FairCognizer: A Model for Accurate Predictions with Inherent Fairness Evaluation (Extended Abstract)*

Adda-Akram Bendoukha[1], **Nesrine Kaaniche**[1], **Aymen Boudguiga**[2] and **Renaud Sirdey**[2]
[1]Samovar, Télécom SudParis, Institut Polytechnique de Paris, France
[2]Université Paris Saclay, CEA List, France
adda-akram.bendoukha@telecom-sudparis.eu, kaaniche.nesrine@telecom-sudparis.eu
aymen.boudguiga@cea.fr, renaud.sirdey@cea.fr

Abstract

Algorithmic fairness is a critical challenge in building trustworthy Machine Learning (ML) models. ML classifiers strive to make predictions that closely match real-world observations (ground truth). However, if the ground truth data itself reflects biases against certain sub-populations, a dilemma arises: prioritize fairness and potentially reduce accuracy, or emphasize accuracy at the expense of fairness. This work proposes a novel training framework that goes beyond achieving high accuracy. Our framework trains a classifier to not only deliver optimal predictions but also to identify potential fairness risks associated with each prediction. To do so, we specify a dual-labeling strategy where the second label contains a per-prediction fairness evaluation, referred to as an unfairness risk evaluation. In addition, we identify a subset of samples as highly vulnerable to group-unfair classifiers. Our experiments demonstrate that our classifiers attain optimal accuracy levels on both the Adult-Census-Income and Compas-Recidivism datasets. Moreover, they identify unfair predictions with nearly 75% accuracy at the cost of expanding the size of the classifier by 45%.

1 Introduction

Machine learning (ML) systems are increasingly pervasive, playing a crucial role in diverse applications like predictive maintenance, autonomous driving, and extending into sensitive domains such as judicial and medical fields [Dressel and Farid2018, Chen *et al.*2019, Cohen *et al.*2020, Ghassemi *et al.*2014]. This broad impact underscores the importance of fair predictions, especially given the biases in historical data [Falletti2023, Mittelstadt *et al.*2016, Fabris *et al.*2022].

Examples like biased skin condition diagnostic tools leading to misdiagnoses for minorities [Seyyed-Kalantari *et al.*2021, Wen *et al.*2022] and unfair recidivism risk prediction algorithms [Dressel and Farid2018] highlight the deep implications of this issue. Regulatory efforts like the EU AI Act aim to address these concerns.

Several works [Chouldechova2016, Kamiran and Calders2011, Feldman *et al.*2015, Fish *et al.*2016, Zafar *et al.*2017, Bendoukha *et al.*2025] show that ensuring fairness in supervised learning is often framed as a trade-off between considering a fair representation or an accurate one, in terms of proximity to ground-truth observations. An accurate classifier learns from historical records, generalizing observed statistical patterns to unseen data. However, if these patterns involve many discriminatory records, the classifier will adopt this biased behaviour. Achieving fair training often requires learning an alternative representation of data, generated via pre-processing techniques to remove biases [Kamiran and Calders2011, Chouldechova2016, He *et al.*2019, Xu *et al.*2018]. This alternative representation does not perfectly mirror reality. Consequently, this distributional drift will inevitably degrade the utility of a classifier trained on this alternative fair representation.

This work introduces FairCognizer [Bendoukha *et al.*2024], a dual-objective classifier learning both accurate and fair representations, providing an unfairness risk assessment for each prediction. Our contributions are threefold: (1) the FairCognizer framework for learning dual labels, (2) a novel sample-level unfairness risk measure to identify "vulnerable" records, and (3) experimental validation on Adult and Compas demonstrating maintained accuracy (86% and 68%) with fairness insights, at a 45% model size increase. This extended abstract of the original paper essentially presents the first contribution.

2 Background

We briefly review the essential concepts for our approach.

