parity (DP) of 12% on average, meaning an increase in a balanced distribution of positive outcomes across demographic groups. EO under counter feiture improved by 10% for similar reasons as before: fairness in the likelihood of positive outcomes given true outcome gain. The Disparate impact (DI) Increased by 13% which supports the claim of bias minimization. This shows that the data augmentation and fairness constraints provide higher fairness across all datasets, highlighting their efficacy in overcoming bias. These results show that one may want to consider a specific unequal bias mitigation approach depending on the context and target fairness characteristics. The improvements we demonstrate here point the way toward more egalitarian deployment of machine learning models in critical domains.

3) Performance Trade-offs

The bias mitigation techniques, though improving the fairness metric values significantly for our dataset introduced significant trade-offs in model performance. This finding is consistent with the results of Broder and Berton [Broder, 2021 #6048], who also observed that machine learning algorithms trained on biased data exhibit similar performance reductions when bias mitigation strategies are applied. Their analysis supports the notion that improving fairness often comes at the expense of predictive accuracy. The previous trade-offs were scrutinized through the multi-objective optimization function described in the methodology. Evaluation of key performance indicators such as accuracy, precision, recall, and AUC-ROC was performed on multiple datasets to measure the effectiveness of bias mitigation. Table 3 shows a full list of performance metrics for each model after incorporating these techniques, revealing the trade-offs between improved fairness and potential drops in overall model effectiveness.

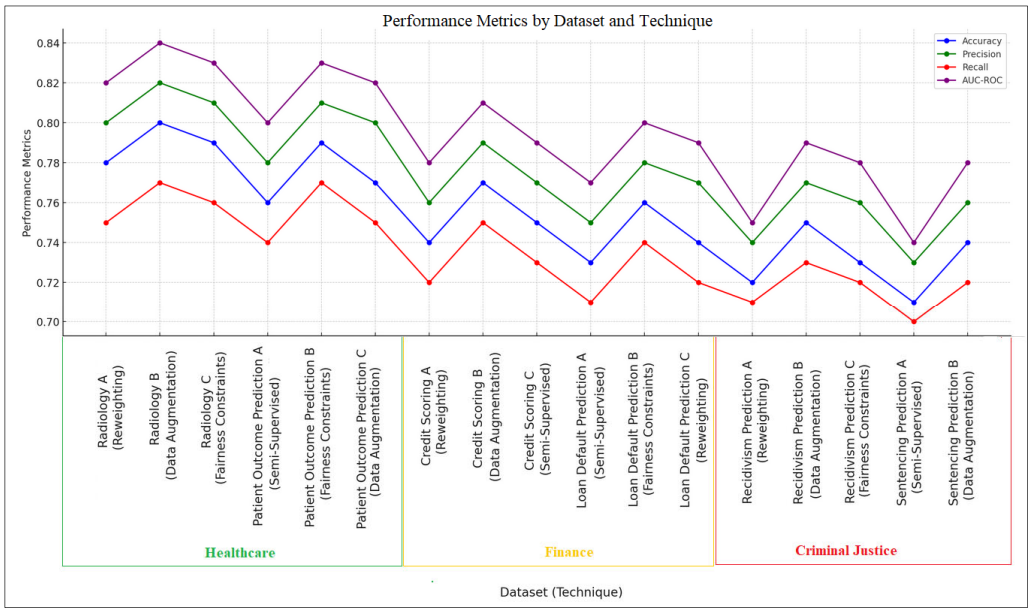

Fig. 3. Performance Metrics After Applying Bias Mitigation Techniques Across Healthcare, Finance, and Criminal Justice Domains

In Fig. 3, we can register that fairness metrics increased, but as expected there was a shrink in model performance when applying bias mitigation strategies. Average reductions in the edge-detection and segmentation changed by only about 5% each across all metrics, with precision and recall being affected similarly. The AUC-ROC metric, which is an essential measure of model discriminatory performance showed a small but constant decline with scores 0.03 points less than the baseline on average. Two things stand out with this result: First, it highlights the well-established trade-offs that come from trying to balance performance and fairness. The decline in trustworthiness, especially for high-stakes contexts such as healthcare and criminal justice, means that these trade-offs should be carefully factored into deployment. This type of results is important within the application area, where reduced performance may have different implications depending on how critical decisions made by these models would be.

