Optimizing fairness and accuracy: a Pareto optimal approach for decision-making

AI and Ethics
https://doi.org/10.1007/s43681-024-00508-4

ORIGINAL RESEARCH

Optimizing fairness and accuracy: a Pareto optimal approach for decision-making

Rashmi Nagpal[1] · Rasoul Shahsavarifar[2] · Vaibhav Goyal[3] · Amar Gupta[1]

Received: 28 December 2023 / Accepted: 22 June 2024

Abstract
In the era of data-driven decision-making, ensuring fairness and equality in machine learning models has become increasingly crucial. Multiple fairness definitions have been brought forward to evaluate and mitigate unintended fairness-related harms in real-world applications, with little research on addressing their interactions with each other. This paper explores the application of a Minimax Pareto-optimized solution to optimize individual and group fairness at individual and group levels on the Adult Census Income dataset as well as on the German Credit dataset. The objective of training a classification model with a multi-objective loss function is to achieve fair outcomes without compromising utility objectives. We investigate the interplay of different fairness definitions, including definitions of performance consistency and traditional group and individual fairness measures, amongst each other coupled with performance. The results presented in this paper highlight the feasibility of incorporating several fairness considerations into machine learning models, which can be applied to use cases with multiple sensitive features and attributes that characterize real-world applications. This research is a valuable step toward building responsible and transparent machine learning systems that can be incorporated into critical decision-making processes.

Keywords Demographic parity · Pareto optimal · Group fairness · Individual fairness

1 Introduction

In machine learning, the primary objective is usually the development of fair and accurate predictive models. The growing importance attached to the pivotal intersection of technology and ethics has prompted researchers and practitioners to explore innovative approaches to challenge the paradigms of predictive model deployment. As the deployment of such models becomes increasingly widespread, concerns about fairness and equity in decision-making have also escalated. This is especially pertinent in critical domains like loan classification, where the consequences of biased decisions can have far-reaching socioeconomic implications.

This paper aims to strike a harmonious balance between predictive accuracy and several proposed fairness definitions in the classification tasks. The Adult Census Income dataset [1] and German Credit dataset [2] is used as the foundation for experimentation. The fundamental premise revolves around implementing the classification model, which utilizes a multi-objective loss function, making it possible to optimize the predictive accuracy, group fairness, and individual fairness simultaneously.

A models prediction accuracy is often employed as the key performance indicator for machine learning models. However, the pursuit of accuracy alone can inadvertently perpetuate biases within machine learning models, leading to unjustified differential treatment of demographic groups

✉ Rashmi Nagpal
rnagpal@mit.edu

Rasoul Shahsavarifar
rasoul.shahsavarifar@ca.ey.com

Vaibhav Goyal
vaibhavgoyal.2606@gmail.com

Amar Gupta
agupta@mit.edu

1 Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, MA 02139, USA

2 Artificial Intelligence Risk Consulting, Ernst & Young LLP, EY Tower, 100 Adelaide Street West, PO Box 1, Toronto, ON M5H 0B3, Canada

3 Don Bosco School, Park Circus, West Bengal, India

or individuals. Regulatory bodies globally are catching up to impose guidelines and restrictions on AI solutions to test and account for fairness. Still, they are non-prescriptive regarding specific procedures to ensure fair and unbiased model performance. From an ethical standpoint, several challenges relating to fairness definitions and their impact on accuracy emerge. Fairness is a rich concept with varying verbal definitions from philosophy, law, and social sciences. Current mathematical formulations can only partially capture these concepts, posing challenges to those intent on creating "fair" AI systems. To implement fairness into a model, practitioners are usually tasked to choose one or more statistical fairness considerations, implementing one definition of "fairness", e.g. "fairness through unawareness" or "outcome parity". These statistical measures guarantee group fairness, or the fairness for the "average" member of each predefined group [3], without regard for the treatment or outcome of any individual. In particular, many M=machine learning models not only affect marginalized populations but can have life-defining effects on individuals, e.g., in the case of credit adjudication use cases, which warrants the additional consideration of potential harm on the individual level. Furthermore, many traditional fairness metrics are mutually exclusive, meaning optimizing for one metric will prohibit the optimization of another in real-world scenarios due to different baseline rates in protected subgroups. Finally, enforcing a given definition of fairness may unequally impact the accuracy of individuals/groups, resulting in solutions with less favorable performance for protected groups. [4] Recognizing the ethical imperative for frameworks to optimize the trade-offs between different fairness metrics and accuracy measures, we introduce a new approach that leverages the Minimax Pareto Optimizer [5] to optimize both individual and group fairness while maintaining the high predictive accuracy of our model.

In essence, this research strives to bridge the gap between the often conflicting objectives of accuracy and fairness in machine learning. By embracing Pareto optimization and investigating the trade-offs with different fairness metrics, we strive to pave the way for more responsible and equitable lending practices within the financial industry and, by extension, in other domains reliant on algorithmic decision-making. Our journey signifies a crucial step toward developing transparent and accountable machine learning systems, fostering trust and ethical standards in our increasingly data-driven world.

The key components of the paper are described below.

1. We expand and test an innovative framework (**Algorithm 1**) that simultaneously addresses fairness and accuracy in machine learning models, focusing on individual and group fairness.
2. We demonstrate the application of Pareto Minimax optimization as a solution to mitigate the accuracy-fairness trade-off, resulting in a minimal decrease in accuracy (less than 0.6%) while significantly improving fairness metrics.
3. Our research highlights the complementary and competing aspects of different fairness metrics within a single optimization framework, demonstrating how a combination of utility-based and traditional fairness metrics can enable practitioners to achieve high performance across subgroups while improving fairness outcomes.

This paper is structured as follows—Sect. 2 offers an extensive literature review, delving into prior research on fairness and Pareto optimization. In Sect. 3, we articulate the architecture of our framework and fairness criteria, along with the dataset particulars and evaluation metrics used across our work. Section 4 provides a comprehensive overview of our research findings with a focus on the performance in terms of accuracy, individual fairness, and achieved group fairness for sensitive features like race and gender. In Sect. 5, we explore potential research directions for our work. Lastly, Sect. 6 encapsulates the conclusion of our paper.

2 Related work

An ever-increasing body of research addresses the trade-off between different group fairness measures and accuracy in machine learning models. The following section provides an overview of selected algorithmic fairness mitigation methods addressing the fairness-accuracy trade-off. Our work aligns with several constraint and multi-objective optimization approaches to find solutions that optimally balance fairness and accuracy.

2.1 Constraint optimization

Corbett-Davies et al. [6] explored the trade-offs between fairness and accuracy in algorithmic decision-making. They discuss the concept of fairness as a constraint optimization problem and present a Bayes optimal framework for quantifying the cost of fairness in terms of accuracy. Constraint optimization can be a powerful tool to optimize trade-offs, but it requires individual thresholds for each constraint and use case, imposing considerable overhead to the implementation of fair AI solutions in the industry. Agarwal et al. [7] build on a similar premise as [6] and present a reduction-based approach to fair classification with finite-sample guarantees. By formalizing statistical group fairness measures as linear constraints, they frame fairness violations as an additional cost to a classification optimization target. This cost sensitive classification applies to a variety of algorithms providing an optimal trade-off between a chosen fairness constraint and overall accuracy.