Multi-output learning Unlike traditional single-label classification, multi-output classification predicts multiple outputs simultaneously from the same input features. The learning process requires data samples to be labeled accordingly. That is, $\mathcal{D} = \{(x_1, y_1^{(1)}, \ldots, y_1^{(k)}), \ldots, (x_n, y_n^{(1)}, \ldots, y_n^{(k)})\}$ and the optimization is performed on the loss of every output and the corresponding label, as such, at each learning iteration t :

*Published at the *European Conference on Artificial Intelligence* ECAI-2024 [Bendoukha *et al.*2024]

$$L_j(\theta) = \frac{1}{|\mathcal{B}|} \sum_{x_i \in \mathcal{B}} \mathcal{L}(x_i, y_i^{(j)}, \theta_t) \ \forall j \in \{1, \cdots, k\}$$

$$\text{and} \quad \theta^{t+1} = \text{argmin}_\theta \ L(\theta) = \frac{1}{k} \sum_{j=1}^{k} L_j(\theta^t)$$

For convex loss functions, a common approach for optimization is to use Stochastic Gradient Descent (SGD). This iterative algorithm processes the data in batches ($\mathcal{B} \subset \mathcal{D}$) and updates the model parameters (θ) through gradient descent. The learning rate (η) controls the step size of these updates, guiding the parameters towards a minimum of the loss function, such that:

$$g_t = \nabla L(\theta_t) \ \text{(Gradients computation)}$$
$$\theta_{t+1} = \theta_t - \eta g_t \ \text{(Parameters update)}$$

Fairness in Machine Learning Group fairness assesses whether a model's predictions $\hat{\mathcal{Y}}$ are independent of a sensitive attribute $\mathcal{S}$ (e.g., gender or race).

We review key fairness metrics at the data and classifier levels, assuming $\mathcal{S}$ is binary with values $\{s_0, s_1\}$.

Classifier unfairness
These metrics assess how data biases affect a model's outputs.

- **Statistical Parity Difference (SPD)** [Chouldechova2016] measures the gap in positive prediction rates:

  $$\text{SPD} = |P(\hat{\mathcal{Y}} = 1 | \mathcal{S} = s_0) - P(\hat{\mathcal{Y}} = 1 | \mathcal{S} = s_1)|.$$

  A dummy classifier predicting all positives yields zero SPD but poor utility.

- **Equal Opportunity Difference (EOD)** [Hardt *et al.*2016] compares true positive rates across groups:

  $$\text{EOD} = \Big| P(\hat{\mathcal{Y}} = 1 \mid \mathcal{S} = s_0, \mathcal{Y} = 1) - P(\hat{\mathcal{Y}} = 1 \mid \mathcal{S} = s_1, \mathcal{Y} = 1) \Big|. \quad (1)$$

Other metrics (e.g., FPD and FND) examine disparities in false-positive and false-negative rates across groups.

3 Related Work

Accuracy vs fairness trade-off Fairness-aware training often leads to reduced predictive performance, with many studies [Kamiran and Calders2011, Liu and Vicente2022, Fish *et al.*2016, Xu *et al.*2018, Wang *et al.*2021] highlighting an inherent trade-off between accuracy and fairness. However, Wick *et al.* [Wick *et al.*2019] suggest this trade-off may be avoidable under certain conditions. Kamiran and Calders [Kamiran and Calders2011] show a linear drop in accuracy from fairness constraints, while Liu *et al.* [Liu and Vicente2022] use Pareto fronts to illustrate the trade-off. Fish *et al.* [Fish *et al.*2016] improve fairness by shifting decision boundaries based on prediction confidence. Wang *et al.* [Wang *et al.*2021] explore fairness in multi-task learning. Overall, there is a growing consensus that improving fairness often compromises accuracy.

Delivering fairness insights along with prediction Recent works on individual fairness [Yadav *et al.*2024, Gajane and Pechenizkiy2018] and explainable AI (XAI) [Jain *et al.*2020, Alikhademi *et al.*2021] explore combining fairness with optimal accuracy. In individual fairness, this is framed as finding the largest neighborhood around an input x^* where predictions remain consistent:

$$\max_{\epsilon} \ \forall x : d(x^*, x) \leq \epsilon \text{ and } f(x) = f(x^*)$$

Yadav *et al.* [Yadav *et al.*2024] define fairness certificates per input based on distances in key features. In XAI, fairness is assessed via feature attribution—for example, Jain *et al.* [Jain *et al.*2020] use Shapley values to quantify the effect of sensitive features. Maughan *et al.* [Maughan and Near2020] introduce *prediction-sensitivity*, which computes the gradient of the model output with respect to sensitive inputs, with its norm indicating local unfairness.