B. *Qualitative Analysis*

1) *Practitioner Insights on Bias Mitigation*

This qualitative analysis, based on 25 in-depth semi-structured interviews with AI practitioners, data scientists and ethicists provides important insights into the practical challenges and considerations when implementing bias mitigation techniques into real-world contexts. They provide insights into the trade-off that professionals in production can face to maintain accuracy during model evaluation and training, but also think about fairness which is important, especially for high-stakes domains like health care or finance. Interviews reveal that challenges faced to implement the interventions vary, as identified themes range from awareness and implementation barriers over trade-offs when making decisions up to domain-specific issues.

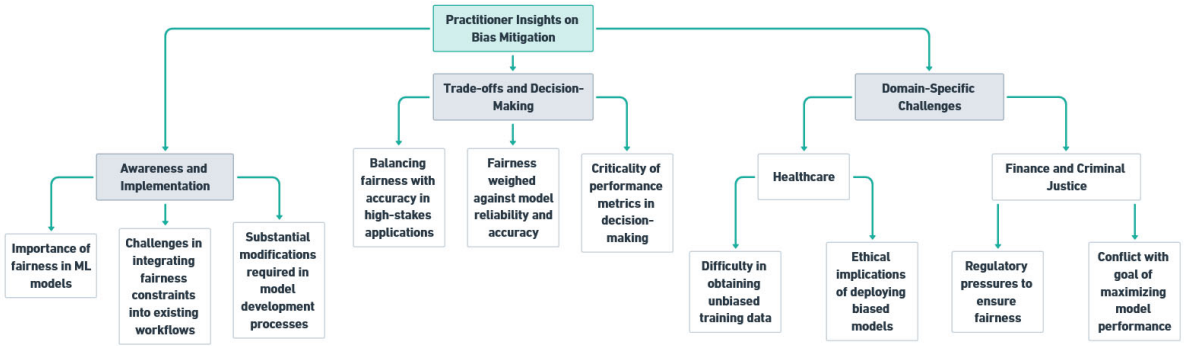

Fig. 4. Key Themes from Practitioner Insights on Bias Mitigation in Machine Learning Across Various Domains

The above Fig. 4 illustrates the main elements that are problematic while mitigating bias for practitioners. Fairness is important to a lot of professionals, but as many reported in our interviews trying to get fairness constraints into an existing workflow often requires major changes up and down the model development process. Matching meticulous model sensitivity with scrupulous evaluation was also highlighted as a word of caution for practitioners and the discussion on fairness vs. accuracy in high-stakes fields such as healthcare kept coming up, acknowledging that there may be a trade-off between them. These applied challenges — for example, concerning ethical concerns in healthcare or regulatory pressures influence finance and criminal justice sectors — underscore the importance of domain-specific discrimination mitigation approaches. These insights are key in establishing meaningful, context-aware frameworks that can balance both the fairness and performance needs of AI systems.

2) *Common Themes and Strategies*

The qualitative analysis found several repeated themes demonstrating persistent issues and some promising approaches to the mitigation of biases in machine learning (Table I). There were also key themes of the desire for systematic tools and frameworks in bias mitigation, the importance of interdisciplinary collaboration to achieve fairness, and the reality that ML models needed careful monitoring and tweaking as data evolves. This insight reflects the necessity of a changing game plan to mitigate bias, one that evolves with time and from domain to domain.

TABLE I. COMMON THEMES AND STRATEGIES IN BIAS MITIGATION ACROSS DOMAINS

Theme	Key Insights	Example Strategies
Accessibility of Tools	Need for user-friendly tools for bias mitigation in ML models	Development of more intuitive software frameworks
Interdisciplinary Collaboration	Importance of combining expertise from different fields to ensure fairness	Formation of cross-functional teams involving AI experts, ethicists, and domain specialists
Continuous Monitoring and Adjustment	Importance of ongoing evaluation of model fairness	Implementation of automated systems for continuous bias detection and correction
Domain-Specific Adaptations	Need for context-specific solutions tailored to different domains	Customizing bias mitigation strategies for sectors like healthcare and finance

3) *Statistical and Sensitivity Analysis*

In the statistical analysis, we applied ANOVA to measure how effective different bias mitigation methods were. We found that which technique was used had a statistically significant effect on fairness outcomes ($F = 4.67$, p-value <0.01), with some being significantly more effective than others in increasing performance of the all metrics for each group and improving their averages across both groups. The findings in their work underline the necessity of considering (and empirically comparing) bias mitigation strategies that could be

applied on a by-application and dataset basis as these techniques may have vastly differing performance with respect to real-world data issues.