Yan et al. [8] take a reweighing approach to optimize accuracy and fairness simultaneously in their FORML framework, which is focused on learning data weights based on their importance to a constraint optimization target. This approach aims to improve fairness by adjusting the weights of the training samples while minimizing the adverse effect on accuracy. The framework represents a data-centric solution for addressing fairness concerns in machine learning models; however, it requires subdividing data into two datasets to prevent overfitting. While the algorithm maintains good levels of accuracy, the reduction in fairness violations is small (1%) in the experiments shown.

2.2 Pre and post processing for group fairness

Other approaches include pre- and post-processing methods to enhance fairness metrics while thriving to maintain acceptable levels of model performance. These different engineering choices carry further ethical implications, where in-processing methods effectively mitigate bias stemming from data set inherent correlation patterns. In contrast, post-processing methods use decision uncertainty as the basis for debiasing [4]. Pre- and post-processing fairness mitigation effectively treat the Machine Learning model as a "black box", giving them the advantage of applicability across different model types. Simultaneously, these approaches typically cannot guarantee to find the best performance-fairness trade-off, as they do not specifically optimize for it at training time. Pleiss et al. [9] discuss simple calibration to achieve fairness and accuracy simultaneously in machine learning models and show that calibration is only suitable for single-error constraints and equivalent to randomizing predictions for a subset of data points. Ethically speaking, these results are of concern, as practitioners leveraging calibration accept occasional random and unfair predictions for individuals. More advanced post-processing approaches include Lohia et al. [10], who propose a novel bias mitigation post-processing framework that increased individual and group fairness metrics by targeting individual biased data points, switching the predicted outcome of likely unfairly treated individuals in the unprivileged group to the outcome of a similar individual in the privileged group. Their algorithm—Individual Group Debiasing (IGD)—achieved an accuracy of 85% across all test data sets. The operational procedure is a pure run-time approach as it does not require ground truth class labels for the validation dataset. While the results are promising, and the algorithm run-time approach is highly relevant for business use cases, its applicability is limited by the need for pairs of "similar" individuals in the privileged and unprivileged groups. Further, the algorithm assumes only individuals in the unprivileged group can be treated unfairly.

Pre-processing methods generally target the input data of a Machine Learning model, aiming to reduce bias. A recent pre-processing trend is learning fair data representations based on fairness constraints, employing Variational Autoencoders of Generative Adversarial Networks [11]. One example is Xu et al. [12], who propose a framework to enhance counterfactual fairness in machine learning models by capturing the causal relationships between variables and interventions. Their results indicate acceptable performance with improved fairness scores for their method on the Adult and Law School datasets. Similarly, Zemel et al. [13] propose a learning algorithm that relies on learning fair representations of the input data to achieve both group and individual fairness while minimizing information loss and, therefore, maintaining high accuracy. In practice, the algorithm cannot guarantee ideal performance- fairness trade-offs and learning fair representations of data hampers the explainability of the AI model, as resulting features cannot be interpreted directly.

Our work focuses on in-processing methods, taking a procedure-driven approach to fairness, focusing on the trade-off between fairness metrics and model performance. We avoid pre-processing methods to maintain the explainability of the data and modeling results, which is a regulatory requirement in many jurisdictions for use cases involving critical decisions, such as healthcare and financial services. Post-processing methods can further enhance the achieved fairness, depending on the specific requirements of a use case, without compromising explainability requirements.

2.3 Trade-offs with individual fairness

Beyond the well-studied group fairness-performance trade-off, the mutual exclusivity of algorithmic fairness formulations, including individual fairness, requires practitioners to consider interactions between different fairness measures. The trade-offs between individual and group fairness or accuracy have been less studied. While group and individual fairness are often framed as competing definitions, Binns [14] discusses how approaches that aim to capture individual fairness could indirectly improve group fairness by reducing the reliance on protected attributes, contesting the notion of competing objectives for group and individual fairness. In line with this finding, Shahsavarifar et al. [15] found preliminary evidence indicating that optimizing individual fairness does not cause significant trade-offs in group fairness. Their work proposes a two-step training process focused on credit adjudication. It involved learning a fair similarity metric from a group sense using a small portion of the raw data and training a 'fair' classifier using the rest of the data, obscuring the sensitive features. The salient characteristic of the two-step training process is its flexibility; the fair metric obtained in the first step will work with any fairness algorithm in the second step.

Work on the individual fairness-accuracy trade-off includes research by Li et al. [16], who propose an in-processing fairness constraint using a Siamese fairness approach and showed that it could minimize the accuracy trade-off while achieving promising levels of individual fairness. Dwork et al. [17] introduce a framework for fair classification that incorporates a task-specific metric for determining the similarity between individuals in the classification task. The framework maximized the utility subject to the fairness constraint so that similar individuals could be treated in similar manner.

2.4 Pareto optimal solutions

To address the potential trade-offs between several competing optimization targets to find an optimal solution, we turn to Pareto optimality, defined as a solution where no single objective can be improved without decreasing another. This approach was previously taken by Liu et al. [18], who propose a stochastic multi-objective optimization framework to evaluate trade-offs between prediction accuracy and traditional group fairness metrics for binary classification. This framework obtains well-spread and accurate Pareto fronts, enabling the decision-maker to define their desired trade-off in an informed way instead of imposing predefined ϵ—based fairness constraints. Martínez et al. [5] leverage a similar mathematical construct but formulate a multi-objective optimization problem, following a "no unnecessary harm" definition of group fairness, circumventing traditional definitions of group fairness. Their work proposes a Pareto optimal classifier with separate risk functions per sensitive subgroup to find a Pareto optimal solution with the smallest worst group conditional risk. Their framework did not require access to sensitive attributes and relies on reweighing "risk scores" per sensitive subgroup to optimize performance by subgroup during training. The "no unnecessary harm" definition of fairness resonates ethically, alleviating concerns around fairness-optimized Machine Learning models at the expense of lower performance for certain population subgroups. Further, Valdivia et al. [19] presented a methodology to assess the boundaries of accuracy and fairness using a Pareto front. In contrast, Yu et al. [20] highlighted the trade-offs and potential approaches to achieve fairness-aware optimization and the importance of considering fairness in the decision-making process. However, it is vital to incorporate various fairness constraints in the optimization process, as discussed by Zafar et al. [21].