In this work, the main challenge is developing classifiers aware of their inherent fairness risk for each input. This risk requires a clear definition and incorporation into the training data while ensuring that classifiers will still learn properly.

4 Towards a Dual Label Fair Learning

We assume a single binary sensitive attribute $\mathcal{S} \in \{s_0, s_1\}$ and a set of non-sensitive attributes denoted $\mathcal{X}$ (as presented in Figure 1). Subsequently, we consider a first binary classification task with $\mathcal{Y} = \mathcal{Y}_{\text{bin}} \in \{0, 1\}$. $\mathcal{Y}_{\text{bin}}$ corresponds to the default classification task of the trained model. Then, we propose to extract a second fairness class label $\mathcal{Y}_{\text{fair}}$, that represents a fair-aware assignation of outcomes with respect to groups $\mathcal{S} = s_0$ or $\mathcal{S} = s_1$.

4.1 Generating $\mathcal{Y}_{\text{fair}}$

We formulate the problem of finding fair-aware class labels, as finding an optimal vector $\mathcal{Y}_{\text{fair}}$ that maximizes the correlation with the default labels $\mathcal{Y}_{\text{bin}}$, and minimizes the correlation with the sensitive attribute vector $\mathcal{S}$ (as presented in Figure 1).

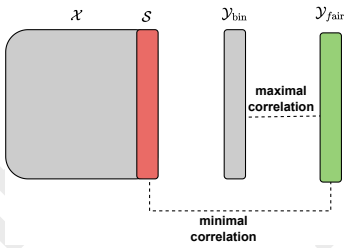


Figure 1: Correlation properties of the generated fairness class labels.

We use the Pearson correlation coefficient to generate $\mathcal{Y}_{\text{fair}}$ as it captures linear dependence and is zero for independent variables, enabling influence modeling without inherent bias. Its linearity also supports optimization tasks [He *et al.*2020]. We recall that Pearson's correlation satisfies:

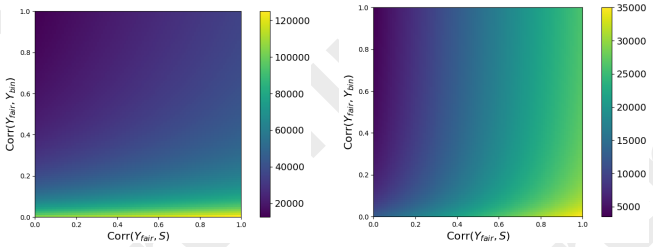
$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Where $\mathsf{Cov}(X, Y)$ is the covariance of X and Y, and σ_X, σ_Y are respectively standard deviations of X and Y. Correlation values range from -1 to 1. A correlation of both 1 and -1 indicates a deterministic functional relationship between the two variables $Y = f(X)$. A positive correlation indicates that X and Y follow the same variation (increasing f, when the correlation is equal to 1). Conversely, a negative correlation indicates opposite variations of X and Y (decreasing f when the correlation is equal to -1).

We leverage the transitive property of Pearson's correlation[1]. Since $\mathcal{Y}_{\text{bin}}$ reflects the non-sensitive labels ($\mathcal{X}$), maximizing its correlation with the fair prediction ($\mathcal{Y}_{\text{fair}}$) helps maintain the relationship between $\mathcal{Y}_{\text{fair}}$ and $\mathcal{X}$. This leads to a high-performing classifier trained on data containing sensitive attributes ($\mathcal{S}$), non-sensitive attributes ($\mathcal{X}$), and fair predictions ($\mathcal{Y}_{\text{fair}}$). Importantly, minimizing the correlation between $\mathcal{Y}_{\text{fair}}$ and the sensitive attribute ($\mathcal{S}$) reduces bias in the fair predictions. This results in data with less disparate treatment based on the sensitive attribute compared to the original data. These requirements are expressed using the function F_λ defined as:

$$F_\lambda(\mathcal{Y}_{\text{fair}}) = \mathsf{dim}(\mathcal{Y}_{\text{bin}}) \mid \mathsf{Corr}(\mathcal{Y}_{\text{fair}}, \mathcal{S}) \mid + \frac{\lambda}{\mid \mathsf{Corr}(\mathcal{Y}_{\text{fair}}, \mathcal{Y}_{\text{bin}}) \mid} \tag{2}$$

Where λ is a trade-off parameter that reflects the relation between both terms of Equation 2. Figures 2a and 2b show F_λ with $\lambda = 150$ and $\lambda = 300$, respectively. These two values of lambda induce different variations of the function. For example, a variation in the x-axis (i.e., in the correlation of $\mathcal{S}$ and $\mathcal{Y}_{fair}$) has a larger impact on F_λ when $\lambda = 300$.



(a) $\lambda = 150$ (b) $\lambda = 300$

Figure 2: Impact of λ on F_λ as a function of $\mid \mathsf{Corr}(\mathcal{Y}_{\text{bin}}, \mathcal{Y}_{\text{fair}}) \mid$, and $\mid \mathsf{Corr}(\mathcal{S}, \mathcal{Y}_{\text{fair}}) \mid$ with $\lambda \in \{150, 300\}$

We deduce from the previous observation that higher values of λ result in highly fair-aware class label $\mathcal{Y}_{\text{fair}}$. But, they sacrifice the accuracy of the induced classifier (trained on the $(\mathcal{X}, \mathcal{S}, \mathcal{Y}_{\text{fair}})$ records) due to the loss of useful correlations. Conversely, lower values of λ prioritize the first term and, therefore, produce a highly accurate classifier with slightly improved fairness.

4.2 Optimization strategy

The F_λ function can be categorized as a pseudo-Boolean function according to the PBO definition. A pseudo-Boolean function f is a function that maps a set of binary variables (0 or 1) to real numbers such as $f : \{0, 1\}^n \rightarrow R$. Several approaches to solving non-linear PBO problems are investigated, including the use of constrained integer programming methods. These methods aim to minimize an objective function subject to constraints on the function variables.

In our case, we introduce hard constraints[2] to limit the search space to the binary space of dimension $|\mathcal{D}|$. We use the Constraint Optimization BY Linear Approximation (COBYLA) solver [Powell1994] which is particularly suited for non-linear cost functions with hard constraints. Since this solver does not handle equality constraints, we define the constraints of the boolean solution as two inequality constraints satisfying:

$$\begin{aligned} &\text{minimize}: \ F_\lambda(\mathcal{Y}_{\text{fair}}) \\ &\text{subject to}: \ \mathcal{Y}_{\text{fair}}[i] \geq 1 - \epsilon \text{ or } \mathcal{Y}_{\text{fair}}[i] \leq \epsilon \ (\forall i \in [\mathsf{dim}(\mathcal{Y}_{\text{fair}})]) \end{aligned}$$

where $\epsilon = 10^{-5}$ characterizes the constraints on solutions within narrow intervals around the values of 0 or 1. Finally, the solver is run with a maximum of $10k$ iterations with $\mathcal{Y}_{\text{bin}}$ given as the initial guess.

4.3 Learning the dual label

Once the fair class label $\mathcal{Y}_{\text{fair}}$ are generated, the learning objective becomes the mapping: $\mathcal{X}, \mathcal{S} \rightarrow (\mathcal{Y}_{\text{bin}}, \mathcal{Y}_{\text{fair}})$. Indeed, the classifier makes two predictions for each data point (x):

- $\hat{y}_{\text{bin}}$: this prediction focuses solely on accurately matching the default label for the given input (x).
- $\hat{y}_{\text{fair}}$: this prediction represents a fairer and more ethical statistical outcome with respect to the sensitive attribute ($\mathcal{S}$).

