MELBA Journal Machine Learning for Biomedical Imaging

A Multi-Objective Evaluation Framework for Analyzing Utility-Fairness Trade-Offs in Machine Learning Systems

Gökhan **Özbulak** [1,2], Oscar **Jimenez-del-Toro** [1], Maíra **Fatoretto** [3], Lilian **Berton** [3], André **Anjos** [1]

1 Idiap Research Institute, Martigny, Switzerland
2 École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
3 Federal University of São Paulo (UNIFESP), São Paulo, Brazil

Abstract

The evaluation of fairness models in Machine Learning involves complex challenges, such as defining appropriate metrics, balancing trade-offs between utility and fairness, and there are still gaps in this stage. This work presents a novel multi-objective evaluation framework that enables the analysis of utility-fairness trade-offs in Machine Learning systems. The framework was developed using criteria from Multi-Objective Optimization that collect comprehensive information regarding this complex evaluation task. The assessment of multiple Machine Learning systems is summarized, both quantitatively and qualitatively, in a straightforward manner through a radar chart and a measurement table encompassing various aspects such as convergence, system capacity, and diversity. The framework's compact representation of performance facilitates the comparative analysis of different Machine Learning strategies for decision-makers, in real-world applications, with single or multiple fairness requirements. In particular, this study focuses on the medical imaging domain, where fairness considerations are crucial due to the potential impact of biased diagnostic systems on patient outcomes. The proposed framework enables a systematic evaluation of multiple fairness constraints, helping to identify and mitigate disparities among demographic groups while maintaining diagnostic performance. The framework is model-agnostic and flexible to be adapted to any kind of Machine Learning systems, that is, black- or white-box, any kind and quantity of evaluation metrics, including multidimensional fairness criteria. The functionality and effectiveness of the proposed framework are shown with different simulations, and an empirical study conducted on three real-world medical imaging datasets with various Machine Learning systems. Our evaluation framework is publicly available at `https://pypi.org/project/fairical`.

Keywords

Machine Learning, Multidimensional Fairness Evaluation, Multi-Objective Optimization, Utility-Fairness Trade-off, Medical Image Analysis

Article informations

`https://doi.org/10.59275/j.melba.2025-ab9a`

Volume 3, Received: 2025-04-21, Published 2025-12-30

Corresponding author: gokhan.ozbulak@idiap.ch

Special issue: Special issue on Fairness of AI in Medical Imaging (FAIMI)
Guest editors: Veronika Cheplygina, Aasa Feragen, Andrew King, Ben Glocker, Enzo Ferrante, Eike Petersen, Esther Puyol-Antón, Melanie Ganz-Benjaminsen

1. Introduction

The increasing integration of Machine Learning (ML) systems in day-to-day activities offers significant opportunities, but it also raises critical concerns regarding demographic fairness and equity (Xinying Chen and Hooker, 2023; Pessach and Shmueli, 2022; Starke et al., 2022; Rabonato and Berton, 2025). Fairness in ML pertains to the ethical imperative of ensuring that algorithms and models do not discriminate or display bias, against individuals or groups based on lawfully demographic attributes such as race, gender, age, and/or other characteristics (Barocas et al., 2023). Fairness is a multi-faceted and complex concept with nuances directly linked to the situation considered (Castelnovo et al., 2022; Pessach and Shmueli, 2022). Consequently, balancing and measuring multiple fairness criteria simultaneously, both at a group and individual level, is a challenging task, resulting in different definitions of fairness appropriate for different contexts (Dutt et al.,

2023).

While Multi-Objective Optimization (MOO) provides a mathematical foundation for balancing objectives in conflict, fairness in ML extends beyond conventional optimization paradigms. Equity encompasses multiple, often conflicting notions, such as group or individual fairness and equality of opportunity, that reflect distinct ethical and social considerations. These criteria cannot be meaningfully represented by a single objective function, as optimizing for one dimension of fairness may amplify disparities in another. In this respect, ML based approaches that model and evaluate the utility-fairness trade-off across multiple objectives can provide valuable insight into how fairness manifests under different operational and demographic conditions. However, as highlighted by Selbst et al. (2019) in five traps, over-abstracting fairness into purely mathematical formalizations can have a risk of detaching it from the reality. Therefore, such analyses should be interpreted as structured, assumption-bound explorations of fairness dynamics, rather than universal solutions to the fairness problem in ML.

An unintended outcome when optimizing for a balanced treatment between genders is the variation in predictive performance across other groups of demographic attributes. Many techniques in real-world scenarios improve fairness at the expense of model utility, with the minimum possible error of any fair classifier bounded by the difference in base rates (Zhao and Gordon, 2022). This fundamental tension in algorithmic fairness has been previously explored to improve the understanding of model bias and the limits of artificial intelligence (AI) (Wei and Niethammer, 2022; Wang et al., 2020). Model biases can be introduced through the optimization of certain objectives, hyperparameter tuning, or simply due to the inherent characteristics of the datasets used to train the models (Yang et al., 2024), such as biased training data, sampling bias, label bias, exclusion, or historical bias. For example, if features correlated with sensitive attributes (*e.g.*, gender, race) are considered during training, the models might latch onto these attributes, potentially resulting in demographically biased outcomes.

The selection and modeling of fairness criteria are relatively new research directions that require clear definitions on the mathematical expression of demographic equity. Most ML approaches are currently evaluated without considering any fairness criteria (Akter et al., 2022; Lu et al., 2020). In recent years, some works started using unique levels of fairness and utility, which fail to characterize ML systems at every level of the utility-fairness trade-off, thus limiting subgroup and intersectional evaluation (Buolamwini and Gebru, 2018). The situation worsens when multiple fairness and utility criteria come into play, as the fairness-aware stakeholders and decision-makers may want to ensure, beyond utility, that multiple fairness criteria are satisfied, such as race, gender, and age aiming to provide fair ML services in production under different demographic settings.

In medical imaging, fairness challenges become more prominent because diagnostic processes often involve multiple and conflicting clinical and operational objectives, as ML models are increasingly used to support decisions in almost every field, such as radiology, ophthalmology, and dermatology. Furthermore, variations in data acquisition conditions, patient demographics (*e.g.*, the deployment of diagnostic tools in a hospital that tends to a heterogeneous community), and disease prevalence can lead to systematic biases in model predictions. For instance, screening models for ophthalmology risk must retain high sensitivity to avoid missed cases, while also ensuring that performance does not systematically degrade for specific demographic subgroups. A representative example arises in glaucoma, an eye disease that exhibits a higher prevalence among Black populations, and within this group, male individuals show greater vulnerability compared to females (Khachatryan et al., 2019; Luo et al., 2024). The high prevalence of glaucoma contrasts with the scarcity of available data from the Black community, contributing to disparities in model performance across racial groups and even between genders within the same group. This mismatch highlights the necessity of evaluating ML models not only for diagnostic utility but also for their equitable behavior across different demographic attributes. Consequently, utility (diagnostic performance) and multiple fairness constraints must be considered simultaneously. A principled way to analyze such conflicting requirements is to employ multi-objective formalizations, where each fairness criterion and the utility metric are treated as separate but jointly optimized objectives rather than in isolation.

On the other hand, the development of ML systems addressing multiple objectives has been extensively studied in the context of MOO systems. In such cases, performance is evaluated by considering all possible trade-offs between the individual objectives, resulting in an N-dimensional graph. This evaluation strategy can provide a basis to better articulate user preferences in model comparison (Gong and Guo, 2023). Although multi-objective measurements have been used in recent works to incorporate single fairness constraints into developed models (Little, 2023), to the best of our knowledge, there are no frameworks that enable a comprehensive comparison of ML systems under multiple utility and fairness criteria. In particular, multiple fairness considerations based on MOO become even more critical in the context of medical imaging. A framework that accounts for multiple fairness constraints simultaneously is therefore essential to ensure equitable diagnostic performance across different patient subgroups

characterized by demographic attributes.

Conceptually, fairness in ML can be regarded as a multidimensional evaluation problem rather than a single optimization objective. Instead of focusing on maximizing a particular fairness metric, it is often more informative to examine how ML systems perform across a spectrum of utility-fairness trade-offs. Such an analysis enables a clearer understanding of both ideal cases, where equity across demographic groups is maximized and practical situations in which certain fairness dimensions may be compromised due to data, operational, or design constraints. Considering fairness in this way allows researchers and practitioners to interpret model behavior within an explicit space of trade-offs, rather than through isolated or averaged/maximized performance scores. This perspective establishes the conceptual foundation for evaluating fairness-aware ML systems in a structured, multi-objective manner, before introducing any specific methodological framework.

In this work, we present an evaluation framework, supported by MOO principles, for the challenging task of comparing ML systems under multiple utility-fairness trade-offs. This approach allows the comparison of multiple systems in a common multidimensional space, where multiple concurrent fairness criteria can come into play. The framework's applicability is demonstrated through use cases based on medical imaging, where fairness constraints play a crucial role in ensuring equitable diagnostic performance. Our contributions can be summarized as follows:

- A model- and task-agnostic evaluation framework with a compact yet comprehensive representation, both qualitatively and quantitatively, of multiple utility-fairness trade-offs resulting from the deployment of ML systems, facilitating system performance analysis and comparison.
- The framework integrates multiple fairness metrics into the evaluation process, providing a more nuanced and multi-faceted assessment of model performance.
- A detailed analysis of the proposed framework and rationale through simulations of typical ML systems on synthetically generated data.
- An empirical study based on three real-world medical imaging datasets demonstrating the effectiveness of the proposed framework.
- An open-source implementation of the framework allowing reproduction of results established in this article, and further reuse[1].

While the framework is demonstrated in the context of medical imaging, its formulation is model-, metric-, and domain-agnostic, and can be readily applied to other high-stakes ML systems (*e.g.*, decision-support in finance, hiring, criminal justice, or homeland security applications for biometrics) where multiple fairness and utility objectives may be in conflict. By providing a unified view of utility-fairness trade-off, the framework establishes a generalizable flow for cross-domain benchmarking and transparent model selection across diverse application areas.

The paper is organized as follows: Section 2 reviews previous works tackling the evaluation of utility-fairness trade-off systems. Section 3 thoroughly describes the proposed evaluation framework, with a set of use cases and MOO principles behind it. The applicability of the framework in real-world scenarios is shown in Section 4, with the analysis of ML systems for three medical imaging tasks. Finally, a discussion and conclusion with the key points and limitations from this study are presented in Section 5 and Section 6.