Many promising algorithms to create fair AI solutions have been developed in the last decade, addressing varying field challenges, from performance trade-offs to the incompatibility of fairness metrics and providing formal guarantees of performance and fairness to users of these models. In this work, we investigate the potential of combining the principle of utility-based fairness (i.e. following the principle of "no unnecessary harm"), tackled by the Minimax Pareto Fairness algorithm introduced by Martinez et al. [5], with constraint optimization to directly address formal definitions of group and individual fairness. While these objectives may directly compete with the original accuracy target by [5], we argue that use cases from highly regulated domains, such as loan adjudication, will require algorithmic solutions that directly target these well-defined fairness metrics.

3 Materials and methods

Three concepts are leveraged to address the joint optimization of group and individual fairness while preserving predictive accuracy across subgroups. First, we introduce a group fairness and individual fairness optimization target to model training. Second, we leverage the Minimax Pareto framework by Martinez et al. [5] and embed the multi-objective optimization into a Pareto framework obeying the principle of "no unnecessary harm." These concepts are explained in the following sections.

3.1 Group fairness

Group fairness (GF) refers to the principle that a predictive model should provide similar performance levels across different groups. The literature defines three important group fairness definitions: demographic (or statistical) parity, equal opportunity difference, and predictive parity. In this work, we limit our exploration to the principle of demographic parity.

Demographic Parity measures the extent to which a predictive model provides similar outcomes across different groups. A perfect score requires the prediction to be statistically independent of the protected attribute. The statistical parity of a binary predictor $\hat{Y}$with one binary sensitive feature A is

$$P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1) \tag{1}$$

An optimization target for Demographic Parity can be obtained by reformulating it as a sum of linear constraints. Two inequality constraints describe the achieved demographic parity for the realization of a sensitive feature. [7].

$$P(\hat{Y}=1|A=a)-P(\hat{Y})\leq k,$$
$$-P(\hat{Y}=1|A=a)+P(\hat{Y})\leq k, \tag{2}$$

where k is a threshold value for the acceptable deviation from Demographic Parity.

3.2 Individual fairness

The principal idea behind Individual Fairness (IF) is to ensure that similar individuals are treated similarly, regardless of their protected attributes. It is an important complementary practice to group fairness, as individuals can be treated unfairly, which cannot be detected at the aggregated group level. Quantifying individual fairness requires the identification of similarities between individuals. Here, the focus is on metric-based individual fairness, where similarity is quantified with a predefined metric measuring the distance between two individual data points. The choice of distance metric strongly depends on the data and the use case. For high-dimensional data, the Manhatten distance metric is resilient to outliers while providing the best contrast among distance metrics with integral norms [22].

In our study, we employed the Manhattan distance metric to define similarity in the experiments to account for the correlations observed in most multivariate real-world data sets.

The Manhattan distance is a multivariate metric that measures the distance between two points as the sum of absolute distances between the coordinates of each point. The Manhattan-distance between two points $\mathbf{a},\mathbf{b}\in\mathbb{R}^n$ is given by

$$M=\sum_i |a_i-b_i|,\text{ where } a=\{a_i\}, b=\{b_i\}, \text{and } 1\leq i\leq n. \tag{3}$$

Individual Fairness is the ratio of similar pairs that receive different predictions to the total number of similar pairs [15].

$$IFM(\epsilon)=\frac{\sum_{(x,x')\in S} I((d(x,x')\leq\epsilon)\cap(h(x)=h(x')))}{\sum_{(x,x')\in S} I(d(x,x')\leq\epsilon)} \tag{4}$$

where S is the set of all possible pairs, I is an indicator function, d is the similarity metric, ϵ is the threshold for similar pairs and h is the classifier. While h(x) represents the prediction of the classifier for the input x, and h(x') represents the classifier's prediction for the input x'. h(x) and h(x') are used to compare the predictions of the classifier for the pairs of inputs (x, x') that are considered similar based on the similarity metric d. The ratio is then calculated based on the number of similar pairs with different predictions over the total number of similar pairs.

3.3 Minimax pareto optimization

Pareto optimization aims to find solutions that simultaneously optimize multiple conflicting objectives without sacrificing performance in one objective for another. It relies on Pareto dominance, where one solution is said to dominate another if it is at least as good in all objectives and better in at least one.

The Minimax Pareto optimization scheme employs Minimax programming to find Pareto optimal solutions, given a set of loss functions.

Minimax optimization is a mathematical optimization approach that aims to minimize the maximum value among a set of functions or variables. It reflects a pessimistic approach to optimization and aims to minimize the worst-case outcome to provide a minimal guaranteed outcome. Due to this property, it is a preferred approach to solve problems in decision-making under uncertainty or in competitive scenarios. In the context of fairness, Minimax Pareto optimization can be leveraged to improve the outcomes for disadvantaged protected groups without impacting the performance of other groups, following a fairness definition of avoiding unnecessary harm.

The general formulation of a Minimax optimization problem for any set of functions $f_i\in F$ from a hypothesis class F, subject to constraints $g_j\in G$, can be represented as follows:

$$\begin{aligned}&\text{Minimize} && \max_i f_i(x)\\ &\text{Subject to} && g_j(x)\leq 0,\quad j=1,2,\dots,m\end{aligned}$$

3.4 Algorithm design

The Minimax Pareto Fairness (MMPF) framework was adopted from Martinez et al. [5]. We expand the design to specifically reduce individual fairness and traditional group fairness metrics by including a constraint optimization target in addition to the accuracy/risk target Martinez et al. employ. The MMPF framework aims to minimize the risk of the worst-performing sensitive subgroup by applying a training weight to each group. It applies to convex optimization targets, such as Cross-Entropy, where the Pareto front can be described as a linear reweighing problem of individual optimizations. In the case of non-convex optimization targets, this weighing approach may only identify solutions to a subset of the Pareto front.

In the first step, a given model is trained based on an accuracy metric such as Binary-Cross-Entropy (BCE). In the second step, loss functions (or risks) are calculated and assessed per sensitive subgroup. Based on the individual group risk scores, the authors propose the Approximate Projection onto the Star Sets (APStar) Algorithm to find a weight vector,

Table 1 Distribution of Gender by Race in Adult Census Income dataset

Race	Female	Male
Amer-Indian-Eskimo	119	192
Asian-Pac-Islander	346	693
Black	1555	1569
White	8642	19174

Table 2 Average Age by Sex & Marital Status

Sex & Marital Status	Avg Age (years)
female, divorced/separated/married	38.84
male, divorced/separated	32.80
male, married/widowed	37.66

Table 3 Distribution of Males and Females by Marital Status

Sex & Marital Status	Count
male, married/widowed	548
male, divorced/separated	310
female, divorced/separated/married	50

which minimizes the risk for the worst-performing subgroup only if APStar converges to a global minimum. APStar is an iterative optimization algorithm, which updates the linear weight vector for all risks in each step to refine the total minimax risk. ([5] for further details on the algorithm) After each iteration, the model is retrained with the new weight vector. The weight vector is accepted if the solution is Pareto dominant for all previous solutions and rejected otherwise. This framework, by definition, achieves the best possible trade-off between group risks without degrading performance to any single group. Still, the underlying utility-based notion of fairness does not necessarily lead to better results in statistical group or individual fairness metrics.

