Journal of Artificial Intelligence Research 81 (2024) 307-335 Submitted 02/2024; published 10/2024

Uncertainty as a Fairness Measure

Selim Kuzucu SELIM.KUZUCU@METU.EDU.TR
Department of Computer Engineering
Middle East Technical University
06800 Ankara, Turkiye

Jiaee Cheong JC2208@CAM.AC.UK
Department of Computer Science
University of Cambridge
Cambridge, CB3 0FD, United Kingdom
The Alan Turing Institute
London, NW1 2DB, United Kingdom

Hatice Gunes HG410@CAM.AC.UK
Department of Computer Science
University of Cambridge
Cambridge, CB3 0FD, United Kingdom

Sinan Kalkan SKALKAN@METU.EDU.TR
Department of Computer Engineering &
ROMER Robotics-AI Center
Middle East Technical University
06800 Ankara, Turkiye

Abstract

Unfair predictions of machine learning (ML) models impede their broad acceptance in real-world settings. Tackling this arduous challenge first necessitates defining what it means for an ML model to be fair. This has been addressed by the ML community with various measures of fairness that depend on the prediction outcomes of the ML models, either at the group-level or the individual-level. These fairness measures are limited in that they utilize point predictions, neglecting their variances, or uncertainties, making them susceptible to noise, missingness and shifts in data. In this paper, we first show that a ML model may appear to be fair with existing point-based fairness measures but biased against a demographic group in terms of prediction uncertainties. Then, we introduce new fairness measures based on different types of uncertainties, namely, aleatoric uncertainty and epistemic uncertainty. We demonstrate on many datasets that (i) our uncertainty-based measures are complementary to existing measures of fairness, and (ii) they provide more insights about the underlying issues leading to bias.

1. Introduction

An impedance to the wide-spread use of machine learning (ML) approaches is the bias present in their predictions against certain demographic groups. The severity and extent of this matter have been considerably investigated for different applications, such as gender recognition (Buolamwini & Gebru, 2018), emotion or expression recognition (Domnich & Anbarjafari, 2021; Xu, White, Kalkan, & Gunes, 2020; Chen & Joo, 2021) and mental health prediction (Cheong, Kuzucu, Kalkan, & Gunes, 2023) etc.

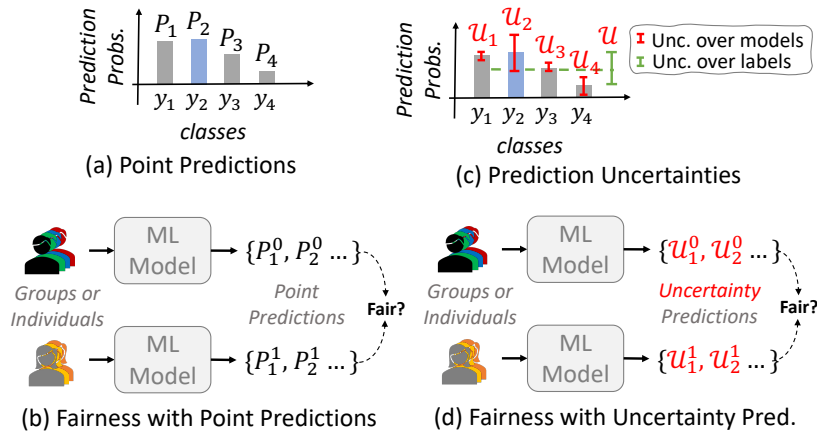


Figure 1: Existing fairness measures utilize point predictions for quantifying fairness, which ignores the uncertainty (variance) of the predictions **(a-b)**. We fill this gap by using uncertainty instead for measuring fairness **(c-d)**.

It has been identified in the literature that fairness is a multi-faceted concept, which has led to different notions and definitions of fairness (Garg, Villasenor, & Foggo, 2020; Verma & Rubin, 2018a; Castelnovo, Crupi, Greco, Regoli, Penco, & Cosentini, 2022; Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012; Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021). For example, a model can be considered fair at a group-level (called *group fairness*) if its predictions are the same for the different demographic groups (a.k.a., *Statistical Parity* – (Dwork et al., 2012; Mehrabi et al., 2021; Garg et al., 2020; Verma & Rubin, 2018a)) or if its false negative rates are the same (a.k.a., *Equal Opportunity* – (Hardt, Price, & Srebro, 2016)). Alternatively, a model can be evaluated for fairness at the level of individuals (called *individual fairness*) by comparing an individual's predictions to similar individuals (Dwork et al., 2012) or to a counterfactual version of the individual (called *counterfactual fairness* – (Kusner, Loftus, Russell, & Silva, 2017; Cheong, Kalkan, & Gunes, 2022)).