2. Background and Related Work

Fairness in machine learning can be categorized based on criteria, sources of bias, perspectives, methodologies, and trade-offs. Fairness criteria include demographic parity, equality of opportunity, equalized odds, and predictive parity (Hardt et al., 2016; Agarwal et al., 2018), each focusing on equitable outcomes or error rates across groups. Sources of bias can stem from data (*e.g.*, under-representation) (Garin et al., 2023), algorithms (*e.g.*, prioritizing accuracy/utility over fairness) (Buolamwini and Gebru, 2018), or human involvement (*e.g.*, subjective labeling) (Zhang et al., 2024). Perspectives of fairness include individual fairness (similar individuals receive similar predictions) (Dwork et al., 2012; Petersen et al., 2021), group fairness (equitable treatment across groups) (Diana et al., 2021; Chan et al., 2024), and subgroup fairness (addressing intersectional identities) (Kuratomi et al., 2025). Methodologies to enforce fairness involve pre-processing data (*e.g.*, balancing representation) (Jang and Wang, 2023; Liu et al., 2021; Lahoti et al., 2020), in-processing adjustments (*e.g.*, modifying loss functions) (Jovanović et al., 2023; Roy and Boddeti, 2019), and post-processing predictions (*e.g.*, calibration techniques) (Hardt et al., 2016; Kim et al., 2020; Jang et al., 2022). Achieving fairness often involves trade-offs, such as balancing it with utility (Liu and Vicente, 2022; Wang et al., 2021) and interpretability (Jo et al., 2023).

Fairness-aware evaluation has an increasing attention in medical imaging tasks, where model predictions can directly affect downstream diagnostic decisions. Recent studies have shown that performance can vary across demographic groups due to acquisition protocols, device related differences, or data imbalance, motivating the inclusion of fairness metrics alongside performance metrics.

1. https://pypi.org/project/fairical

For instance, in glaucoma detection, ML systems have been evaluated not only for accuracy but also for their fairness performance using equity scaling measurements (Luo et al., 2024) or with respect to demographic features such as gender (Akter et al., 2022). Similarly, multi-objective formalizations have been explored to jointly optimize imaging quality, diagnostic performance and fairness (Lu et al., 2020). However, these works typically address a single fairness notion and do not provide a general framework capable of comparing multiple ML systems under several concurrent fairness constraints, which is the gap our framework aims to bridge.

While several tools and frameworks exist for fairness assessment, their scope differs significantly from the evaluation problem addressed in this work. For instance, Fairlearn and its dashboard (Weerts et al., 2023) and FACET (Gustafson et al., 2023) provide analytics for understanding model behavior across demographic groups, but they evaluate individual model performance rather than assessing the aggregated structure of a utility-fairness trade-off system achieved with our framework. This is an important aspect as the assessment of individual models drawn from a hypothesis class $\mathcal{H}$ may be unstable, whereas evaluating the joint behavior of an entire set of models provides a more stable, comprehensive and complete representation of the fairness performance of the ML system in consideration. Little (2023) proposes a metric that summarizes the utility-fairness trade-off along a single fairness criterion; however, it does not generalize to multi-objective scenarios (multiple utility/fairness). Liu and Vicente (2022) model utility-fairness interactions from a multi-objective perspective, but their formalization focuses on optimization rather than evaluation, and does not provide model-agnostic tools for analyzing the geometry or quality of the trade-off optimality. To the best of our knowledge, no existing framework offers a unified, multi-objective evaluation protocol capable of assessing multiple utility and fairness criteria jointly, and this aspect motivates the contribution of our proposed framework.

Whereas ML systems are typically developed (and evaluated) using a single utility criterion, they are often deployed in scenarios where multiple objectives must be respected. A modern example of this condition relates to the deployment of ML systems under one or multiple demographic fairness constraints (Liu and Vicente, 2022; Zhang et al., 2021; Padh et al., 2021). In this context, we argue that evaluation techniques cross-pollinated from multi-objective optimization (MOO) offer a rich set of primitives allowing for a comprehensive performance characterization under multiple criteria that can streamline system evaluation in this realm.

The principal aim of MOO is to find solutions that lie on, or are proximate to, the set of the optimal performance points called the Pareto Front (PF), resulting in a spectrum of ideal trade-offs among the various objectives. This methodology equips decision-makers with the means to select the most favorable compromise amidst conflicting goals, fostering more informed and balanced decision-making (Wu and Azarm, 2001). The trade-off selection procedure in MOO is therefore critical and affects the quality of service for the deployed system, especially in the case of conflicting objectives. Assessing the quality of these trade-off systems is comparative and encompasses criteria such as proximity to the Pareto optimal set (convergence), the distribution/spread of the points in the objective space (diversity), and the cardinality of solutions (capacity) (Zitzler et al., 2003). These criteria are evaluated by MOO specific performance indicators that have been studied in previous works (refer to Section 3.2.1 for details) (Tan et al., 2002; Wu and Azarm, 2001; Van Veldhuizen and Lamont, 2000; Coello Coello and Reyes Sierra, 2004).

Even though the modeling of trade-off for demographic fairness-accuracy PF is well-known (Wei and Niethammer, 2022; Zietlow et al., 2022), performance indicators for the quality of the PF have rarely been exploited in the context of fairness. Yang et al. (2023) developed a bias mitigation framework that incorporated the Area Under the Curve (AUC) metric, while considering both inter- and intra-group AUC simultaneously. However, the bias mitigation framework does not provide an evaluation protocol for the utility-fairness trade-off; instead, it leverages the AUC to address fairness performance. Little (2023) proposed a scalar measure of the area under the curve from the trade-off between fairness and accuracy. The generated curve outlines the empirical Pareto frontier consisting of the highest attained accuracy within a collection of fitted models at every level of fairness. Although Little (2023) focuses on similar issues as in this study, it does not address the challenge of comparing multiple ML systems in high dimensions. Additionally, the analysis of the PF is superficial, ignoring important performance indicators for diversity and capacity, providing an incomplete evaluation of compared ML strategies. To tackle the aforementioned issues, a more flexible evaluation framework is needed to accommodate different fairness criteria and utility metrics, facilitating a straightforward comparison and analysis of results from different algorithms. The method should be model-agnostic, allowing for real-world comparisons among trade-off systems that may have been optimized using different objectives. Furthermore, since the utility goals of the model across multiple objectives often diverge from fairness goals, performance indicators of the optimal PF solutions can provide a deeper understanding of the trade-offs across these objectives (Wang et al., 2021). The proposed evaluation framework bridges the gap between these issues and their solutions, providing a comprehensive

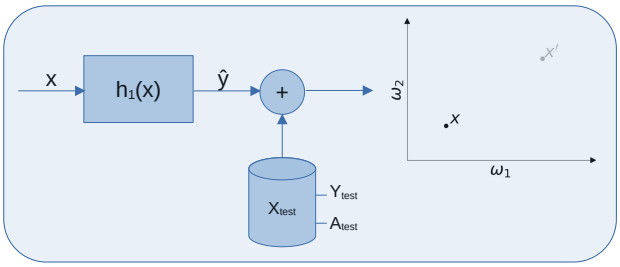


Figure 1: Black-box system evaluation (Scenario 1).

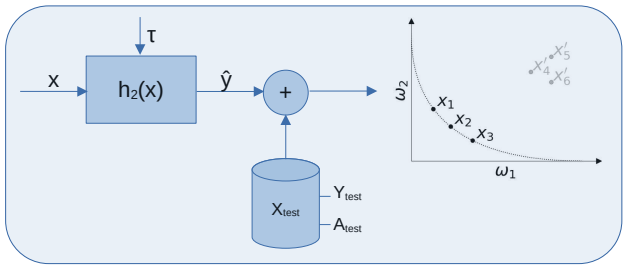


Figure 2: White-box system evaluation (Scenario 2).

guideline on how this can be achieved in the following sections.

While prior works typically quantify fairness outcomes into a scalar metric or report the best performing ML system, such simplifications may create ambiguity on the relationship between different demographic attributes. The evaluation perspective considered in this study bridges this gap by encouraging analysis across the full spectrum of utility-fairness trade-offs.

3. Methodology

In MOO, each individual objective is considered a distinct characteristic that needs to be optimized to its fullest potential. The trajectory of the optimization is determined by the objective functions in a cooperative way. Conflicting objectives increase the complexity of optimization forcing cooperation, thus making it harder to achieve optimal solutions. This trade-off is measured at evaluation time by using multiple metrics directly related to each of the objectives. Likewise, utility and fairness can result in conflicting objectives challenging the optimization of both through the same ML strategies. A good level of utility is typically achieved by sacrificing fairness or through less biased models that might reduce utility.

The proposed evaluation framework considers the trade-off between objectives for assessment and evaluates the performance of each dimension with a performance metric tailored for it. There are no limitations on using performance measurements, so any typically used metric is applicable.

3.1 Use Cases

To formalize the proposed method, we evaluate three ML use cases, which are based on two scenarios as black-box and white-box, typically found in the literature, exclusively from a deployment perspective. We explicitly assume that the ML models are trained and one is only seeking to characterize their performance from a multi-objective perspective including the model's utility and one to many fairness objectives.

The first type of scenario considers a "black-box" ML system $h_1(x) \in \mathcal{H}$ to provide binary outcomes for an input x such that $\hat{y} = h_1(x)$ where $\hat{y} \in \{0, 1\}$. To measure the approximate Pareto solution S, we assume the availability of a dataset X_{test} that carries annotations for all considered objectives, *i.e.*, the expected output of the classification Y_{test}, and demographic attributes A_{test}. Fig. 1 contains a representation of this scenario in two optimization dimensions as ω_1 and ω_2. As there is no tuning possibility to select the model ($\tau = \emptyset$), this test evaluates the solution in the deployed ML system as it is provided.

The second scenario defines the evaluation in a "white-box" manner for an ML system $h_2(x) \in \mathcal{H}$ that is tunable over prediction scores (logits) as $\hat{y} = h_2(x)$ where $\hat{y} \in [0, 1]$. Model selection in S may be achieved via τ so that a set of non-dominated solutions filtered by this parameter is available for performance assessment of the given ML system by using X_{test} alongside Y_{test} and A_{test}. This scenario is illustrated in Fig. 2 for two optimization dimensions, ω_1 and ω_2.

In the following, we demonstrate each combination of these two scenarios in three assessment based use cases. We work with two synthetically generated approximate PF solutions, *System1* and *System2*, which represent different systems across the use cases, and evaluate them in a two-dimensional setting to simulate an assessment in one direction as utility and the other as fairness. Thus, we have an overall insight into the comparative trade-off performance of different ML systems by applying possible evaluation strategies commonly encountered during the assessment.

3.1.1 UC-1 - The comparative evaluation of two black-box systems

In this use case, *System1* and *System2* are considered in a black-box manner to assess their comparative performance using the proposed evaluation framework. Since both systems are assumed to be black-box, we only have the model for each as provided and assess the trade-off performance without any tuning.