For ML algorithms in highly regulated areas such as the financial domain, parity-based fairness metrics are increasingly used to monitor outcome fairness. It is, therefore, vital to directly address these metrics in the optimization process without degrading model performance.

We propose and test an integrated approach to discourage biased inference, where a classification model is trained directly with a multi-objective loss function, including accuracy, group fairness, and individual fairness objectives. Simultaneously, we apply the MMPF framework to find a Pareto optimal reweighing vector for data points from each protected group and ensure that the fairness constraints do not lead to unfair accuracy reductions in any sensitive subgroup.

Algorithm 1 Minimax Pareto Optimization [5] with Fairness Objectives integrated

Data: data X, target Y, sensitive attributes A, initial weights μ, risk function $r()$, fairness objectives $L_{fair}()$, K_{min}
Result: μ^*, $r^*(\mu^*)$
Initialize:
  $h, r(\mu), L_{fair}(h,\mu) \leftarrow \min_h \sum_a \mu_a r_a(h) + L_{fair}(h,\mu)$;
$\overline{r} = ||r(\mu)||_\infty$;
$K \leftarrow 1$;
repeat
  APStar-algorithm:
   $1_\mu \leftarrow \{l(r_a(\mu) \leq \overline{r}\}_{a=1}^{|A|}$
   $\mu \leftarrow (\alpha\mu + \frac{1-\alpha}{K||1_\mu||_1} 1_\mu)\frac{K}{(K-1)\alpha+1}$
   Train model on multi-objective loss function:
   $h, r(\mu), L_{fair}(h,\mu) \leftarrow \min_h \sum_a \mu_a r_a(h) + L_{fair}(h,\mu)$; $K \leftarrow K+1$;
  if $||r(\mu)||_\infty \leq \overline{r})$ **then**
    update Pareto-optimal weights:
    $h^*, \mu^*, r^*, L_{fair} \leftarrow h, \mu, r(\mu), L_{fair}$;
  end
until *convergence*;

Table 4 Classifier Performance Summary

Classifier	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.71	0.93 / 0.64	0.46 / 0.96	0.62 / 0.77
Random Forest	0.80	0.88 / 0.75	0.71 / 0.90	0.79 / 0.82
Logistic Regression	0.85	0.88 / 0.83	0.82 / 0.88	0.85 / 0.86
Decision Tree	0.86	0.87 / 0.85	0.85 / 0.87	0.86 / 0.86
K-Nearest Neighbors	0.87	0.94 / 0.81	0.79 / 0.94	0.86 / 0.87
SVM	0.87	0.90 / 0.84	0.82 / 0.91	0.86 / 0.87
XGBoost	0.90	0.91 / 0.89	0.89 / 0.91	0.90 / 0.90

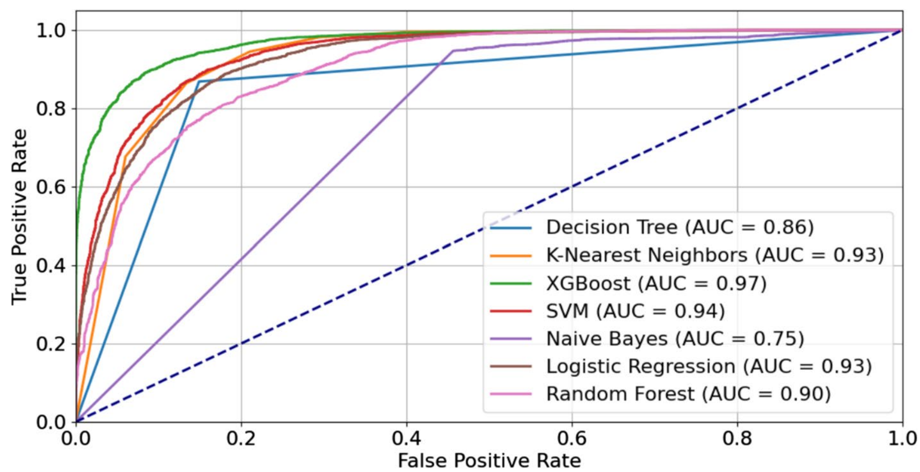

Fig. 1 AUC-ROC Curve on Adult Income dataset

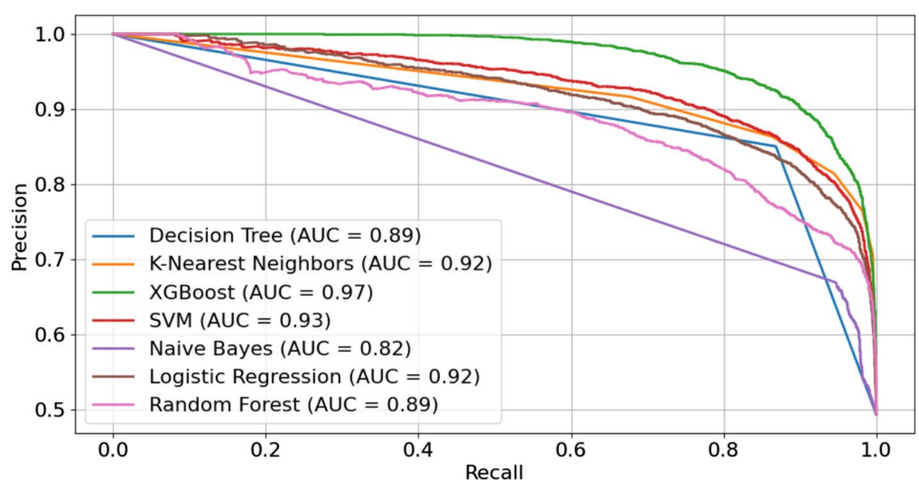

Fig. 2 Precision-Recall Curve on Adult Income dataset

Table 5 Classifier Performance Summary

Classifier	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.725	0.726 / 0.700	0.726 / 0.727	0.726 / 0.727
Random Forest	0.760	0.726 / 0.675	0.726 / 0.675	0.726 / 0.675
K-Nearest Neighbors	0.660	0.543 / 0.523	0.543 / 0.523	0.543 / 0.523
SVM	0.760	0.724 / 0.680	0.724 / 0.680	0.724 / 0.680
Decision Tree	0.725	0.689 / 0.707	0.689 / 0.707	0.689 / 0.707
Random Forest	0.760	0.726 / 0.675	0.726 / 0.675	0.726 / 0.675
XGBoost	0.765	0.726 / 0.701	0.726 / 0.701	0.726 / 0.701