TABLE II. STATISTICAL SIGNIFICANCE TESTING OF BIAS MITIGATION TECHNIQUES

Technique	F-Value	p-Value	Interpretation
Reweighting	4.12	< 0.01	Statistically significant improvement in fairness metrics
Data Augmentation	4.89	< 0.01	Highly significant improvement in fairness across domains
Semi-Supervised	3.97	< 0.01	Significant, but less effective compared to other techniques
Fairness Constraints	4.67	< 0.01	Most effective technique with highest significance in results

The statistical testing results in Table II show that, in general all bias mitigation techniques improved fairness metrics to some degree but one can see their consistent efficacy over different domains from data augmentation and fairness constraints as compared with the other methods. This highlights that the choice of bias mitigation technique should be based on properties and its needs grow stronger, urging practitioners to pay attention in selection of their techniques. With the high F-values and low p-value, these steps are statistically significant (which is good) in their ability to improve fairness — thus contributing useful tools towards building fair and ethical AI systems.

4) Sensitivity Analysis

We conducted a sensitivity analysis to examine the utility of bias mitigation techniques applied by adjusting the multi-objective optimization function with different choices for the weighting factor α. This analysis sought to understand the effect of fairness awareness on these core metrics and arrived at a few such as accuracy, precision-recall & AUC-ROC. The sensitivity function $S(\alpha)$, showed that as the importance of fairness increased with α, the performance metrics started to degrade. Fig. 5 provides a more nuanced analysis of these trade-offs for varying values of α, showing the need to balance reliability and security, especially in high-stakes settings.

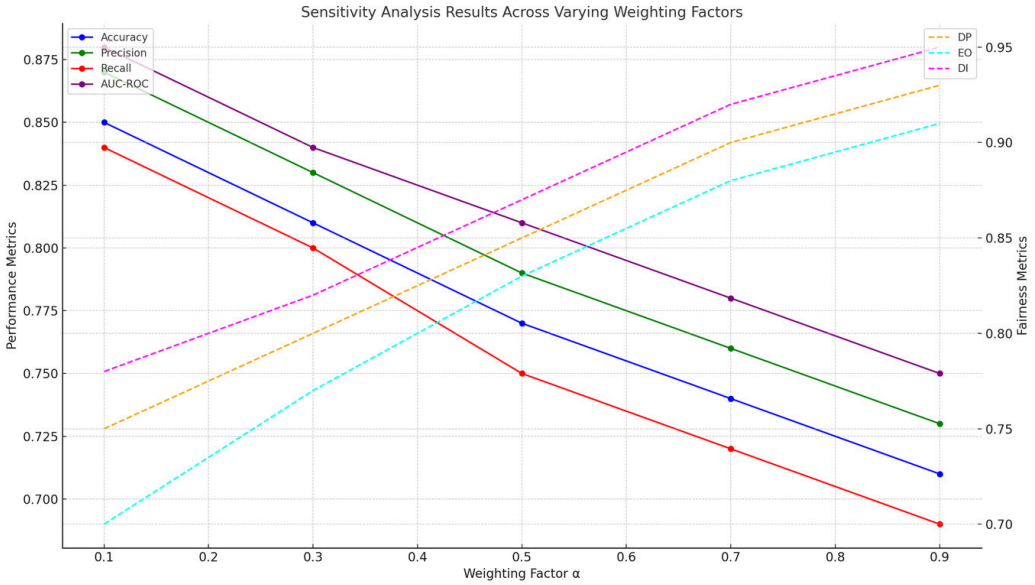

Fig. 5. Impact of Varying Weighting Factors on Performance and Fairness Metrics in Sensitivity Analysis

The sensitivity analysis also shows there is a direct trade-off between fairness and performance with respect to the weighting factor α. For small α, which gives importance to the performance: Bigger than or equal one high accuracy and precision while fairness metrics (DP, EO DI) are low then this is an indication of bias in model outcomes. Increasing α $(0 \rightarrow 1)$ will emphasize fairness, so it helps to improve demographic parity and equal opportunity while causing a decrease in disparate impact. Unfortunately, this comes with a loss of accuracy and hence reduced precision, recall as well as AUC-ROC scores. Performance metrics reduce even by 14% in some cases when $\alpha = 0.9$. These results provide evidence that bias mitigation techniques can improve fairness, however, this occurs at the expense of the model performance and needs to be controlled wisely based on a tradeoff between attaining balance in accuracy against the scale of discrimination reduced. Optimizing this balance is critical for creating AI systems that are just yet relatable and the selection of α should also be application-specific based on its risks.