3.1.2 UC-2 - The comparative evaluation of one black-box system with one white-box system as a hybrid case

This is the use case where the comparative performance of *System1* and *System2* is evaluated in a hybrid manner by applying black-box and white-box scenarios. *System1* is assumed to be deployed as it is without any tuning capability (black-box), and *System2* can be modified to have different settings based on user preference (white-box).

3.1.3 UC-3 - The comparative evaluation of two white-box systems

This use case considers the assessment of *System1* and *System2* as white-box by tuning them according to specified preferences. Thus, we demonstrate how the trade-off capacities of the systems can be assessed when tuning is feasible and how model selection is achieved by fully leveraging their capabilities.

We perform simulations for these use cases in Section 3.6 to exemplify them quantitatively so that it is clarified how the proposed evaluation framework can be applied for such different assessment strategies given ML systems in comparison. These simulations are based on the synthetically generated systems, *System1* and *System2*, and exhibit the PF trend with non-dominated and dominated points as expected from the utility-fairness trade-off systems.

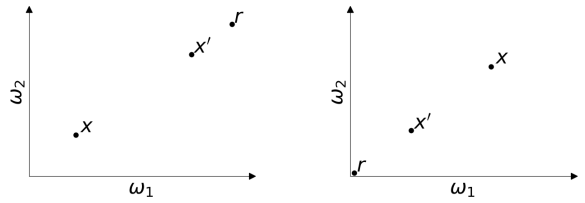
3.2 MOO Based Performance Indicators

Central to MOO is the concept of the Pareto Front (PF), which delineates the set of all Pareto optimal solutions. A solution is deemed Pareto optimal if no other solution can enhance one objective without degrading another. In this regard, solutions residing on the PF are referred to as non-dominated solutions. More formally, given two points $\mathbf{x}, \mathbf{x}'$ in the multidimensional solution space Ω ($\mathbf{x}, \mathbf{x}' \in \Omega$), the sample $\mathbf{x}$ is said to dominate $\mathbf{x}'$ ($\mathbf{x} \prec \mathbf{x}'$), if $\mathbf{x}$ is no worse than $\mathbf{x}'$ in all considered objective dimensions and is strictly better in at least one of them. Geometrically, this implies that $\mathbf{x}$ lies farther from the reference point $r \in R$, also known as the *nadir* point, which represents the worst possible outcome in the objective space. In a minimization problem, for example, $\mathbf{x}$ provides a smaller combined value for the target objective compared to $\mathbf{x}'$, see the illustration in Fig. 3.

The Pareto optimal set ($\mathcal{P}$) is the set containing all the solutions that are non-dominated with respect to Ω, defined as:

$$\mathcal{P} := \{\mathbf{x} \in \Omega \mid \nexists \mathbf{x}' \in \Omega \quad \text{such that} \quad \mathbf{x}' \prec \mathbf{x}\} \tag{1}$$

Pareto optimal solutions are called the Pareto set and the image of the Pareto set constructs the Pareto Front



(a) Minimization problem (b) Maximization problem

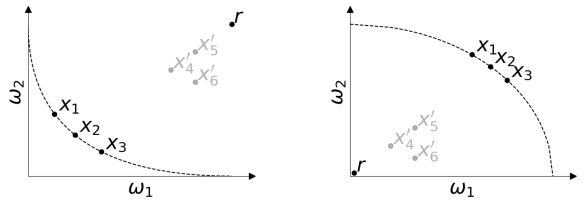
Figure 3: Dominance in bi-objective minimization (a) and maximization (b) problems: x' is dominated by x with respect to the reference point r.

(PF) (Audet et al., 2021). We note that, in real-world problems, the PF is rarely achievable. We refer to suboptimal solutions approximating the PF as S (Zitzler et al., 2003) as shown in Fig. 4. We propose expanding on this approach for evaluating ML systems under multiple fairness constraints. This approach is analogous to the analysis of Receiver Operating Characteristic (ROC) or detection-error trade-off curves in classical ML.

Whereas interpretation of a solution set S considering two optimization dimensions (ω_1, ω_2) is straightforward (Little, 2023), concurrent analysis of multiple fairness constraints is typically done (Buolamwini and Gebru, 2018; Luo et al., 2024) as a single degree of fairness by treating equity performances in isolation from one another. In this type of analysis, the dependency/correlation between different fairness criteria is not considered and the evaluation remains oversimplified. However, in a multi-task setting, every objective may conflict with each other, as one may not be improved without deteriorating others. This dependency between objectives, as is also the case for multiple fairness criteria alongside utility, should be projected into one shared space so that the multiple degrees of evaluation may be achieved in a fused way. To address this, we propose to characterize the solution set S, representing an ML system using multiple criteria from MOO. These indicators will be assembled in an easy to interpret table and a plot. Qualitative analysis can still be carried out when the number of concurrently analyzed objectives is small ($N = 2$) or when visual clutter is minimal in systems with $N > 2$. In all cases, the proposed evaluation framework via a PF characteristic remains usable.

3.2.1 The Performance Indicators

In the design of metrics for MOO, four complementary performance criteria are typically considered to analyze the PF optimality: convergence, diversity, convergence-diversity, and capacity (or cardinality) (Audet et al., 2021; Jiang et al., 2014). Measuring strict convergence, which denotes the proximity of the solution to the true PF, is not often



(a) Minimization problem (b) Maximization problem

Figure 4: The approximate PF S is shown for both bi-objective minimization (a) and maximization (b) tasks: x'_i is dominated by x_j with respect to the reference point r.

attainable; we therefore focus on the other three properties, and describe them next.

Diversity This indicator measures how well the non-dominated points are distributed or spread along the candidate solution set. Uniform Distribution (*UD*) and Overall Pareto Spread (*OS*) are some diversity measurements based on distribution and spread characteristics, respectively.

The *UD* indicator (Tan et al., 2002) evaluates the deviation characteristic of the distribution for non-dominated solutions, denoted as X_n, and is formulated as:

$$UD(S,\sigma) = \frac{1}{1 + D_{nc}(S,\sigma)} \tag{2}$$

where

$$D_{nc}(S,\sigma) = \sqrt{\frac{1}{|X_n| - 1} \sum_{i=1}^{|X_n|} \Big(nc(x^i,\sigma) - \mu_{nc(x,\sigma)} \Big)^2} \tag{3}$$

and

$$nc(x^i,\sigma) = |\{x \in X_n \mid \|x - x^i\| < \sigma\}| - 1 \tag{4}$$

σ is the niche radius that is problem dependent and can be adjusted based on the distribution of the candidate solution in the space. $\mu_{nc(x,\sigma)}$ is the mean of the niche counts, *nc*, and is defined as

$$\mu_{nc(x,\sigma)} = \frac{1}{|X_n|} \sum_{j=1}^{|X_n|} nc(x^j,\sigma) \tag{5}$$

The *UD* indicator is expected to be higher for a more uniform solution set. This indicator evaluates how uniform the solution set is spanned in the metric space based on an upper-bound distance, σ. For instance, a system with the highest *UD* value among others exhibits the best performance as its solutions are the most uniformly distributed. Having a trade-off system with a higher *UD* value corresponds to a more uniformly spanned set of non-dominated points. This increases the likelihood of achieving a desired combination of utility with fairness in tuning, compared to a system with a lower *UD*. Although the *UD* measures the coverage of the solution space by the candidate set, it fails to characterize PF as any type of uniformly distributed solution (whether Pareto optimal or not) may yield high performance in terms of this indicator.

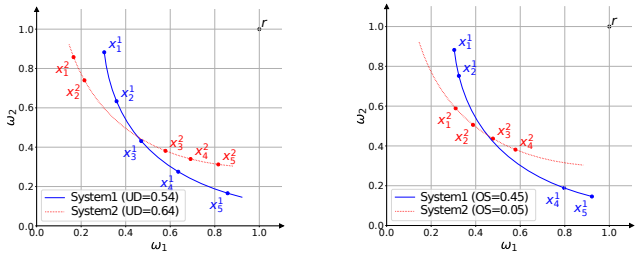
The *OS* indicator (Wu and Azarm, 2001) assesses the spread of the solutions obtained by the trade-off system. For a minimization problem evaluated in N different dimensions, this indicator is formulated as:

$$OS(S,\mathcal{P}) = \prod_{i=1}^{N} \left| \frac{\max\limits_{s \in S} s_i - \min\limits_{s \in S} s_i}{\max\limits_{p \in \mathcal{P}} p_i - \min\limits_{p \in \mathcal{P}} p_i} \right| \tag{6}$$

where the nominator and denominator are the absolute difference between the worst and best points for the candidate solution S and Pareto optimal set $\mathcal{P}$, respectively. A higher *OS* value indicates a more widely spread solution. This indicator assesses how well the points from the candidate set spread towards the ideal of the optimal PF. For instance, a system with a higher *OS* score compared to others has more points close to the ideal point and fewer ones near the *nadir* (here, we can access the *nadir* and ideal points without having the exact PF solution, so there is no requirement to know the PF *a priori*). Having a higher *OS* value exhibits a more spread characteristic for non-dominated solutions, leading to an improvement in tuning performance for the trade-off system when the selection of models around the ideal point is expected. In this study, *OS* is in the range of $[0, 1]$ and there is no transformation applied as it is scaled compatible with other indicators. Similarly to distribution, this measurement also fails to analyze Pareto optimality in a comprehensive manner as it only assesses the extreme cases without considering the entire PF space. In Fig. 5, both *UD* and *OS* indicators are exemplified in synthetically generated data. *System1* is said to have less uniformity but more spread than *System2* as its points are more equally distributed, $UD_{System1} = 0.54 < UD_{System2} = 0.64$, (Fig. 5a) and closer to the extreme points, $OS_{System1} = 0.45 > OS_{System2} = 0.05$, (Fig. 5b).

During the study, we observed that *OS* can decrease drastically when any objective fails to cover the full extent of the true PF. This sensitivity arises from the multiplicative characteristic of *OS* as a small value in one dimension sharply reduces the final score. To smooth this behavior, we introduce the Average Spread (*AS*) as a less sensitive variant of *OS*. *AS* simply replaces the multiplicative operator with summation, and is defined as:

$$AS(S,\mathcal{P}) = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\max\limits_{s \in S} s_i - \min\limits_{s \in S} s_i}{\max\limits_{p \in \mathcal{P}} p_i - \min\limits_{p \in \mathcal{P}} p_i} \right| \tag{7}$$



(a) Distribution (b) Spread

Figure 5: Diversity: (a) *System1* (blue) provides solutions that are less uniformly distributed than *System2* (red) and therefore has lower *UD*. (b) *System1* (blue) better covers the extremes of the PF approximations and therefore has better spread (larger *OS*) than *System2* (red).