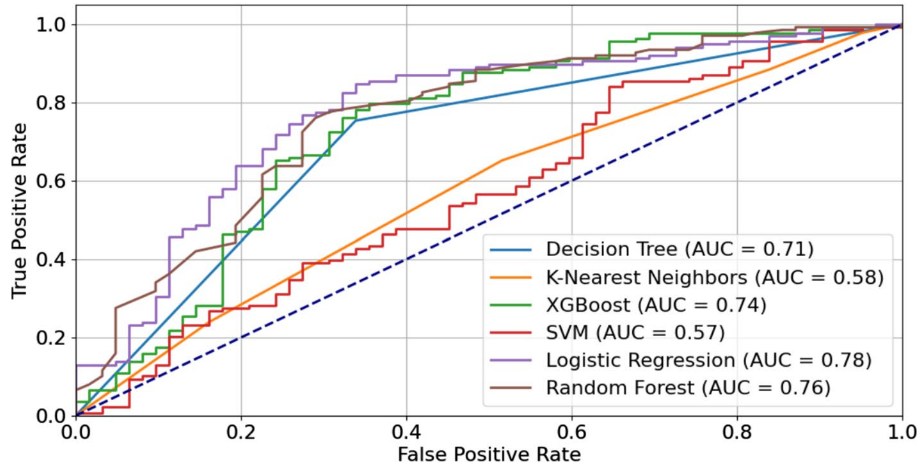

Fig. 3 AUC-ROC Curve on German Credit dataset

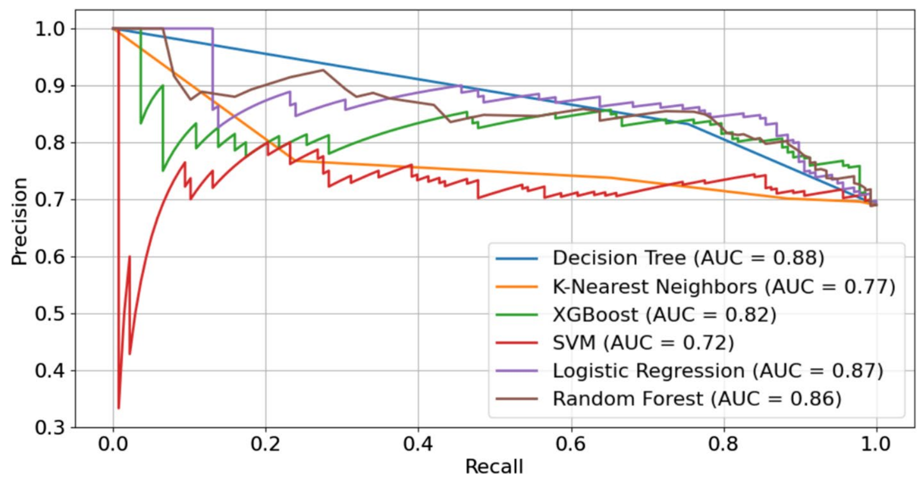

Fig. 4 Precision-Recall Curve on German Credit dataset

The target functions for model loss and subgroup risks are exchangeable and can be adapted to reflect the requirements of a particular use case.

The risk per subgroup represents a utility function for which we guarantee "no fairness" by optimizing the worst group risk. We optimize for accuracy and other potential targets, including recall and precision. The optimization method and loss function to train the model are independent of the Pareto optimization and, therefore, fully customizable. Our experimentation is based on Demographic Parity as a group fairness objective, but Equalized Odds can be used interchangeably in our implementation, and further extensions are possible. Likewise, we utilize a metric-based measure for individual fairness, though in principle, the framework allows for different individual fairness mitigation techniques to be applied.

3.5 Multi-objective loss function

The multi-objective loss function employed in the above Minimax Pareto Fairness (MMPF) algorithm is designed to address accuracy, group fairness, and individual fairness simultaneously. It is expressed as an unweighted sum of the Binary Cross-Entropy (BCE), Group Fairness and Individual Fairness.

The multi-objective loss function is given by:

$$
\begin{aligned}
&\text{Multi-Objective Loss} \\
&= -\frac{1}{N}\sum_{i=1}^{N}\left(y_i\log(p_i) + (1-y_i)\log(1-p_i)\right) \\
&\quad + \lambda_1 \cdot \max\left(0, P(\hat{Y}=1|A=a) - P(\hat{Y}) - k\right) \\
&\quad + \lambda_2 \cdot \max\left(0, -P(\hat{Y}=1|A=a) + P(\hat{Y}) - k\right) \\
&\quad + \frac{\sum_{(x,x')\in S} I((d(x,x')\leq\epsilon)\cap(h(x)=h(x')))}{\sum_{(x,x')\in S} I(d(x,x')\leq\epsilon)}
\end{aligned}
\tag{5}
$$