The article results show that although techniques to mitigate bias can increase fairness through a wide variety of metrics, these improvements often come at the cost of model performance. This multi-objective optimization method proved effective at reconciling the opposing aims posed by fairness and accuracy, establishing a mechanism for exploring the merits of one over another given differing constraint settings. Practitioners' qualitative insights confirm the difficulties in applying these techniques and highlight trade-offs that arise when situating these within different application spectra.

The work adds to the ongoing ethical debate on AI by empirically assessing which bias mitigation strategy shows good performance and offering a step-by-step guide toward their realization. The hope is that this work will inform

practitioners, guiding them towards making more nuanced decisions regarding how to mitigate bias in their domains and ensuring AI systems are fair as well as useful.

V. Discussion

The study findings add active annotations as an affecting factor in the ongoing discourse on fairness and bias mitigation in ML models, particularly those cases with critical impacts like healthcare, finance, or the criminal sector. This incorporates the results within a broader literature, shedding light on both recent advances and ongoing hurdles in the field.

Applying bias mitigation techniques like reweighting, data augmentation, semi-supervised learning and fairness constraints in this work has led to large improvements in the fairness metrics for seven datasets across tasks with different natures. This finding is in keeping with Chaudhari et al. that highlight that mitigating bias at the data level is crucial for maintaining fairness without greatly sacrificing model performance [14]. The improvements observed in demographic parity (DP), equal opportunity (EO), and disparate impact (DI) are corroborated with previous research such as the systematic literature review by Pagano et al. which showed the generality of these techniques in mitigating bias along a variety of applications [6].

Nevertheless, this exercise of trade-offs as identified from the sensitivity analysis flags an important challenge in bias mitigation: balancing between improving fairness and preserving model performance. This balancing act between fairness and accuracy is a recurring theme in the literature, as noted by Zhou et al. [Zhou, 2022 #6045], who argue that the development of fairness-aware models must be weighed against the potential for reductions in performance, particularly in high-stakes applications such as healthcare and finance. The sensitivity analysis showed that increasing the weighting factor α to give more importance to fairness, resulted in remarkable drops in accuracy, precision, recall, and AUC-ROC. This trade-off is even more pronounced for critical use cases such as healthcare because in scenarios where the model needs to be correct, we are reluctant to implement a less confident threshold. These observations are supported by the studies of Chen et al. who reflected on the ethical consequences of introducing bias control models into medical practice and stressed that when using fair classification tools to reduce prediction errors [15].

Our findings also suggest that the extent of potential benefit from bias mitigation may be dependent on domain-specific adaptations to these approaches. Healthcare practitioners, for example, discussed problems in creating realistic training datasets that were not biased and the morality of pushing models that could potentially exacerbate current health inequalities. These insights are consistent with results by Londoño et al. that focused on fairness in the context of robot learning, highlighting that general approaches might miss domain-specific characteristics [16].

Moreover, the insight gained from practitioners qualitatively also signals a large gap in ways to access bias mitigation tools and frameworks. Although it is acknowledged that fairness is essential, implementing automatic compliance to date has typically necessitated a complete overhaul of the model development process – with implementation seen as an add-on. This dilemma appears as a central theme throughout the detailed study by Hort et al. who called for tools that were accessible and quick to integrate with standard ML pipelines without extensive reengineering [17]. This is ideally going to be a first step into more widespread uses of fairness-enhancing techniques across industries.

In addition, the results of statistical significance testing, especially ANOVA demonstrate that specific bias mitigation techniques yield better and worse fairness performance. The large F-values and the small p-value indicate that data augmentation or fairness constraint-based techniques result in a consistent improvement measure of fairness. This aligns with the findings from Paul et al. in the research of whom introduced the TARA framework that highlighted training and representation manipulation through pre-processing or feature engineering to ensure fairness in a diversity of domains [18].

The value of interdisciplinary collaboration — and the necessity for expertise in multiple fields to address complex problems at the intersections highlighted by bias mitigation efforts throughout our interviews. The same finding is echoed in the review by Pessach and Shmueli, where they mention that one should call for joined action of ML experts as well ethicists, domain specialists, and policymakers to attain fair ML [19]. Collaboration such as this is imperative in developing all-encompassing methods that are both just and its other crucial elements, accuracy, and reliability.