Convergence-Diversity This measurement evaluates both convergence and diversity together so that the proximity (convergence) alongside the distribution/spread (diversity) of the candidate solution set is projected into a single scalar score. This unary metric measures the volume of the objective space covered by an approximation set, relying on a reference point for calculation. The Hypervolume (*HV*) (Zitzler and Thiele, 1998) takes distribution, spreading, and convergence into account at the same time, making it unique in this regard. Recognized for its distinctive properties, *HV* is Pareto-compliant, ensuring that any approximation set achieving maximum quality for a MOO contains all Pareto optimal solutions. The reference point can be simply attained by constructing a solution of worst objective function values. Given a minimization based MOO problem with two objectives, as shown in Fig. 6, it is expected to have solution sets with points that are in the best achievable state in the objective space. This should be the case even if the objectives are conflicting with each other. PF is one possible setting for such cases with non-dominated solutions. In Fig. 6, x_1 and x_2 are two non-dominated solutions drawn from the PF-like solution set (represented as a dashed curve) with one dominated solution, x_3'. The performance of such a solution set may be evaluated by the *HV* indicator to analyze how optimal the set is in terms of convergence and diversity. By discarding the dominated solution x_3', which should not be part of an optimal solution set, the *HV* indicator is calculated as the union of two volumes constructed between each of the non-dominated solutions, x_1 and x_2, and the reference point r that is chosen as one of the poorly performing solutions in the space. The *HV* formulation is then as follows:

$$HV = vol_1 \cup vol_2 = VOL\left(\prod_{i=1}^{2}[x_1^i, r^i] \cup \prod_{i=1}^{2}[x_2^i, r^i]\right) \quad (8)$$

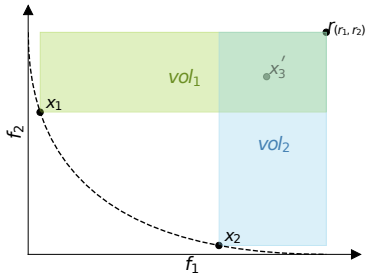


Figure 6: An approximate PF solution with two non-dominated solutions. *HV* is calculated as the union of two volumes associated with these solutions.

The formulation in (8) may be generalized as (Navon et al., 2020):

$$HV(S) = VOL\left(\bigcup_{\substack{x\in S \\ x\prec r}} \prod_{i=1}^{N}[x^i, r^i]\right) \quad (9)$$

In this study, *HV* is on the scale of $[0, 1]$ as every point in the solution space is represented by measurements between 0 and 1. An illustrative example in Fig. 7 shows how two systems are evaluated in terms of *HV*. *System1* occupies a larger volume in 2D space compared to *System2* as it is further away from the reference point. *HV* reflects this situation with $HV_{System1} = 0.55 > HV_{System2} = 0.21$.

Capacity (or Cardinality) This measurement quantifies the number of non-dominated points in the candidate solution set. The Overall Nondominated Vector Generation (*ONVG*) and Overall Nondominated Vector Generation Ratio (*ONVGR*) are commonly used capacity indicators. *ONVG*, proposed by Van Veldhuizen and Lamont (2000), is the number of non-dominated solutions, X_n, in the candidate solution set, S, and is formulated as:

$$ONVG(S) = |X_n| \quad (10)$$

As similarly proposed by Van Veldhuizen and Lamont (2000), *ONVGR* is the ratio of the non-dominated solution cardinality to that of S, and is defined as:

$$ONVGR(S) = \left|\frac{X_n}{S}\right| \quad (11)$$

Both *ONVG* and *ONVGR* yield higher scores for solution sets with greater capacity. These capacity-based indicators do not provide an extensive analysis like convergence or diversity do; they are only used as auxiliary indicators when other measurements are not discriminative. For instance, we can select a system with a higher *ONVG* over other systems when the convergence and diversity are the same for all. Furthermore, they may help analyze the effectiveness of the optimization, as a higher number of

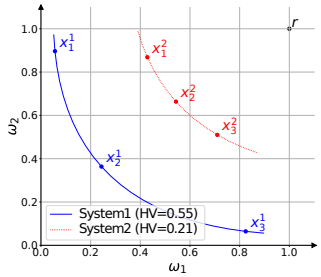


Figure 7: Convergence-Diversity: *System1* (blue) covers a larger volume of the solution space than *System2* (red), relative to the reference r (top-right). Thus, *System1* performs better in terms of *HV*.

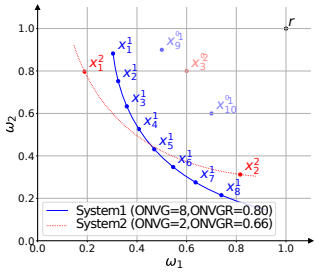


Figure 8: Capacity: *System1* (blue) has more capacity than *System2* (red) as it provides more non-dominated solutions, absolutely (*ONVG*), and relative to the total number of solutions per system (*ONVGR*).

non-dominated points compared to dominated ones is a good indicator of how well the objective is approximated. Having a larger number of non-dominated points may also improve tuning the trade-off system as the possibility of finding an expected combination of utility alongside fairness would increase due to more optimal solutions in the objective space. We apply a transformation for *ONVG* by normalizing over the maximum value of it for the systems as $\widehat{ONVG} = \frac{ONVG}{\max(ONVG)}$ to make it in the same range as others. On the other hand, *ONVGR* is in the range of $[0, 1]$ with 0 and 1 indicating the absence of the non-dominated and dominated solutions, respectively. However, these measurements fail to capture PF optimality as the number of solutions does not provide information about the Pareto characteristic. Fig. 8 highlights that *System1* exhibits a more capacity characteristic compared to *System2* in terms of *ONVG* and *ONVGR* as it has more non-dominated solutions, $ONVG_{System1} = 8 > ONVG_{System2} = 2$, and a bigger ratio on overall solutions, $ONVGR_{System1} = 0.80 > ONVGR_{System2} = 0.66$.

3.3 Radar Chart: Compact Visualization

The assessment of a utility-fairness trade-off system with the aforementioned performance indicators can be reported as a measurement table. However, it's also possible to convey this information in different ways such as the illustration in a chart summarizing all the performance indicators. A radar (spiderweb) chart is such a compact plot that compares different characteristics in the same projection and allows for easy comparative analysis of several systems over the same attributes.

The qualitative analysis resulting from the comparison of utility-fairness trade-offs with a radar chart makes it possible to select the optimal ML system showing more capacity, diversity, and convergence-diversity. This overall characteristic of the systems can also be quantified by calculating the areas occupied by each of them in the radar chart. The calculation of these areas allows for different systems to be comparable over very compact quantities compressed by the performance indicators in the table.

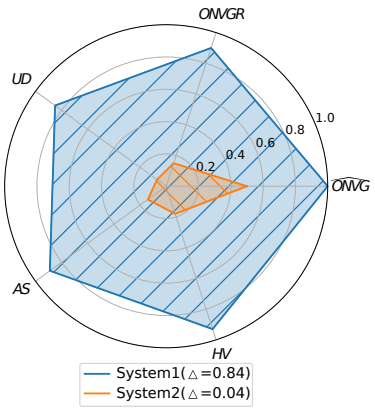
We consider the areas in the radar chart as polygons and use the Surveyor's formula (Braden, 1986) for calculation. Given a polygon *Poly* with n ordered points as counterclockwise in the Cartesian coordinate system, $v_i \in V$, we define $Poly = v_1, v_2, \ldots, v_n$ alongside $v_i = (x_i, y_i)$. As we work in the polar coordinate system to represent the systems in the radar chart, we define $v_i = (r_i, \Theta_i)$ where r_i is the radius and Θ_i is the angle. In this stage, we need to convert the point in polar coordinates to the counterpart in Cartesian one by $x_i = r_i \cos(\Theta_i)$ and $y_i = r_i \sin(\Theta_i)$. After switching to the Cartesian coordinate system, we apply the Surveyor's formula as shown below:

$$\Delta_{Poly} = \frac{1}{2}\left\{\begin{vmatrix} x_1 & x_2 \\ y_1 & y_2 \end{vmatrix} + \ldots + \begin{vmatrix} x_{n-1} & x_n \\ y_{n-1} & y_n \end{vmatrix} + \begin{vmatrix} x_n & x_1 \\ y_n & y_1 \end{vmatrix}\right\} \tag{12}$$

where $|.|$ is the 2×2 determinant. The calculation of *Area* (Δ) is then min-max normalized by $\widehat{\Delta} = \frac{\Delta - \min(\Delta)}{\max(\Delta) - \min(\Delta)}$ to transform the area range to $[0, 1]$. Given a pentagon with 5 dimensions as shown in Fig. 9a, the theoretical lower and upper bounds of the area are 0.00 and ≈ 2.37, respectively. This is analogous to the concept of the Area Under the Curve (AUC) over Receiver Operating Characteristic (ROC), where an area of 1.00 is expected to be the best situation for a system. In the illustrative example of Fig. 9a, it can be easily seen from the radar chart that *System1* outperforms *System2* in every dimension. This can be verified by their respective areas of 0.84 and 0.04 (Table 9b). We can also clearly observe that *System1* is closer to the ideal performance than *System2*, relying only on these areas.

3.4 Deduplication of Solutions

In our study, each point on the utility-fairness trade-off curve corresponds to an ML model generated under a



(a) Dominance of *System1* (blue) with respect to *System2* (orange) is clearly visible as it occupies a larger volume of the plot. Normalized MOO indicators are used to bind axes to the [0,1] scale, improving visual analysis.

	Convergence-Diversity	Capacity		Diversity		
				Distribution	Spread	
System	HV	$\widehat{ONVG}$	$ONVGR$	UD	AS	$\widehat{\Delta}$
$System1$	0.93	1.00	0.90	0.85	0.89	0.84
$System2$	0.18	0.50	0.15	0.07	0.14	0.04

(b) Quantitative results for (a) alongside the inner area of the radar chart.

Figure 9: A sample evaluation with the proposed radar chart. The radar chart provides a summarized assessment of the optimal utility-fairness trade-offs from the evaluated ML systems.

specific preference vector. These points form a solution set, S, in a multi-objective space, where the dimensions represent utility and fairness metrics. However, multiple preference vectors may cause to build ML models that exhibit nearly identical behavior, and the resulting trade-off system may have (near-)duplicate points in S. Such redundant ML models distort performance indicators by artificially increasing density or changing the characteristics of the approximated PF.