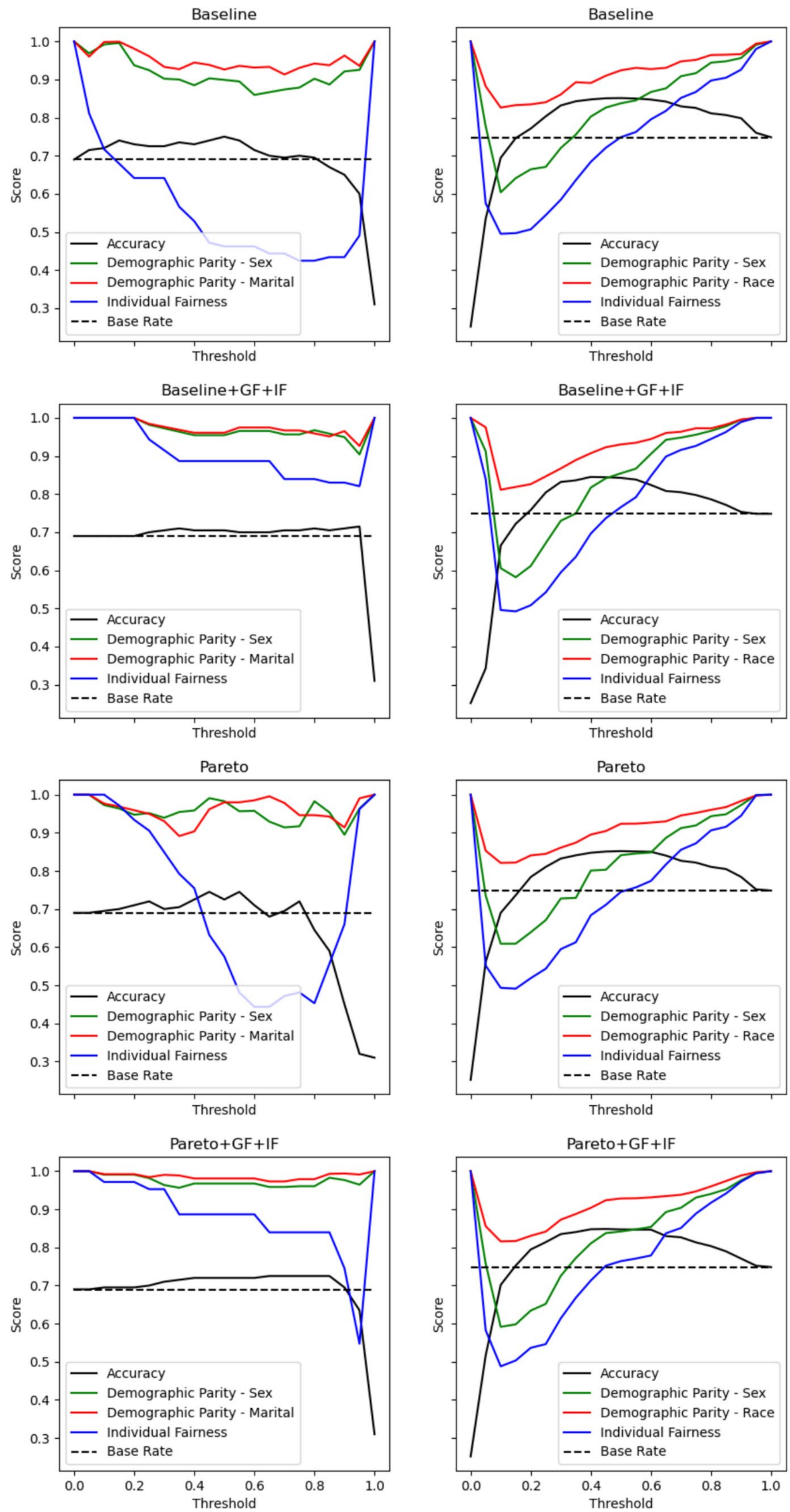

Fig. 5 Threshold plots for accuracy, demographic parity, and individual fairness for the German dataset (left) and Adult dataset (right). All scores are transformed so that a score of 1.0 is optimal

Table 6 Performance Assessment - Adult. We report overall accuracy as well as accuracy and precision by sensitive subgroup (1.Caucasian-male 2.Caucasian-female 3.Non-Caucasian-male 4.Non-Caucasian-female")

Classifier	Acc.	Acc. Gen—Ethnic	Precision Gen—Ethnic
Boosted Tree model	0.853	0.927/0.817 — 0.906/0.846	0.856/0.802 — 0.0839/0.804
Boosted Tree model + GF+IF	0.841	0.921/0.802 — 0.900/0.831	0.854/0.804 — 0.862/0.805
Minimax Pareto	0.851	0.926/.816 — 0.906/0.842	0.813/0.795 — 0.832/0.793
Minimax Pareto + GF+IF	0.850	0.924/0.815 — 0.905/0.841	.0902/0.826 — 0.888/0.829

where λ_1 and λ_2 are regularization parameters that control the importance of each constraint in the overall loss. We conducted a simple grid search to optimize the regularization parameters for the loss function, observing that high regularization resulted in uninformative predictions due to the imbalanced nature of the dataset, where positive outcomes were underrepresented.

3.6 Dataset description

3.6.1 Adult census income dataset

The Adult Census Income dataset comprises 32561 instances and 15 features, which include a mix of categorical and numerical variables. Before the analysis, the dataset underwent standard preprocessing, where missing values were handled through mean/mode imputation, categorical features were one-hot encoded, and numerical features were scaled to ensure uniformity. The distribution of race and gender is shown below (Tables 1, 2, 3).

Exploratory data analysis revealed a large imbalance between high and low income, where many individuals were earning less than $50,000. Additionally, we observed that education level and marital status are strongly correlated with income.

3.6.2 German credit dataset

The German Credit Data contains data on 20 variables and the classification of whether an applicant is considered a Good or Bad credit risk for 1000 loan applicants. The distribution of gender and marital status with age is shown below.

The German Credit dataset contains information on individuals categorized by their "Sex & Marital Status." We have observed the following distribution:

From the above breakdown, it is evident that most individuals in the dataset are males, particularly those who are married or widowed. On the other hand, the number of females, especially those who are divorced, separated, or married, is notably lower.

Table 7 Loss Assessment—Adult. We report Binary Cross-Entropy (BCE), Demographic Parity Loss and Individual Fairness Loss based on logit-predictions

Classifier	BCE	DP Loss	Ind.F.Loss
Boosted Tree model	0.326	0.061	0.088
Boosted Tree model + GF+IF	0.357	0.046	0.086
Minimax Pareto	0.327	0.066	0.100
Minimax Pareto + GF+IF	0.341	0.061	0.103

Table 8 Fairness Outcomes—Adult. We report the highest accuracy disparity between subgroups, demographic parity for gender and ethnicity respectively as well as Individual Fairness

Classifier	Acc. Disparity	Dem.Parity	Ind.Fairness
Boosted Tree model	0.145	0.161/0.067	0.006
Boosted Tree model + GF+IF	0.157	0.142/0.061	0.008
Minimax Pareto	0.152	0.155/0.068	0.006
Minimax Pareto + GF+IF	0.146	0.149/0.064	0.007

4 Results

4.1 Adult income dataset

Multiple tree-based classifiers were tested for accuracy, precision, Recall, and F1 scores to select the baseline model for the classification task on the out-of-sample dataset.

Table 4 evaluates the performance of different classifiers in terms of their accuracy as well as balanced precision, recall and F1 scores for positive and negative outcomes respectively. Since the XGBoost tree-based classifier model outperforms other models, it becomes the obvious choice

Table 9 Performance Assessment—German. We report overall accuracy as well as accuracy and precision by sensitive subgroup (1.married/divorced males 2.single males 3.married/divorced females 4.single females)

Classifier	Acc.	Acc. Gen - Marital	Precision Gen
Boosted Tree model	0.75	0.67/0.71 — 0.8/.58	0.67/0.71 — 0.80/0.58
Boosted Tree model + GF+IF	0.72	0.67/0.71 — 0.75/.58	0.67/0.71 — 0.76/0.58
Minimax Pareto	0.76	0.56/0.73 — 0.82/.67	0.63/0.73 — 0.8/0.64
Minimax Pareto + GF+IF	0.74	0.67/0.71 — 0.78/.58	0.67/0.71 — 0.78/0.58

to be the baseline model. This baseline model achieves an accuracy of 0.90 and an AUC score of 0.97 (see Fig. 1), thereby demonstrating its strong predictive capability.

In Fig. 1, XGBoost stands out with the highest AUC score of 0.97, followed by SVM with an AUC score of 0.94. On the other hand, the Decision Tree and Naive Bayes exhibit lower AUC scores which indicate their inability to effectively distinguish the lower income groups in the ADULT Census Income dataset.

In Fig. 2, the higher AUC-PR values, such as those achieved by XGBoost (0.97), imply a very strong ability to balance precision and recall thus offering a compelling choice to be utilized as the baseline model.

4.2 German credit dataset

Multiple classifier architectures were tested for accuracy, precision, Recall, and F1 scores to select the baseline model for the classification task on the out-of-sample dataset.