This essentially converts the original binary classification task into a multi-output classification. Figure 3 depicts a neural network architecture based on our proposed framework for dual labeling. Applied to the Adult dataset, the figure illustrates how the network achieves this task with two separate branches. Ideally, both predictions, $\hat{y}_{\text{bin}}$ and $\hat{y}_{\text{fair}}$, would be the same. However, in cases where these predictions differ, the classifier identifies these discrepancies and acts as a per-prediction unfairness alert system, prompting further human evaluation of the specific data point.

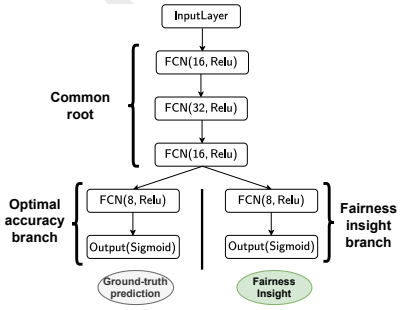


Figure 3: FairCognizer architecture for Adult.

[1] $\mathcal{Y}_{\text{bin}}$ acts as a statistical proxy of the non-sensitive labels $\mathcal{X}$.

[2] Hard constraints must be satisfied by all variables, while soft constraints only impose penalties on variables that fail to meet them.

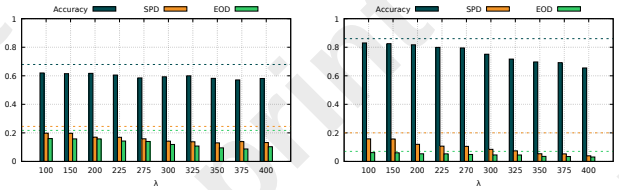
4.4 FairCognizer implementation

To evaluate the performance of our framework, we conduct various experiments on two datasets: Adult (25k samples) and Compas (5k samples). For our analysis, we focus on the binary groups male and female within these datasets.

Training analysis

First, we compute the $\mathcal{Y}_{\text{fair}}$ vector using the COBYLA solver implementation of the SciPy package [Virtanen *et al.*2020]. We generate $\mathcal{Y}_{\text{fair}}$ for our training subsets, i.e., $\frac{3}{4}$ of both datasets, for the binary groups.

Second, for validation, we train a three-layer MLP classifier on the records $(\mathcal{X}, \mathcal{S}, \mathcal{Y}_{\text{fair}})$ for 10 values of λ, and compare their predictive capabilities and fairness metrics (EOD and SPD) with a baseline classifier trained on the original data. Figure 4 depicts the obtained results.



(a) Compas (b) Adult

Figure 4: Accuracy and fairness measures of classifiers trained on $(\mathcal{X}, \mathcal{S}) \rightarrow \mathcal{Y}_{\text{fair}}$, with $\mathcal{Y}_{\text{fair}}$ generated using different values of λ. Dashed lines represent the baseline measures (color-wise) of a classifier trained on the original dataset.

It shows greater fairness improvement using $\mathcal{Y}_{\text{fair}}$ on the Adult dataset compared to Compas. Indeed, our method mainly removes the disparate treatment (direct discrimination), but does not act on the disparate impact (indirect correlation) contained in the dataset. The predominant source of unfairness in the Adult dataset is disparate treatment, whereas in Compas, the primary origin of unfairness is disparate impact[3]. Finally, we train larger FairCognizer classifiers (from 2500 to nearly 3800 trainable parameters) on data-records $(\mathcal{X}, \mathcal{S}, (\mathcal{Y}_{\text{bin}}, \mathcal{Y}_{\text{fair}}))$. We measure their predictive performances and fairness with respect to the initial labels, the fair ones, and the dual-label. Table 1 presents the obtained results on both outputs of the FairCognizer classifier. We note from Table 1 that FairCognizer achieves optimal accuracy for the default class label. This means it prioritizes fairness for the second prediction, $\hat{y}_{fair}$, without compromising accuracy on the original prediction.