To alleviate this issue, we use a *deduplication operator*, denoted by $deduplicate_{\varepsilon}(.)$, which filters out ML models that lie within ε-neighborhood of each other given a utility-fairness trade-off system. We leverage the DBSCAN clustering algorithm (Schubert et al., 2017) to retain only the representative ML models by eliminating redundant ones. Given $S = s_1, s_2, \ldots, s_M \subset \mathcal{R}^N$ where M and N refer to the number of solutions and objectives, respectively, applying *deduplication operator*, $S' = deduplicate_{\varepsilon}(S)$, eliminates models in similar performance within ε formulated as below:

$$|s_i - s_j| > \varepsilon, \quad s_i, s_j \in S' \tag{13}$$

We empirically set $\varepsilon = 1e-6$ based on experiments conducted on synthetic and real-world datasets.

3.5 *A Priori* and *A Posteriori* Analysis

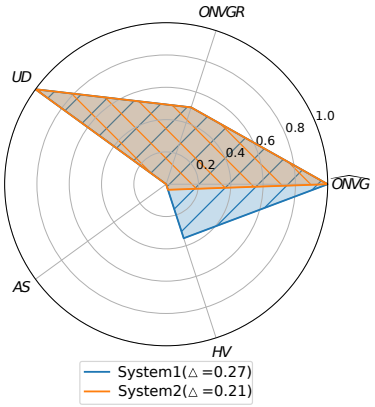
In ML evaluation, it is essential to distinguish between *a priori* and *a posteriori* analysis, as each serves a different role in assessing the generalization and stability of the systems in comparison. In the context of demographic fairness, we adapt this concept into our evaluation framework to achieve a similar analysis on utility-fairness trade-off systems compatible with ML assessment.

In our context, *a priori* analysis refers to identifying Pareto-optimal operating points by using a validation set, prior to observing the final test data. This procedure allows the selection of operating combinations of thresholds and sub-ML models (derived from the trade-off system) that constitute the estimated non-dominated solution set on the validation set. The goal of this analysis is to perform a real deployment scenario in which system parameters must be fixed before test-time evaluation. In contrast, *a posteriori* analysis relies directly on test-set evaluations and assesses the performance of all sub-ML models without any pre-selection. While *a posteriori* evaluations provide a complete characterization of the attainable utility-fairness spectrum, they do not represent a real generalization scenario, since the selection step is already considered with test data.

Our framework supports both protocols: *a priori* evaluation performs a realistic operating mode, whereas *a posteriori* evaluation provides a full diagnostic understanding of the system's trade-off characteristics. To have *a priori* evaluation in practice, we allow users to assess two subsets as validation and test by using the same trained ML model. Based on the validation set, the framework can determine the threshold/sub-ML combinations that define the Pareto-optimal estimate. These selections can then be directly applied to the test set to measure how well the trade-off structure generalizes. In Section 4.3, we provide use cases for both analyses based on real-world medical problems.

3.6 Simulations for Use Cases

The first use case, UC-1, focuses on black-box testing of *System1* and *System2*, corresponding to the first scenario (Fig. 1). As both systems have just 1 non-dominated solution, they have the same $\widehat{ONVG}$ and *ONVGR* values of 1.00 and 0.50 respectively. In terms of the diversity measurements, it is not possible to evaluate both systems as there exists only 1 non-dominated solution. *System1* has a higher *HV* score than *System2* because its non-dominated solution is farther from the *nadir* point than the non-dominated solution of the other, resulting in a larger volume. Table 10b and Fig. 10a show this comparison quantitatively and qualitatively. The radar chart in Fig. 10a illustrates the performance gap and *HV* dominance of *System1* over *System2*. The difference in the area is

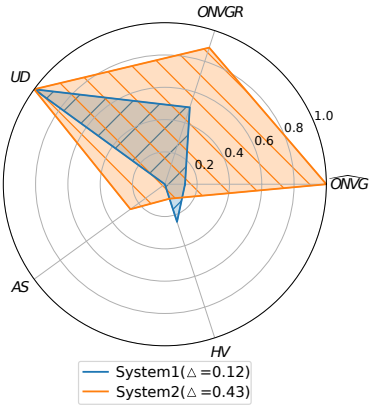


(a) Dominance of *System1* (blue) with respect to *System2* (orange) is clearly visible from the non-overlapped area.

	Convergence-Diversity	Capacity		Diversity		
				Distribution	Spread	
System	HV	$\widehat{ONVG}$	$ONVGR$	UD	AS	$\widehat{\Delta}$
System1	0.35	1.00	0.50	1.00	0.00	0.27
System2	0.04	1.00	0.50	1.00	0.00	0.21

(b) Quantitative results for (a) as black-box.

Figure 10: Simulation for UC-1.



(a) Dominance of *System2* (orange) with respect to *System1* (blue) is visible as it occupies a larger volume of the plot.

	Convergence-Diversity	Capacity		Diversity		
				Distribution	Spread	
System	HV	$\widehat{ONVG}$	$ONVGR$	UD	AS	$\widehat{\Delta}$
System1	0.24	0.12	0.50	1.00	0.00	0.12
System2	0.09	1.00	0.89	1.00	0.26	0.43

(b) Quantitative results for (a) as black- and white-box.

Figure 11: Simulation for UC-2.

$\widehat{\Delta} = \widehat{\Delta_1} - \widehat{\Delta_2} = 0.27 - 0.21 = 0.06$ and it arises solely from the *HV* differences between the two systems. In this case, the proposed evaluation framework simplifies the comparison using a single indicator, *HV*, as the other indicators do not play a role in the model selection decision.

In the second use case, UC-2, we perform hybrid testing with black- (*System1*, Fig. 1) and white-box (*System2*, Fig. 2) cases. 8 different non-dominated solutions are considered out of 25, adjusting τ for *System2* against a single non-dominated solution for *System1*, as it is not tunable. Table 11b contains the indicator scores for both systems. *System2* outperforms *System1* for $\widehat{ONVG}$ as it has more non-dominated solutions but results in a lower *HV* with a score of 0.09. The distribution indicator is not informative as *UD* is same for both systems, and the spread is better for *System2* with an *AS* of 0.26 compared to 0.00 for *System1*. This can be observed in the radar chart shown in Fig. 11a. We can end up with a decision that *System2* outperforms *System1* overall, as seen in the difference of the areas $\widehat{\Delta} = \widehat{\Delta_2} - \widehat{\Delta_1} = 0.43 - 0.12 = 0.31$.

The last use case, UC-3, considers white-box testing for both *System1* and *System2*, matching the second scenario (Fig. 2). 10 and 6 different non-dominated solutions were considered out of 25 for *System1* and *System2*, respectively. As seen in Table 12b, *System1* outperforms *System2* in terms of $\widehat{ONVG}$, *ONVGR*, *HV*, and *AS* with same performance in terms of *UD*. This can also be interpreted through a visual inspection of the radar chart in Fig. 12a. Finally, the areas of both systems confirm this as well, $\widehat{\Delta} = \widehat{\Delta_1} - \widehat{\Delta_2} = 0.61 - 0.31 = 0.30$.

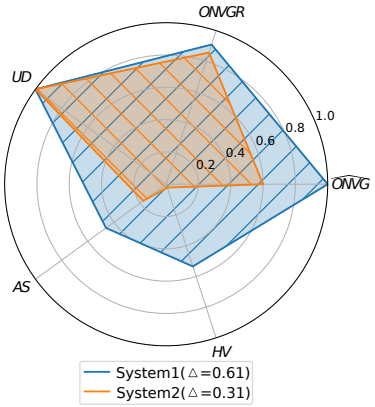
We can derive some conclusions about performance indicators based on the observations regarding the use cases above. Firstly, the black-box case only gives a partial characterization of the assessment since the ML systems are not tunable. In this case, we obtain non-discriminative values for $\widehat{ONVG}$, *ONVGR*, *UD*, and *AS* as seen in Fig. 10. Secondly, diversity is another issue when working with a low number of solutions. A system with only 1 solution does not allow to evaluate *AS* as there must be at least two solutions to measure the distance between the extreme points from the system and PF. A single point solution generates an *AS* score of 0.00 as seen in *System1* shown in Table 11b. Thirdly, *HV* may be the first decision point to select the system, which exhibits more PF characteristics, as it covers every aspect of convergence, diversity and capacity. There may be opposite cases between *HV* and the other indicators as seen in Fig. 11. A good indication of $\widehat{ONVG}$, *ONVGR*, *UD*, and *AS* does not work as well as evaluating the performance over *HV*, which has been proven to be a more reliable measurement for PF (Audet et al., 2021; Jiang et al., 2014).

4. Empirical Validation

In this section, we demonstrate the effectiveness of the proposed evaluation framework empirically using three fairness-aware medical imaging datasets.

4.1 Dataset Description

To empirically validate the proposed evaluation framework, we use three medical imaging datasets with demographic attributes: the Harvard Glaucoma Fairness (HGF)



(a) Dominance of *System1* (blue) with respect to *System2* (orange) is clearly visible from the non-overlapped area as it occupies a larger volume of the plot.

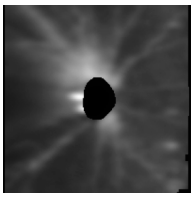
	Convergence-Diversity	Capacity		Diversity		
				Distribution	Spread	
System	HV	$\widehat{ONVG}$	$ONVGR$	UD	AS	$\widehat{\Delta}$
System1	0.54	1.00	0.91	1.00	0.46	0.61
System2	0.02	0.60	0.86	1.00	0.18	0.31

(b) Quantitative results for (a) as white-box.

Figure 12: Simulation for UC-3.

dataset (Luo et al., 2024), the Shenzhen Chest X-ray dataset (Jaeger et al., 2014), and the mBRSET retinal dataset (Wu et al., 2025). These datasets were selected because they jointly capture demographic and clinical diversity, and enable the evaluation of utility-fairness trade-offs under both modality-based and population-based conditions.

Harvard Glaucoma Fairness (HGF). The HGF dataset includes cases with a retinal nerve disease called glaucoma as well as samples of healthy patients. Glaucoma is twice as common in Black patients compared to other races, and more prevalent in men (Khachatryan et al., 2019; Luo et al., 2024). The dataset comprises 3300 two-dimensional retinal nerve fiber layer thickness (RNFLT) maps from three racial groups, namely Asian, Black, and White, with a resolution of 200×200 pixels. Color intensity represents retinal nerve fiber thickness in micrometers around the optic disc. The glaucoma/non-glaucoma ratio in the dataset is $53.0\%/47.0\%$. The prevalence of glaucoma in Asians, Blacks, and Whites in the HGF dataset is equally balanced at 33.3%, and the gender distribution is as 54.9% and 45.1% for females and males, respectively. The ophthalmologic images from HGF are linked to sensitive attributes for race, gender, age, and ethnicity. This dataset provides a clear setting for assessing demographic disparities, as the higher glaucoma prevalence among Black patients contradicts the relative scarcity of such samples in ophthalmology data. This fact creates a fairness challenge that represents real-world imbalance in healthcare AI systems and provides a solid use case for our framework.