From Table 5, we can observe that the XGBoost tree-based classifier model outperforms other models thus, it reaffirms our obvious choice to be the baseline model.

In Fig. 3, we used bunch of classifiers on the German Credit dataset and calculated the AUC-ROC graph.

In Fig. 4, the higher AUC-PR values, such as those achieved by XGBoost (0.97), imply a very strong ability to balance precision and recall thus offering a compelling choice to be utilized as the baseline model.

Table 10 Loss Assessment—German. We report Binary Cross-Entropy (BCE), Demographic Parity Loss and Individual Fairness Loss based on logit-predictions

Classifier	BCE	DP Loss	Ind.F.Loss
Boosted Tree model	0.598	7.908	0.081
Boosted Tree model + GF+IF	1.112	0.06	0.0
Minimax Pareto	0.487	4.16	0.021
Minimax Pareto + GF+IF	0.832	0.256	0.003

4.3 Fairness-enhanced Pareto optimization

We applied our proposed framework (**Algorithm 1**) to the ADULT Census Income dataset and the German Credit risk dataset to assess the performance of the different optimization targets and demonstrate the achievable fairness and accuracy levels when considering individual and group fairness during training. We utilize Binary-Cross-Entropy (BCE) as performance loss, Demographic Parity (DP loss) for group fairness, and an individual fairness loss (Ind.F.loss, see Sect. 3.2). For the group fairness assessment, we consider gender (Male/Female) and ethnicity (White/Other) for the Adult dataset and gender (Male/Female) and Marital status (single vs married/divorced) to be sensitive attributes. Individual fairness is defined here as being independent of sensitive attributes, and we argue that differences in sensitive features should not be used to define a similarity metric in this context. The objective is maintaining the model's predictive power while concurrently improving fairness conditions in the outcomes. To this end, we assess model performance regarding achieved loss, accuracy, precision, group fairness and individual fairness. The highly imbalanced nature of both dataset limits the group fairness improvements achievable through in-training mitigation methodology without degrading performance. In practical applications, additional data pre-processing steps are called for to

Table 11 Fairness Outcomes—German. We report the highest accuracy disparity between subgroups, demographic parity for gender and marital status, respectively as well as Individual Fairness

Classifier	Acc.Disparity	Dem.Parity	Ind.Fairness
Boosted Tree model	0.219	0.042 — 0.04	0.104
Boosted Tree model + GF+IF	0.128	0.031 — 0.042	0
Minimax Pareto	0.256	0.0 — 0.005	0.028
Minimax Pareto + GF+IF	0.197	0.042 — 0.04	0

achieve better fairness outcomes on datasets with a large representational bias. In this work, we focus on in-training mitigation only to directly assess the effect of the proposed framework without confounding pre- and post-processing methods.

We begin by assessing the trade-offs between accuracy, group fairness by sensitive subgroup and individual fairness for different thresholds for both datasets (Fig. 3). For the German dataset, we observe a clear positive effect of including fairness targets as constraints on individual and group fairness scores achieved across thresholds. The profile of trade-offs changes significantly across optimization approaches. Of these four approaches, the Pareto optimization with Fairness constraints is the only one yielding good group and individual fairness scores while effectively limiting performance trade-offs. In change, the trade-off profile for the Adult dataset is largely independent of the optimization method in our experiments. Depending on the threshold, we observe improvements in one or more metrics. In the following section, we will dissect these results in more detail, where we choose a threshold with an optimal trade-off profile for each experiment. Thresholds constitute the model with a predictive power above the outcome base rate of the test dataset. (Fig. 5)

4.3.1 Results for adult income dataset

With the selected thresholds, we examine the model performance on the held-out validation dataset, assessing whether fairness performance generalizes beyond the training population. Before applying the Pareto Minimax optimization, we observe an accuracy-fairness trade-off, where the protected groups for both gender and ethnicity achieve lower accuracy after applying unfairness mitigation in the training process (Table 6, rows 1 and 2). This highlights the important role of utility-based fairness metrics, as higher group fairness scores do not directly translate into desirable performance improvements for members of protected subgroups. Indeed, here we find all groups are treated more poorly as a result of mathematical fairness constraints.

Combining Pareto optimization with fairness mitigation effectively mitigates the performance decline. Here, we observe accuracy decreases of less than.6 % and increased precision for all subgroups compared to the results achieved without Pareto optimization (Table 8).

To examine the direct impact of the optimization targets on performance, we also examine the loss achieved on the validation dataset based on the raw probabilistic outputs of the model candidates. The group fairness metrics in this setting improve by 7.5% for gender and 4.5% for ethnicity, albeit not to the same level as before the employment of Pareto optimization (11.8% for gender, 9% for ethnicity). This is also apparent in the achieved demographic parity loss before and after including the Pareto Minimax optimization for each sensitive feature (Table reft7). While this shows that the no-harms definition of fairness is a competing optimization target to the traditional measure of Demographic Parity in this example, the high-performance scores and the improved group fairness scores indicate the merit of our multi-objective approach. The Minimax Pareto algorithm with group and individual fairness objectives (Pareto+GF+IF) achieves the best balance between equal subgroup performance and Demographic parity scores out of the compared frameworks.

Comparing the Pareto Minimax optimization before and after the inclusion of fairness loss functions, we find nearly equal performance with improved group fairness (Table 6, rows 3 and 4), indicating that the additional consideration of group fairness does not harm performance while improving fairness.

The achieved levels of individual fairness based on the binary classification outcome are effectively constant for all tested models. The examination of the individual loss function, based on the predicted probabilities instead of binary classification, reveals a small effect of the constraint in maintaining or improving individual fairness outcomes for models that concurrently seek to optimize group and individual fairness.

We have assessed the effects of combining three different fairness definitions within one optimization framework, highlighting complementary and competing aspects of the optimization. The results suggest that a combination of utility-based and traditional fairness metrics may enable practitioners to achieve high performance across subgroups while improving fairness metrics.

4.3.2 Results for German credit dataset

With the selected thresholds, we examine the model performance in terms of binary predictions on the held-out validation dataset (German Credit) similarly. Before applying the Pareto Minimax optimization, we observe an accuracy-fairness trade-off, where the protected groups for both gender and ethnicity achieve lower accuracy after applying unfairness mitigation in the training process (Table 9, rows 1 and 2) of 3%, driven by poorer performance in the subgroup of single males, which represents the group with highest overall performance.

Combining Pareto optimization with fairness mitigation minimizes the performance decline without hurting any of the other groups. Here, we observe accuracy decreases of less than 1 % (Table 9). Interestingly, Pareto optimization without fairness constraints increases the overall accuracy slightly while reversing the performance imbalance between married/divorced males and single females. While all examined fairness metrics indicate that the Boosted Tree model with Fairness constraints most effectively addresses unfairness in the model, this comes at the cost of reduced accuracy(Tables 6, 7, 8, 9). Pareto optimization without fairness constraints decreases group and individual unfairness metrics while achieving high accuracy but does not reduce the accuracy disparity between the highest and lower-performing subgroups.