Classification interpretation

The four possible predictions made by the dual-label classifier on test data are examined, especially data-records for which the classifier's prediction is $(1, 0)$ or $(0, 1)$ $(\hat{y}_{\text{bin}} \neq \hat{y}_{\text{fair}})$. We measure the prediction inconsistency rate (PIR). That is, we compute the probability: $P(\lceil h(x) \rfloor \neq \lceil h(\overline{x}^{\mathcal{S}}) \rfloor)$, where h is a classifier trained on the samples $(\mathcal{X}, \mathcal{S} \rightarrow \mathcal{Y}_{\text{bin}})$ and $\overline{x}^{\mathcal{S}}$ is the sample x where the binary value of $\mathcal{S}$ is flipped.

[3]For reference, the BER in Adult for gender as the sensitive attribute $\mathcal{S}$ is 0.183, compared to 0.087 in Compas.

Measures		Label		
		$\mathcal{Y}_{\text{bin}}$	$\mathcal{Y}_{\text{fair}}$	$(\mathcal{Y}_{\text{bin}}, \mathcal{Y}_{\text{fair}})$
Adult	Acc	**0.8543**	0.8258	0.9100
	Precision	0.7308	0.7660	0.8211
	Recall	0.6790	0.6331	0.8182
	f1-score	0.7039	0.6932	0.8196
	SPD	0.1624	**0.1146**	–
	EOD	0.0599	**0.0230**	–
Compas	Acc	**0.6898**	0.6749	0.8388
	Precision	0.6768	0.6823	0.6795
	Recall	0.6098	0.5896	0.6723
	f1-score	0.6415	0.6326	0.6758
	SPD	0.1830	**0.1150**	–
	EOD	0.1510	**0.0879**	–

Table 1: Dual-output model predictive performances with respect to $\mathcal{Y}_{\text{bin}}$, $\mathcal{Y}_{\text{fair}}$ and the dual label $(\mathcal{Y}_{\text{bin}}, \mathcal{Y}_{\text{fair}})$, and fairness. $\mathcal{Y}_{\text{fair}}$ is obtained with $\lambda = 250$. Dual-label metrics are average-weighted by the number of true instances in each class.

We observe that for the subset of data-records with dual predictions $\hat{y}_{\text{bin}} \neq \hat{y}_{\text{fair}}$, the prediction inconsistency rate is significantly higher compared to data-records for which $\hat{y}_{\text{bin}} = \hat{y}_{\text{fair}}$.

	$(\hat{y}_{\text{bin}} \neq \hat{y}_{\text{fair}})$	$(\hat{y}_{\text{bin}} = \hat{y}_{\text{fair}})$
Adult PIR	58.3%	4.1%
Compas PIR	64.7%	19.0%

Table 2: Prediction inconsistency rates across 1000 sampled data-records for which the fair prediction equals the accurate one, and on records for which the fair prediction is different from the accurate one.

Table 2 shows that the dual-label model can identify unfair predictions, even if they are accurate. These predictions occur for data points similar to the ones from the opposite sensitive group ($\mathcal{S}$), but with different labels. By similar, we refer to close data-records with respect to non-sensitive attributes $\mathcal{X}$. Indeed, flipping the sensitive attribute value for such a point is likely to flip the model prediction (as indicated by the high PIR values on these points).

5 Conclusion

In this work, we explore a novel paradigm of fairness-aware learning that can be succinctly described as follows: If a classifier cannot simultaneously achieve optimal accuracy and group fairness, it can still provide valuable per-prediction insights about fairness risks. We provide two different hybrid pre-processing and in-processing approaches to implement this paradigm. Our study introduces a nuanced analysis of unfairness that encompasses both the classifier and the individual data records. Specifically, individuals exhibit varying degrees of vulnerability to the group-unfairness of a classifier. This nuanced perspective enables a targeted approach for enhancing fairness by directing efforts towards the most susceptible subset of data records affected by classifier unfairness.

Acknowledgments

This work was supported by the France ANR project ANR-22-CE39- 0002 EQUIHID

References

[Alikhademi *et al.*, 2021] Kiana Alikhademi, Brianna Richardson, Emma Drobina, and Juan E. Gilbert. Can explainable ai explain unfairness? a framework for evaluating explainable ai, 2021.