(a) HGF RNFLT map.


(b) Shenzhen X-ray.


(c) mBRSET fundus.

Figure 13: Sample images from the datasets used in the empirical study: (a) RNFLT map from HGF for glaucoma detection; (b) chest X-ray from Shenzhen for pulmonary disease screening; and (c) retinal fundus image from mBRSET for diabetic retinopathy classification.

Shenzhen Chest X-ray Dataset. The Shenzhen dataset is a public chest X-ray collection designed for computer-aided screening of pulmonary diseases, particularly tuberculosis (TB). The dataset contains 662 frontal chest X-rays with a distribution of normal and TB-positive cases as 326 and 336, respectively. Image resolutions vary with an approximation of 3000×3000 pixels. The dataset has sensitive attributes for age and gender, providing a complementary use case for evaluating fairness across clinical subgroups with chest radiographs.

mBRSET Retinal Dataset. The mBRSET dataset is a retinal imaging resource and designed to enable the benchmarking of automated ophthalmologic screening models under different demographic conditions. The dataset consists of 5164 retinal images from 1291 subjects, including fundus photography modality. Each sample is annotated for ocular diseases such as diabetic retinopathy and macular edema, alongside demographic attributes such as patient age, gender, and obesity. The image resolution varies in height and width, ranging from 874 and 951 to 2304 to 2984 pixels, respectively. The dataset exhibits a gender imbalance, with female patients compared to male ones as $65.1\%/34.9\%$. The patient age has an average of 61.4 years with a standard deviation of 11.6.

By jointly employing these three datasets, our empirical evaluation covers complementary fairness perspectives across medical imaging. HGF provides a clear setting for analyzing demographic and disease prevalence imbalances in glaucoma with race and gender labels. The Shenzhen dataset extends this perspective by providing chest X-ray imagery annotated with sensitive attributes such as gender, enabling fairness assessment in a different clinical modality. Finally, mBRSET provides clinical diversity across ocular conditions by having a broader diagnostic context with different retinal diseases. Together, these datasets enable a more comprehensive examination of how the proposed framework evaluates utility-fairness trade-offs

across different medical imaging modalities, disease types, and demographic distributions.

4.2 System Definitions and Configuration

To properly demonstrate the capabilities of the proposed evaluation framework, we formulate the utility-fairness trade-off as a bi-objective optimization problem through two distinct systems denoted as *System1* and *System2*, which differ from the systems considered in previous experiments. For the HGF dataset, both systems are developed using Pareto HyperNetworks (PHNs), which are based on hypernetworks (Ha et al., 2016), and are designed to learn Pareto-optimal solutions across multiple objectives (Navon et al., 2020). The PHN formulation allows the generation of sub-neural networks (sub-NNs) that represent different levels of compromise between diagnostic utility and fairness, enabling continuous exploration along the PF. In particular, each sub-NN corresponds to a specific utility-fairness trade-off level, *i.e.*, one sub-model represents a diagnostic performance with higher utility but lower fairness, whereas another has higher equity at the cost of classification accuracy. Each objective combination is represented by a preference vector that encodes the relative weighting between the two objectives. In the HGF setting, all sub-NNs are evaluated directly on the test set, and the full spectrum of utility-fairness outcomes is characterized as *a posteriori*, without a validation-driven selection of operating points.

Each PHN employs a ResNet-18 backbone as a shared encoder that generates parameters for sub-NNs corresponding to different preference vectors. A total of 25 preference vectors are uniformly sampled from a Dirichlet distribution with $\alpha = 0.2$ to represent distinct utility-fairness trade-offs. The utility objective is defined by Binary Cross-Entropy (BCE) loss, and the fairness objective is modeled using a differentiable relaxation over Equalized Odds (EO) criterion, which captures disparities in true positive and false positive rates (TPR, FPR) across demographic subgroups. The combined objective is jointly optimized through backpropagation to dynamically generate Pareto-optimal solutions that balance diagnostic utility and fairness.

For the mBRSET and Shenzhen datasets, *System1* and *System2* use two different model architectures. *System1* is based on the DenseNet topology (Huang et al., 2017), and this setting follows a training mechanism performed directly under a bi-objective loss combining BCE for utility and EO-based fairness penalties to represent an independent ML model that corresponds to a specific preference. Unlike the joint optimization characteristic of the PHN, here, each utility-fairness trade-off is modeled as a separate ML model. *System2* employs a LoRA-enabled ViT-Small model (Hu et al., 2022; Dosovitskiy, 2020), where low-rank adaptation layers are inserted into the self-attention blocks to enable lightweight fairness-aware fine-tuning. For both datasets, the hyperparameters (*e.g.*, learning rate, optimizer, and batch size) follow a fixed configuration within each system. For the mBRSET, PF operating points are selected on a validation set and then evaluated on the test set, following *a priori* analysis of the utility-fairness trade-offs, whereas for Shenzhen, all operating points are evaluated directly on the validation set, corresponding to an *a posteriori* analysis.

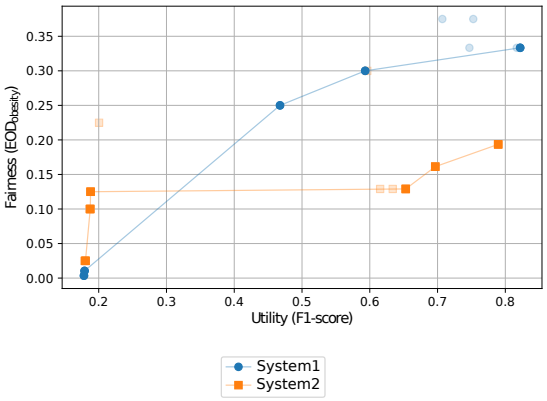
System1 serves as a baseline fairness configuration across all datasets. It focuses on optimizing fairness with respect to a *binary sensitive attribute* in each case. For the HGF dataset, this system minimizes EO disparities between *male* and *female* patients, targeting *gender fairness*. For the mBRSET, the system similarly minimizes EO disparities between obese and non-obese patients, capturing *obesity-related fairness*. For the Shenzhen dataset, the system again focuses on *gender fairness* by minimizing EO disparities between same subgroups.

System2 extends this baseline to investigate dataset specific demographic variation. In the HGF setting, this system shifts the fairness objective from gender to race by minimizing EO disparities between *Asian*, *Black*, and *White* subgroups. For the mBRSET dataset, the LoRA-enabled ViT-Small model uses the same EO-based *obesity fairness* objective. Similarly, for the Shenzhen dataset, the same topology minimizes EO disparities between *male* and *female* patients.

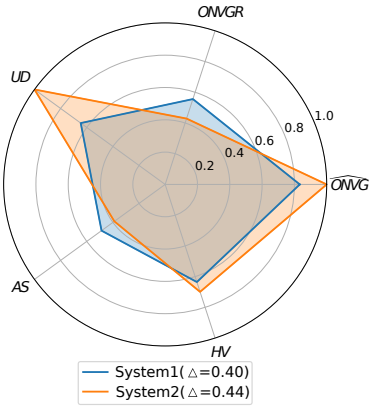
All systems within the same dataset share identical training procedures and hyperparameters to ensure comparability. For PHN-based systems, optimization is performed using Adam optimizer with a learning rate of $1e-4$ and batch size of 256. DenseNet- and LoRA-ViT-Small-based systems employ dataset-specific but fixed hyperparameters across *System1* and *System2* using Adam optimizer with a batch size of 16. However, learning rates differ as $5e-5$ and $1e-4$ for the mBRSET and Shenzhen datasets, respectively. This shared setup ensures that differences in observed performance arise solely from the fairness objective being optimized to provide a controlled analysis of utility-fairness trade-offs.

4.3 Evaluation Results

The comparative evaluation of the DenseNet- and LoRA-ViT-Small-based systems on the mBRSET dataset is illustrated in Fig. 14, which reports the distribution of utility (F1-score) and fairness outcomes (equalized odds difference (EOD) for the obesity attribute) across all ML models in the trade-off system. We note that group-wise comparisons based on prevalence-sensitive metrics such as F1-score should be interpreted with caution, as they are influenced by the underlying base rate (class ratio) and



(a) Pareto plot.

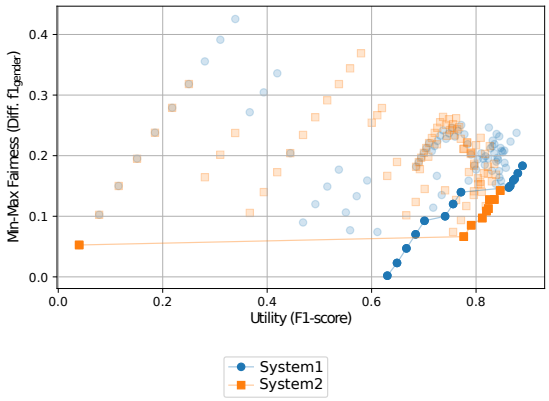


(b) Radar chart.

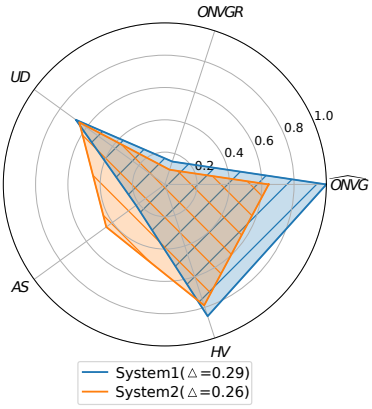
	Convergence-Diversity	Capacity		Diversity		
				Distribution	Spread	
System	HV	$\widehat{ONVG}$	$ONVGR$	UD	AS	$\widehat{\Delta}$
System1	0.64	0.83	0.56	0.65	0.49	0.40
System2	0.70	1.00	0.43	1.00	0.39	0.44

(c) Quantitative results.

Figure 14: DenseNet/LoRA-ViT-Small on mBRSET.



(a) Pareto plot.



(b) Radar chart.

	Convergence-Diversity	Capacity		Diversity		
				Distribution	Spread	
System	HV	$\widehat{ONVG}$	$ONVGR$	UD	AS	$\widehat{\Delta}$
System1	0.86	1.00	0.15	0.68	0.22	0.29
System2	0.79	0.64	0.10	0.65	0.45	0.26

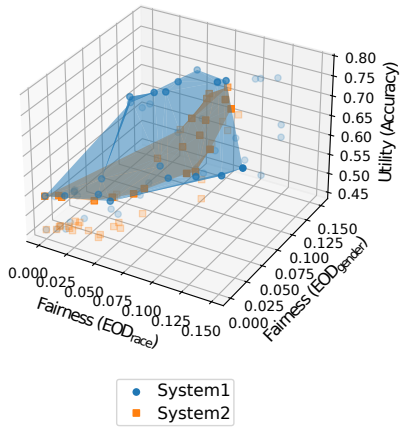
(c) Quantitative results.