To further examine the training procedure independent of the threshold, we look at the accuracy and fairness losses achieved on the validation dataset (Table 10). Here, we observe the relative performance of all model candidates. Including fairness constraints in the baseline model (Boosted Tree model+GF+IF) leads to decreased fairness loss at the expense of performance. Pareto optimization achieves the highest performance while slightly improving traditional fairness. Integrating Pareto optimization with fairness constraints (Pareto+GF+IF) achieves a balance in favor of fairness, decreasing the performance trade-off while achieving notable fairness improvements compared to the baseline (Table 11).

5 Future work

Our work provides preliminary evidence supporting the combination of different in-training unfairness mitigation strategies to train fair models with optimal accuracy-fairness trade-offs. We find the differential performance of the algorithms in the Adult Income and the German Credit datasets likely due to different data-inherent patterns and biases. To guide practitioners in selecting appropriate fairness considerations for a given dataset, it is important to begin to identify the specific disbalances and patterns that an algorithm can effectively address and those where an optimization schema fails to produce good results. Further validation on additional datasets with different patterns and imbalances is required to understand the specific trade-offs of accuracy and the proposed fairness definitions, including statistical and utilitarian group fairness and individual fairness.

Theoretically, our framework can include any fairness metric that can be expressed as a constraint; however, we considered only two fairness metrics in our framework. Hence, integrating other group and individual fairness metrics is of our interest to include in our future work in this domain. Furthermore, combining the explored in-training mitigation with pre-and/or post-processing mitigation strategies is another future work direction we identified to be studied. This strategy may further improve the achievable results for performance and fairness. The presented research's purpose was to explore a framework that combines different fairness-aware training methods to obtain optimal performance-fairness trade-offs. Within the presented framework, differential weights for each fairness objective on the level of the model may allow for fine-tuning of the desired trade-offs. Another potential avenue for further development is to adjust the Pareto-reweighing procedure to allow for individual weighting of data points.

Further, alternative optimization methods, such as grid search or linear programming approaches, may offer different benefits than the chosen Pareto solution to address the trade-off between fairness and performance measures. In the context of responsible AI, exploring the effects of different fairness considerations in terms of explainability and interoperability would also be of interest.

6 Conclusion

This research has examined the application of the Minimax Pareto framework to optimize individual and group fairness within the context of classification tasks. We have assessed the specific performance of Pareto-enhanced Fairness frameworks on the Adult Census and the German Credit dataset, comparing the achieved results to a fairness-unaware baseline model and simple Pareto and Fairness-constraint optimization. While we find the specific trade-offs to be dataset dependent, the combination of explicit individual and group fairness constraints with Pareto optimization does lead to an improved accuracy/fairness trade-off, favouring fairness in our experiments. The results also emphasize the need for use case specific fairness-aware optimization. Fairness cannot be addressed with a one-size-fits-all mentality, but requires careful consideration of the risks and goals of a use case and an examination of the data-inherent patterns. This work lays the foundation for future work to emphasize the need for a deeper understanding of the optimization trade-offs between various fairness objectives and accuracy. Pareto-reweighing of fairness objectives at the model level may serve to fine-tune this trade-off. Additionally, the potential to customize the Pareto-reweighing procedure for individual data points represents a promising avenue of exploration. In conclusion, our work contributes to the ongoing effort to build responsible and transparent machine-learning systems that can be applied to critical decision-making processes.

Funding Open Access funding provided by the MIT Libraries.

Data availability References [1] (Adult Income Dataset) and [2] (German Credit Dataset) link the datasets we use for this research work, which are open-sourced and publicly available.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

1. Adult census income dataset, uci machine learning repository. UCI Machine Learning Repository (1996).
2. Hofmann, H.: German credit data. UCI Machine Learning Repository (1994).
3. Carey, A.N., Wu, X.: The fairness field guide: Perspectives from social and formal sciences. arXiv, Mar **8**, 2022 (2023).
4. Tubella, A.A., Barsotti, F., Koçer, R.G., Mendez, J.A.: Ethical implications of fairness interventions: what might be hidden behind engineering choices? Ethics Inf Technol **24**, 1 (2022). https://doi.org/10.1007/s10676-022-09636-z
5. Martinez, N., Bertran, M., Sapiro, G.: Minimax pareto fairness: a multi objective perspective. arXiv, Nov **3** (2023).
6. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. Jun **09** (2017). https://doi.org/10.1145/3097983.309809
7. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. arXiv, Jul **16**, 2018 (2023).
8. Yan, B., Seto, S., Apostoloff, N.: Forml: Learning to reweight data for fairness. arXiv, Jul **19**, 2022 (2023).
9. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. arXiv, Nov **3**, 2017 (2023).
10. Lohia, P.K., Ramamurthy, K.N., Bhide, M., Saha, D., Varshney, K.R., Puri, R.: Bias mitigation post- processing for individual and group fairness. arXiv, Dec **14**, 2018 (2023).
11. Z. Zhang, S.W., Meng, G.: A review on pre-processing methods for fairness in machine learning. Springer International Publishing, 1185–1191 (2023).
12. Xu, Z., Liu, J., Cheng, D., Li, J., Liu, L., Wang, K.: Disentangled representation with causal constraints for counterfactual fairness. arXiv, Aug **19**, 2022 (2023).
13. Zemel, R., Wu, Y.L., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. arXiv, Jul **28**, 2013 (2013).
14. Binns, R.: On the apparent conflict between individual and group fairness. arXiv, Dec **14**, 2019 (2023).
15. Shahsavarifar, R., J.Chandran, Inchiosa, M., A.Deshpande: Identifying, measuring, and mitigating individual unfairness for supervised learning models and application to credit risk models. arXiv, Nov **11** (2022).
16. Li, X., Wu, P., Su, J.: Accurate fairness: improving individual fairness without trading accuracy. AAAI **37**(12), 14312–14320 (2023). https://doi.org/10.1609/aaai.v37i12.26674
17. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. arXiv, Nov **28**, 2011 (2023).
18. Liu, S., Vicente, L.N.: Accuracy and fairness trade-offs in machine learning: a stochastic multi- objective approach. arXiv, Mar **18**, 2022 (2023).
19. Valdivia A, C.J. Sánchez-Monedero J: How fair can we go in machine learning? assessing the boundaries of accuracy and fairness. Int J Intell Syst. **36**, 1619–1643 (2021). https://doi.org/10.1002/int.22354
20. Yu, G., Ma, L., Du, W., Du, W., Jin, Y.: Towards fairness-aware multi-objective optimization. arXiv, Jul **22**, 2022 (2023).
21. Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. arXiv, Mar **23**, 2017 (2023).
22. C. C. Aggarwal, A.H., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. Springer Berlin Heidelberg **1973**, 420–434 (2001). https://doi.org/10.1007/3-540-44503-X_27

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.