Despite advances in fairness quantification measures using point predictions (Fig. 1(a,b)) and bias mitigation methods, the utility of such measures or methods are often limited when exposed to real-world data. This is because (**P1**) first, they often do not account for real-world problems such as missing data (Goel, Amayuelas, Deshpande, & Sharma, 2021), biased labeling (Jiang & Nachum, 2020) and domain or distribution shifts (Chen, Raab, Wang, & Liu, 2022). (**P2**) Second, they are susceptible to fairness gerrymandering. For instance, depending on how a group is defined, a key challenge with existing statistical-parity point-based fairness measures is that it is implausible to ensure they hold for every subgroup of the population. Any classifier can be deemed unfair to the subgroup of individuals defined ex-post as the set of samples it misclassified (Kearns, Neel, Roth, & Wu, 2018). (**P3**) Third, recent works have demonstrated how traditional bias mitigation methods do not necessarily lead to fairer outcomes as measured using traditional parity-based measures nor do they shed light on the source of bias. For instance, larger or more balanced datasets did not mitigate the embedded disparities in *real-world tabular datasets* (Ding, Hardt, Miller, & Schmidt, 2021) and balancing samples across gender did not produce fairer predictions for

females (Cheong et al., 2023). We propose addressing these challenges by measuring fairness using prediction uncertainties (Fig. 1(c,d)).

1.1 Uncertainty-based Fairness for Social Impact.

An uncertainty-based definition of fairness has the potential of addressing the aforementioned drawbacks (P1-P3):

1. **Addressing P1**: Point predictions calculated using $P(Y|X)$ are often unreliable (Guo, Pleiss, Sun, & Weinberger, 2017; Baltaci, Oksuz, Kuzucu, Tezoren, Konar, Ozkan, Akbas, & Kalkan, 2023; Mukhoti, Kulharia, Sanyal, Golodetz, Torr, & Dokania, 2020) and uninformative in real-world problems (Naik, Kalkan, & Kruger, 2024; Han, Canli, Shah, Zhang, Dino, & Kalkan, 2024) with missing data, labeling or data noise and distribution shifts. Uncertainty-based fairness addresses P1 by quantifying the level of unreliability present to provide practitioners with an indication of the potential source of underlying bias present. As we will show in our paper, (i) machine learning models have different prediction uncertainties for different demographic groups and (ii) these differences can provide useful insights, e.g., about a lack of data or a presence of noise affecting one demographic group more than others, which are fundamental to fairness.

2. **Addressing P2**: Prediction uncertainties quantify variance over multiple predictions for the same input, which make them less susceptible or less vulnerable towards manipulation. They represent the inherent uncertainty about the model and the data, which is largely immutable for a given model and a dataset.

3. **Addressing P3**: Quantification of different types of uncertainty by definition provides insights about the underlying issues with the data and the model, which can shed light on e.g. when adding more data does not necessarily lead to fairer outcomes.

1.2 Contributions.

In summary, our main contributions are:

- We **introduce uncertainty-based fairness measures at the group and individual-level**. To the best of our knowledge, our paper is the first to use uncertainty as a fairness measure.

- We **prove that an uncertainty-based fairness measure is complementary to point-based measure**s, suggesting that both uncertainty and point predictions should be taken into account when analyzing fairness of models.

- We show on many datasets that (i) uncertainty fairness can vary significantly across demographic groups and (ii) it **provides insight about the sources of bias**.

2. Related Work

2.1 Fair ML

The seminal work of (Buolamwini & Gebru, 2018) and the follow-up studies (Domnich & Anbarjafari, 2021; Xu et al., 2020; Chen & Joo, 2021; Cheong, Kalkan, & Gunes, 2023) have exposed significant bias present in many applications of ML models. To address such biases and obtain fairer ML models, the ML community have proposed a plenitude of pre-processing, in-processing or post-processing strategies with promising outcomes – see (Barocas, Hardt, & Narayanan, 2017; Mehrabi et al., 2021; Cheong, Kalkan, & Gunes, 2021) for surveys.