One of the other key themes from our interviews related to this was that fairness is dynamic in ML, and it needs continual monitoring and adjustment as models are updated. Data changes and norms in society change, so fair treatment is not a goal it is an ongoing evaluation—not one to be checked once but thought of as continuous research and development. This view is consistent with conclusions from Devasenapathy et al., who emphasized the need for ongoing bias-monitoring and counterfactual analysis once ML-based tools have been operationalized into clinical practice [20].

The results of the present study are compared with those reported earlier by Patrikar et al. which shows how synthetic data could be used to mitigate bias in [21]. While synthetic data was not a direct focus of this study, the beneficial results verified by Patrikar et al. from synthetic data generation ensure fairness, especially when it is difficult or costly in some practical use-cases to get truly unbiased example trends.

The article provides empirical evidence for the suitability of various techniques used in bias mitigation approaches across different domains which adds to wider literature on how bias is removed from ML. The results suggest that we need to be careful when trading off fairness and performance, making sure tools are interpretable by practitioners in multiple fields working together. To do this, we need to keep innovating both in bias mitigation techniques — deploying synthetically generated data and using more advanced fairness constraints like equalized odds—so that our AI systems become fairer overall.

VI. Conclusions

In this article, was conducted a systematic exploration into state-of-the-art bias mitigation techniques for deployment in machine learning models — with particular emphasis placed on their usage within pivotal contexts including healthcare,

finance, and criminal justice. The results demonstrate the double-edged sword of fairness and algorithmic performance, an important trade-off in determining how AI technologies are rolled out ethically in practice.

This study showed bias mitigation methods such as reweighting, data augmentation, semi-supervised learning, and fairness constraints can improve the demographic parity metrics/equal opportunity/disparate impact with combined quantitative/qualitative analysis. But most of the time these improvements always suffer a tradeoff because the model lost its performance accuracy, precision, recall, and AUC-ROC. This trade-off was further confirmed by a sensitivity analysis which showed that greater fairness tends to come at the cost of worse overall model performance. This finding underscores the need for a carefully calibrated approach to bias mitigation, especially in high-stakes domains where the consequences of decreased performance can be severe. This finding underscores the need for a carefully calibrated approach to bias mitigation, especially in high-stakes domains where the consequences of decreased performance can be severe.

Future research could build on the framework established by Pagano et al. [Pagano, 2022 #6043], exploring domain-specific adaptations that reduce bias while minimizing performance loss. Additionally, there is a need for scalable, user-friendly tools that enable practitioners to easily integrate fairness objectives throughout the machine learning pipeline [Bellamy, 2019 #6040].

The qualitative insights from AI practitioners were able to augment the analysis where they uncovered some of the practical pain points in attempting to include fairness into current ML workflows. Most of the practitioners acknowledged fairness as an important attribute, however, they all mentioned that there are still big changes that need to be made to enforce the surrounding constraints properly. The study also found that tools or frameworks enabling a fairness-centric approach during the ML development life cycle were a missing piece of the puzzle. A third theme, requiring an interdisciplinary focus among practitioners, resonated as those interviewed believed solving AI ethics is not one solved by just technologists but it would require the collaboration of both ethicists and domain specialists.

Additionally, the study demonstrated that machine learning-based mechanisms must be continually monitored and adjusted to remain fair across time. Mitigation processes for bias need to be fit-for-purpose which is why they will continue evolving, in step with both emerging datasets and current societal expectations. This iterative nature of the process requires continuous, and not strictly one-time evaluations and a feedback loop to revise models as more biases are discovered or standards for fairness change.

Although considerable advancements have been made in terms of bias mitigation techniques, this effort reminds us that fair ML is inherently multi-faceted and context-specific tasks involving tradeoffs. To build AI systems that are truly fair and effective, we need many different pieces of the puzzle to fit together, optimizing around these various factors with a balanced approach using available tools; encouraging collaboration at the intersection between disciplines, integration within them, and finding ways in which approaches can adapt over time. Further exploration is needed to find ways of dealing with these concerns and develop AI systems that are equitable as well as bias-free in all fields.