[Bendoukha *et al.*, 2024] Adda-Akram Bendoukha, Nesrine Kaaniche, Aymen Boudguiga, and Renaud Sirdey. FairCognizer: a model for accurate predictions with inherent fairness evaluation. In *ECAI 2024 : 27TH EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE*, Frontiers in Artificial Intelligence and Applications, Santiago de Compostela, SPAIN, Spain, October 2024. IOS Press.

[Bendoukha *et al.*, 2025] Adda-Akram Bendoukha, Didem Demirag, Nesrine Kaaniche, Aymen Boudguiga, Renaud Sirdey, and Sébastien Gambs. Towards privacy-preserving and fairness-aware federated learning framework. In *PETs 2025 : Privacy Enhancing Technologies*, volume 2025, pages 845–865, Washinghton, DC, United States, July 2025.

[Chen *et al.*, 2019] Irene Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21:E167–179, 02 2019.

[Chouldechova, 2016] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016.

[Cohen *et al.*, 2020] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Duong, and Marzyeh Ghassem. Covid-19 image data collection: Prospective predictions are the future. *Machine Learning for Biomedical Imaging*, 1(December 2020):1–38, December 2020.

[Dressel and Farid, 2018] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018.

[Fabris *et al.*, 2022] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152, September 2022.

[Falletti, 2023] Elena Falletti. Algorithmic discrimination and privacy protection. *Journal of Digital Technologies and Law*, 1:387–420, 06 2023.

[Feldman *et al.*, 2015] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact, 2015.

[Fish *et al.*, 2016] Benjamin Fish, Jeremy Kun, and Ádám D. Lelkes. A confidence-based approach for balancing fairness and accuracy, 2016.

[Gajane and Pechenizkiy, 2018] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning, 2018.

[Ghassemi *et al.*, 2014] Marzyeh Ghassemi, Tristan Naumann, Finale Doshi velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. *KDD: Proc Int Con Knowl Discov Data Mining.*, 2014:75–84, 08 2014.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

[He *et al.*, 2019] Yuzi He, Keith Burghardt, and Kristina Lerman. Learning fair and interpretable representations via linear orthogonalization, 2019.

[He *et al.*, 2020] Yuzi He, Keith Burghardt, and Kristina Lerman. A geometric solution to fair representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 279–285, New York, NY, USA, 2020. Association for Computing Machinery.

[Jain *et al.*, 2020] Aditya Jain, Manish Ravula, and Joydeep Ghosh. Biased models have biased explanations, 12 2020.

[Kamiran and Calders, 2011] Faisal Kamiran and Toon Calders. Data pre-processing techniques for classification without discrimination. *Knowledge and Information Systems*, 33, 10 2011.

[Liu and Vicente, 2022] Suyun Liu and Luis Nunes Vicente. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach, 2022.

[Maughan and Near, 2020] Krystal Maughan and Joseph P. Near. Towards a measure of individual fairness for deep learning, 2020.

[Mittelstadt *et al.*, 2016] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679, 2016.

[Powell, 1994] M. J. D. Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. 1994.

[Seyyed-Kalantari *et al.*, 2021] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

[Virtanen *et al.*, 2020] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[Wang *et al.*, 2021] 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 amp; Data Mining*, KDD '21. ACM, August 2021.

[Wen *et al.*, 2022] David Wen, Saad M Khan, Antonio Ji Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos de Blas Perez, Alastair K Denniston, Xiaoxuan Liu, et al. Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health*, 4(1):e64–e74, 2022.

[Wick *et al.*, 2019] Michael Wick, swetasudha panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[Xu *et al.*, 2018] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks, 2018.

[Yadav *et al.*, 2024] Chhavi Yadav, Amrita Roy Chowdhury, Dan Boneh, and Kamalika Chaudhuri. Fairproof : Confidential and certifiable fairness for neural networks, 2024.

[Zafar *et al.*, 2017] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification, 2017.