Figure 15: DenseNet/LoRA-ViT-Small on Shenzhen.

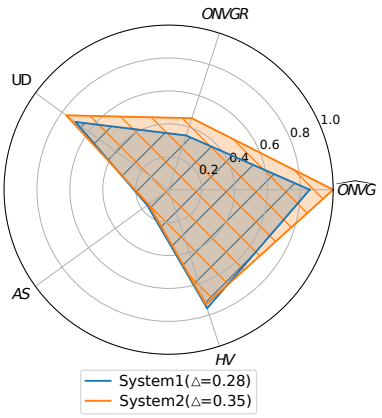
may partially reflect distributional characteristics rather than purely model-induced disparities. The curves show that *System2* has lower EOD scores for a broader range of utility levels, indicating improved fairness consistency under varying preference settings. Table 14c further summarizes the evaluation for PF. *System2* achieves higher scores in both $\widehat{ONVG}$ and *UD*. The *HV* metric also favors *System2* (0.70 vs. 0.64), reflecting a closer alignment with the *ideal* reference point. In contrast, *System1* exhibits a higher *ONVGR* and *AS* values. When aggregated using the area score $\widehat{\Delta}$, *System2* achieves 0.44 compared to 0.40 for *System1*, confirming that *System2* exhibits a higher overall trade-off performance structure when balancing diagnostic utility and *obesity fairness* on this dataset.

The performance of the DenseNet- and LoRA-ViT-Small-based systems on Shenzhen dataset is summarized in Fig. 15 that visualizes the utility performance (F1-score) and fairness disparity (min-max difference for the gender groups). As shown in the plots, *System1* demonstrates a more favorable balance between fairness variability and utility consistency. Specifically, *System1* maintains a higher performance on the F1-score without hitting very low scores while achieving comparable EOD values against *System2*. A quantitative comparison using PF-based indicators is provided in Table 15c. *System1* achieves higher performance in $\widehat{ONVG}$ with *ONVGR*, and outperforms *System2* in *UD* and *HV* (0.86 vs. 0.79). While *System2* exhibits a larger *AS* indicating a broader spread of trade-off solutions, its distribution characteristic is weaker. Aggregating all indicators through the area score shows *System1* achieving 0.29 compared to 0.26 for *System2*, validating that *System1* provides a higher trade-off structure for *gender fairness* in the Shenzhen dataset.

For the HGF experiment, Fig. 16 summarizes the evaluation of the two PHN-based systems considering one utility performance (accuracy) and two fairness metrics



(a) 3D plot in metric space of ML system performance for one utility metric (accuracy) and two fairness criteria (gender and race).



(b) *System2* (orange) slightly dominates *System1* (blue) with a margin.

	Convergence-Diversity	Capacity		Diversity			Model 1				Model 10				Model 25			
				Distribution	Spread		Utility		Fairness		Utility		Fairness		Utility		Fairness	
System	HV	$\widehat{ONVG}$	$ONVGR$	UD	AS	$\widehat{\Delta}$	Acc	$F1$	DP	$EOdd$	Acc	$F1$	DP	$EOdd$	Acc	$F1$	DP	$EOdd$
System1	0.76	0.86	0.35	0.70	0.16	0.28	0.78	0.77	0.07	0.09	0.68	0.75	0.06	0.08	0.54	0.70	0.00	0.00
System2	0.73	1.00	0.46	0.77	0.15	0.35	0.76	0.76	0.08	0.10	0.56	0.71	0.02	0.05	0.54	0.70	0.00	0.00

(c) Quantitative results.

Figure 16: PHN on HGF.

(EOD for the gender and race). Visual inspection alone illustrated by the 3D performance plot in Fig. 16a is insufficient to reliably differentiate the systems when multiple fairness criteria are considered alongside utility. At this point, the radar visualization in Fig. 16b provides a clearer qualitative aspect by showing a slight performance improvement for *System2* over *System1*. This observation is further confirmed by the quantitative indicators reported in Table 16c. Although there is an inconsistency between *HV* and the other performance indicators ($\widehat{ONVG}$, *ONVGR* and *UD*), an area of 0.35 over 0.28 shows that *System2* constitutes the better-performing utility-fairness trade-offs than *System1*. In this case, it is not possible to differentiate the systems over *AS* as they both exhibit average spreads of 0.15 and 0.16. Finally, it is straightforward to interpret that both systems are far from optimality as they are well below the ideal performance of $\widehat{\Delta} = 1.00$.

Table 16c also reports the performance of individual sub-NNs generated by the PHN (Models 1, 10, and 25), providing concrete examples of model-level behavior within each utility-fairness trade-off system. These examples illustrate how operating points guided by preference along the PF can differ substantially in their balance between utility and fairness. For instance, in *System1* (*gender fairness*), the first sub-model achieves high utility (accuracy $(Acc) = 0.78$ and F1-score $(F1) = 0.77$), but exhibits demographic disparity (demographic parity difference $(DP) = 0.07$ and equalized odds difference $(EOdd) = 0.09$), whereas later models, such as 25th sub-NN, perform substantially reduced unfairness ($DP = 0.00$ and $EOdd = 0.00$) at the cost of lower accuracy and F1-score ($Acc = 0.54$ and $F1 = 0.70$). A similar trend is observed for *System2* (*race fairness*), where earlier preference vectors favor utility while later ones build more equitable outcomes across racial groups. However, evaluating these systems solely through individual model outputs is insufficient, as meaningful assessment requires understanding their behavior under multiple fairness criteria. Although *System1* and *System2* are trained and evaluated on gender and race fairness, respectively, a complete analysis also demands examining cross-criterion disparities to capture the broader fairness landscape. These examples highlight how the proposed framework supports aggregated, system-level interpretation rather than relying on isolated operating points. Ultimately, this reinforces that fairness-aware evaluation must adopt a multi-objective perspective that jointly reflects utility and multiple fairness criteria.

5. Discussion

This work presents the requirements and advantages of an evaluation framework assessing multidimensional utility-fairness trade-offs obtained with ML systems. The framework enhances the comparison of different modeling strategies, with the goal of selecting optimal solutions for real-

world applications that require the assessment of multiple fairness criteria. The ease of use and effectiveness of the proposed framework are explored through comprehensive simulations and empirical studies using medical imaging datasets designed to test fairness optimization approaches. Unlike previous evaluation frameworks, this work provides a series of steps for the selection process of ML systems in the context of multidimensional fairness exploring different criteria, supported by MOO principles. Characterizing the optimal PF is particularly useful in tasks where contradictory fairness performance indicators cannot be avoided, and trade-offs should be tuned to specific fairness requirements from decision-makers. Our framework could guide the tuning of ML systems to different notions of fairness for any sensitive attribute and metric in a simple and transparent manner. Fairness is interpreted as a multidimensional evaluation space, and the spectrum of trade-offs is captured across demographic attributes rather than enforcing a single optimal fairness configuration, which may fail to represent the overall capability of a fairness-aware ML system.

While the utility-fairness trade-off is commonly reported in fairness-aware ML studies, it does not universally manifest across all settings. As discussed in previous works (Chouldechova, 2017; Kleinberg et al., 2016), different fairness definitions such as *independence*, *separation*, and *sufficiency* are mutually incompatible under certain statistical assumptions, which means that optimizing multiple fairness notions simultaneously is impossible to achieve. Furthermore, recent works also show that the perceived utility-fairness tension may not always exist when dataset-level biases are addressed (Wick et al., 2019; Dutta et al., 2020). Conversely, a more recent work by Dehdashtian et al. (2024) suggests that utility-fairness trade-offs can indeed arise from intrinsic dataset characteristics by empirically observing the persistence of such tension. Taken together, these findings indicate that utility-fairness interactions are highly context dependent and influenced by both data distribution and model assumptions. In this ongoing research landscape, our model- and metric-agnostic evaluation framework contributes by providing a structured way to *capture a snapshot* of how different methods achieve the balance/compensation between utility and fairness. By quantifying these dynamics in a unified evaluation space, the framework allows for a systematic comparison of fairness-aware ML systems independent of their underlying architectures or optimization strategies, and it provides a baseline for future work that needs a benchmark for multi-objective fairness evaluation. Moreover, the framework is designed flexibly for practitioners so that they can select and evaluate the specific utility-fairness trade-offs relevant to their application, rather than imposing any pre-determined balance among objectives or assuming a universal trade-off structure.

As mentioned in Section 2, the comprehensive analysis of PFs in multiple dimensions is a limitation of previously proposed fairness evaluation approaches. A summarized objective representation of performance indicators from the PFs, both as a radar chart and as a measurement table, overcomes the limitations from visualizing performance plots in multiple dimensions resulting from the assessment of different ML systems. Furthermore, analogous to AUC over ROC, our evaluation framework can transform the qualitative trend into a quantitative measurement, thus gathering all necessary information for the interpretation of the performance gap between the trade-off systems in consideration and against an ideal solution.

Beyond the medical imaging domain, the proposed framework has broader implications for fairness evaluation across ML applications, as our approach generalizes the evaluation process by integrating multiple fairness and utility criteria within a single multi-objective formulation, independent of the domain under consideration. This generalization enables the framework to serve as a common evaluation protocol for comparing utility-fairness trade-offs in different contexts, such as financial risk assessment and biometric systems. In this regard, our results complement existing fairness benchmarking protocols by providing a structured, quantitative, and qualitative means of summarizing utility-fairness trade-off performance across domains.

There may be some limitations of the proposed evaluation framework. A restriction is the exponential cost for computing the MOO-based performance indicators as the number of objectives increases (Audet et al., 2021). This scalability challenge is particularly relevant when fairness must be evaluated across many demographic axes or multiple fairness notions simultaneously, which may arise in complex real-world deployments. However, we argue that the majority of tasks studied in the literature focus on a small set of sensitive attributes, *i.e.*, gender, age, and race. The use of alternative performance indicators, such as Inverted Generational Distance (*IGD*) (Coello Coello and Reyes Sierra, 2004), could be explored to improve efficiency. Another issue to consider is the equal weighting of all performance indicators, which normalizes the contribution of the different indicators in the final evaluation. This may not be desirable in situations where an indicator can evaluate the systems in suboptimal performance and needs to have less impact on the final decision compared to others. The proposed evaluation framework could be extended to support the dynamic weighting of the indicators, so that the re-weighting can be performed based on the use case. As another limitation, there may also be situations in which the indicator measurements are the same or not feasible (refer to *AS* and *UD* in Fig. 10) for all systems in comparison. In this case,

such an indicator is not informative for the final decision and the proposed framework simplifies the evaluation to the joint contribution of the rest. This issue could be alleviated by using different indicators that discriminate more, as the proposed framework supports seamlessly including/excluding different types of indicators. Finally, as emphasized by Selbst et al. (2019), algorithmic modeling and evaluation of fairness cannot be fully abstracted from its social context. Accordingly, the proposed evaluation framework should be interpreted as an assessment tool for examining utility-fairness trade-offs under clearly defined assumptions and constraints, rather than as a universal fairness solution applicable to all real-world scenarios.