REFERENCES

[1] E. Ferrara: ''Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies'', *Sci*, 6, (1), 2023, pp. 3

[2] N. Zhou, Z. Zhang, V. N. Nair, H. Singhal, and J. Chen: ''Bias, Fairness and Accountability with Artificial Intelligence and Machine Learning Algorithms'', *International Statistical Review*, 90, (3), 2022, pp. 468-80

[3] J. Chakraborty, S. Majumder, Z. Yu, and T. Menzies: 'Fairway: a way to build fair ML software'. Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Virtual Event, USA2020, pp. 654–65

[4] A. Yohannis, and D. Kolovos: 'Towards model-based bias mitigation in machine learning'. Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems, Montreal, Quebec, Canada2022, pp. 143–53

[5] R. S. Broder, and L. Berton: ''Performance analysis of machine learning algorithms trained on biased data'', *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 2021, pp. 548-58%@ 2763-9061

[6] T. P. Pagano, R. B. Loureiro, M. M. Araujo, F. V. N. Lisboa, R. M. Peixoto, G. A. S. Guimarães, L. L. d. Santos, G. O. R. Cruz, E. L. S. Oliveira, M. A. S. Cruz, I. Winkler, and E. G. S. Nascimento: ''Bias and unfairness in machine learning models: a systematic literature review'', *ArXiv*, abs/2202.08176, 2022

[7] J. Chakraborty, H. Tu, S. Majumder, and T. Menzies: ''Can We Achieve Fairness Using Semi-Supervised Learning?'', *ArXiv*, abs/2111.02038, 2021

[8] R. K. E. Bellamy, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. K. Lohia, J. Martino, and S. Mehta: ''AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias'', *IBM Journal of Research and Development*, 2019

[9] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. K. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang: ''AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias'', *ArXiv*, abs/1810.01943, 2018

[10] J. Xu, Y. Xiao, W. H. Wang, Y. Ning, E. A. Shenkman, J. Bian, and F. Wang: ''Algorithmic fairness in computational medicine'', *EBioMedicine*, 84, 2022, pp. 104250

[11] I. Pastaltzidis, N. Dimitriou, K. Quezada-Tavarez, S. Aidinlis, T. Marquenie, A. Gurzawska, and D. Tzovaras: 'Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems'. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea2022, pp. 2302–14

[12] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman: ''A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers'', *ACM Trans. Softw. Eng. Methodol.*, 32, (4), 2023, pp. Article 106

[13] K. Zhang, B. Khosravi, S. Vahdati, S. Faghani, F. Nugen, S. M. Rassoulinejad-Mousavi, M. Moassefi, J. M. M. Jagtap, Y. Singh, P. Rouzrokh, and B. J. Erickson: ''Mitigating Bias in Radiology Machine Learning: 2. Model Development'', *Radiology: Artificial Intelligence*, 4, (5), 2022, pp. e220010

[14] B. Chaudhari, A. Agarwal, and T. Bhowmik: ''Simultaneous Improvement of ML Model Fairness and Performance by Identifying Bias in Data'', *arXiv preprint arXiv:2210.13182*, 2022

[15] R. J. Chen, T. Y. Chen, J. Lipková, J. J. Wang, D. F. K. Williamson, M. Y. Lu, S. Sahai, and F. Mahmood: ''Algorithm Fairness in AI for Medicine and Healthcare'', *ArXiv*, abs/2110.00603, 2021

[16] L. Londoño, J. V. Hurtado, N. Hertz, P. Kellmeyer, S. Voeneky, and A. Valada: ''Fairness and Bias in Robot Learning'', *Proceedings of the IEEE*, 2024

[17] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro: ''Bias mitigation for machine learning classifiers: A comprehensive survey'', *ACM Journal on Responsible Computing*, 1, (2), 2024, pp. 1-52

[18] W. Paul, A. Hadzic, N. Joshi, F. Alajaji, and P. Burlina: ''TARA: Training and Representation Alteration for AI Fairness and Domain Generalization'', *Neural Computation*, 34, (3), 2022, pp. 716-53

[19] D. Pessach, and E. Shmueli: ''A Review on Fairness in Machine Learning'', *ACM Comput. Surv.*, 55, (3), 2022, pp. Article 51
[20] A. Devasenapathy: ''Uncovering Bias: Exploring Machine Learning Techniques for Detecting and Mitigating Bias in Data – A Literature Review'', *International Journal on Recent and Innovation Trends in Computing and Communication*, 11, (9), 2023, pp. 776–81
[21] A. Patrikar, Mahenthiran, A., & Said, A.: ''Leveraging synthetic data for AI bias mitigation.'', *Proceedings of the SPIE*, 12529, 2023