6. Conclusions

This paper proposes a multi-objective evaluation framework for utility-fairness trade-offs resulting from ML systems, using performance indicators based on MOO. The proposed framework is model- and task-agnostic, allowing for high flexibility in the comparison of ML strategies, even when they have been optimized for different objectives. This is an adaptive assessment framework that supports any kind of performance indicators, including the proposed method, for convergence, diversity and capacity analysis. The proposed framework is able to perform a comprehensive analysis of ML systems with a measurement table and radar chart, overcoming the limitations resulting from the qualitative assessment of solutions with multiple fairness requirements. These tools provide a structured and visual means to evaluate and compare multiple fairness metrics in ML systems. The measurement table allows for a clear, organized presentation of the data, while the radar chart offers a visual representation of how well the system performs across various utility and fairness criteria using MOO indicators. The effectiveness of the evaluation approach is verified by performing simulations and empirical analyses for a variety of use cases, with both black- and white-box ML systems. In particular, the empirical results on medical imaging based use cases illustrate how the framework can expose fairness disparities between diagnostic models, guiding the selection of appropriate trade-offs for ML systems in healthcare applications. The proposed system is made available for public access to be applied in the context of multi-objective evaluation for any domain.

Acknowledgments

This work was supported by the Swiss National Science Foundation (SNSF) through the project FairMI - Machine Learning Fairness with Application to Medical Images under grant number 214653. We also thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), grants 21/14725-3 and 2023/12468-9.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

Data availability

All datasets used in this study are publicly available:
Harvard Glaucoma Fairness (HGF): `https://github.com/Harvard-Ophthalmology-AI-Lab/Harvard-GF`;
Shenzhen Chest X-ray: `https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/Shenzhen-Hospital-CXR-Set/index.html`;
mBRSET: `https://physionet.org/content/mbrset/1.0/`.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

Nahida Akter, John Fletcher, Stuart Perry, Matthew P Simunovic, Nancy Briggs, and Maitreyee Roy. Glaucoma diagnosis using multi-feature analysis and a deep learning technique. *Scientific Reports*, 12(1):8064, 2022.

Charles Audet, Jean Bigeon, Dominique Cartier, Sébastien Le Digabel, and Ludovic Salomon. Performance indicators in multiobjective optimization. *European journal of operational research*, 292(2):397–422, 2021.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Bart Braden. The surveyor's area formula. *The College Mathematics Journal*, 17(4):326–337, 1986.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific reports*, 12 (1):4209, 2022.

Eunice Chan, Zhining Liu, Ruizhong Qiu, Yuheng Zhang, Ross Maciejewski, and Hanghang Tong. Group fairness via group consensus. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1788–1808, 2024.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Carlos A Coello Coello and Margarita Reyes Sierra. A study of the parallelization of a coevolutionary multi-objective evolutionary algorithm. In *Mexican international conference on artificial intelligence*, pages 688–697. Springer, 2004.

Sepehr Dehdashtian, Bashir Sadeghi, and Vishnu Naresh Boddeti. Utility-fairness trade-offs and how to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12037–12046, 2024.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Raman Dutt, Ondrej Bohdal, Sotirios A Tsaftaris, and Timothy Hospedales. Fairtune: Optimizing parameter efficient fine tuning for fairness in medical image analysis. *arXiv preprint arXiv:2310.05055*, 2023.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*, pages 2803–2813. PMLR, 2020.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Sean P Garin, Vishwa S Parekh, Jeremias Sulam, and Paul H Yi. Medical imaging data science competitions should report dataset demographics and evaluate for bias. *Nature medicine*, 29(5):1038–1039, 2023.

Hao Gong and Chunxiang Guo. Influence maximization considering fairness: A multi-objective optimization approach with prior knowledge. *Expert Systems with Applications*, 214:119138, 2023.

Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. Facet: Fairness in computer vision evaluation benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20370–20382, 2023.

David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

Taeuk Jang and Xiaoqian Wang. Difficulty-based sampling for debiased contrastive representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24039–24048, 2023.

Taeuk Jang, Pengyi Shi, and Xiaoqian Wang. Group-aware threshold adaptation for fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6988–6995, 2022.

Siwei Jiang, Yew-Soon Ong, Jie Zhang, and Liang Feng. Consistencies and contradictions of performance metrics in multiobjective optimization. *IEEE transactions on cybernetics*, 44(12):2391–2404, 2014.

Nathanael Jo, Sina Aghaei, Jack Benson, Andres Gomez, and Phebe Vayanos. Learning optimal fair decision

trees: Trade-offs between interpretability, fairness, and accuracy. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 181–192, 2023.

Nikola Jovanović, Mislav Balunovic, Dimitar Iliev Dimitrov, and Martin Vechev. Fare: Provably fair representation learning with practical certificates. In *International Conference on Machine Learning*, pages 15401–15420. PMLR, 2023.

Naira Khachatryan, Maxwell Pistilli, Maureen G Maguire, Rebecca J Salowe, Raymond M Fertig, Tanisha Moore, Harini V Gudiseva, Venkata RM Chavali, David W Collins, Ebenezer Daniel, et al. Primary open-angle african american glaucoma genetics (poaagg) study: gender and risk of poag in african americans. *PloS one*, 14(8):e0218804, 2019.

Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. Fact: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning*, pages 5264–5274. PMLR, 2020.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Alejandro Kuratomi, Zed Lee, Panayiotis Tsaparas, Evaggelia Pitoura, Tony Lindgren, Guilherme Dinis Junior, and Panagiotis Papapetrou. Subgroup fairness based on shared counterfactuals. *Knowledge and Information Systems*, pages 1–39, 2025.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Camille Little. To the fairness frontier and beyond: Identifying, quantifying, and optimizing the fairness-accuracy pareto frontier. Master's thesis, Rice University, 2023.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

Suyun Liu and Luis Nunes Vicente. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science*, 19(3):513–537, 2022.

Zhichao Lu, Ian Whalen, Yashesh Dhebar, Kalyanmoy Deb, Erik D Goodman, Wolfgang Banzhaf, and Vishnu Naresh Boddeti. Multiobjective evolutionary design of deep convolutional neural networks for image classification. *IEEE Transactions on Evolutionary Computation*, 25(2): 277–291, 2020.

Yan Luo, Yu Tian, Min Shi, Louis R Pasquale, Lucy Q Shen, Nazlee Zebardast, Tobias Elze, and Mengyu Wang. Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization. *IEEE Transactions on Medical Imaging*, 43(7):2623–2633, 2024.

Aviv Navon, Aviv Shamsian, Gal Chechik, and Ethan Fetaya. Learning the pareto front with hypernetworks. *arXiv preprint arXiv:2010.04104*, 2020.

Kirtan Padh, Diego Antognini, Emma Lejal-Glaude, Boi Faltings, and Claudiu Musat. Addressing fairness in classification with a model-agnostic multi-objective algorithm. In *Uncertainty in artificial intelligence*, pages 600–609. PMLR, 2021.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55 (3):1–44, 2022.

Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34:25944–25955, 2021.

Ricardo Trainotti Rabonato and Lilian Berton. A systematic review of fairness in machine learning. *AI and Ethics*, 5(3):1943–1954, 2025.

Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2586–2594, 2019.

Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.

Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. Fairness perceptions of algorithmic

decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2):20539517221115189, 2022.

Kay Chen Tan, Tong Heng Lee, and Eik Fun Khor. Evolutionary algorithms for multi-objective optimization: Performance assessments and comparisons. *Artificial intelligence review*, 17(4):251–290, 2002.

David A Van Veldhuizen and Gary B Lamont. On measuring multiobjective evolutionary algorithm performance. In *Proceedings of the 2000 congress on evolutionary computation. CEC00 (Cat. No. 00TH8512)*, volume 1, pages 204–211. IEEE, 2000.

Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H Chi. Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1748–1757, 2021.

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.

Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems. *Journal of Machine Learning Research*, 24(257):1–8, 2023.

Susan Wei and Marc Niethammer. The fairness-accuracy pareto front. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(3):287–302, 2022.

Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32, 2019.

Chenwei Wu, David Restrepo, Luis Filipe Nakayama, Lucas Zago Ribeiro, Zitao Shuai, Nathan Santos Barboza, Maria Luiza Vieira Sousa, Raul Dias Fitterman, Alexandre Durao Alves Pereira, Caio Vinicius Saito Regatieri, et al. A portable retina fundus photos dataset for clinical, demographic, and diabetic retinopathy prediction. *Scientific Data*, 12(1):323, 2025.

Jin Wu and Shapour Azarm. Metrics for quality assessment of a multiobjective design optimization solution set. *J. Mech. Des.*, 123(1):18–25, 2001.

Violet Xinying Chen and John N Hooker. A guide to formulating fairness in an optimization model. *Annals of Operations Research*, 326(1):581–619, 2023.

Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, 30(10):2838–2848, 2024.

Zhenhuan Yang, Yan Lok Ko, Kush R Varshney, and Yiming Ying. Minimax auc fairness: Efficient algorithm with provable convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11909–11917, 2023.

Qingquan Zhang, Jialin Liu, Zeqi Zhang, Junyi Wen, Bifei Mao, and Xin Yao. Fairer machine learning through multi-objective evolutionary learning. In *International conference on artificial neural networks*, pages 111–123. Springer, 2021.

Yixuan Zhang, Boyu Li, Zenan Ling, and Feng Zhou. Mitigating label bias in machine learning: Fairness through confident learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16917–16925, 2024.

Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research*, 23(57):1–26, 2022.

Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10421, 2022.

Eckart Zitzler and Lothar Thiele. Multiobjective optimization using evolutionary algorithms—a comparative case study. In *International conference on parallel problem solving from nature*, pages 292–301. Springer, 1998.

Eckart Zitzler, Lothar Thiele, Marco Laumanns, Carlos M Fonseca, and Viviane Grunert Da Fonseca. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on evolutionary computation*, 7(2):117–132, 2003.