2.2 Fairness Measures

Fairness has multiple facets, which have been recognized by the ML community with different notions and measures of fairness. One prominent notion of fairness is *group fairness*, which pertains to comparing a model's predictions across different demographic groups. Statistical Parity (Dwork et al., 2012; Mehrabi et al., 2021; Garg et al., 2020), Equal Opportunity (Hardt et al., 2016), and Equalized Odds (Hardt et al., 2016) are commonly used measures of group fairness. Alternatively, ML model predictions can be evaluated for *individual fairness* (Dwork et al., 2012). Such fairness can be measured e.g. by comparing an individual's predictions with those of similar individuals (Dwork et al., 2012) or with those of a counterfactual version of the individual (Kusner et al., 2017).

Wang et al. (Wang, He, Gao, & Calmon, 2023) study algorithmic discrimination with two measures of group fairness that are relevant to our fairness measures: Aleatoric discrimination, for inherent biases in data, and epistemic discrimination, for model or algorithmic biases. These discrimination measures are based on the gap between the performance of a model and the fairness Pareto frontier for that model. The fairness Pareto frontier represents the best achievable performance for a certain fairness constraint. The gap between this frontier and the 100% performance would characterize irreducible (aleatoric) discrimination of the model whereas the gap between the frontier and the current model's performance would represent reducible (epistemic) discrimination. Although these measures are valuable, obtaining the fairness Pareto frontier requires solving a sophisticated optimization problem. Wang et al. address this issue by making simplifications about the decision boundaries or the machine learning model, which limit the applicability of their approach in practice. Moreover, their approach is limited to only measuring group-level discrimination.

2.3 Bayesian Neural Networks (BNNs)

Bayesian Neural Networks (MacKay, 1992; Neal, 1995) operate by placing a prior distribution over the weights of a neural network, such that each weight is represented by a distribution parameterized by a mean and a standard deviation: $\omega_i = (\mu_i, \sigma_i)$. In addition to being robust against over-fitting (Gal et al., 2016; Blundell, Cornebise, Kavukcuoglu, & Wierstra, 2015), BNNs are known to work well with small datasets and propagate reliable uncertainty estimates as they bring a natural framework to estimate the first two moments of the predictive distribution (Gal et al., 2016).

Against these desirable properties, one key challenge with BNNs is to perform inference. Inference arise as a challenge due to the need to find the most probable weights (in the form of distributions) that have generated the data. Specifically, when we attempt to apply the Bayes' Theorem to obtain the true posterior on the weights, we often fail to do so as marginalizing the prior on the weights does not have an analytical solution for the complex cases (Gal et al., 2016). Owing to this issue, there is a need for approximating the posterior, which is often achieved by utilizing variational Bayesian approximation techniques such as Monte Carlo (MC) sampling. In practice, MC sampling is used not only for optimizing the BNNs (Section 5.2) but also for uncertainty quantification (Section 4.1).

2.4 BNNs and Uncertainty Quantification

Modern decision making systems should possess an awareness of unknowns and propagate this information to the people at the end of the decision making pipelines. In recent literature, this goal is aimed to be achieved by producing reliable uncertainty estimates, commonly referred to as uncertainty quantification (Gal & Ghahramani, 2016; Kendall & Gal, 2017; Mukhoti, Kirsch, van Amersfoort, Torr, & Gal, 2023; Van Amersfoort, Smith, Teh, & Gal, 2020b).

In Bayesian modeling, it is possible to disentangle the overall predictive uncertainty into two unique components: epistemic uncertainty to account for the lack of data and aleatoric uncertainty to account for any irreducible uncertainty associated with the data (Kendall & Gal, 2017). Epistemic uncertainty is modeled by capturing how much the weights vary given a set of data based on a prior placed on the weights whereas aleatoric uncertainty is modeled by quantifying the variance of the distribution placed over the outputs of the model (Kendall & Gal, 2017; Kwon, Won, Kim, & Paik, 2020). Due to the aforementioned intractability issue, both of these uncertainties are commonly captured using MC sampling (Gal & Ghahramani, 2016; Kendall & Gal, 2017; Kwon et al., 2020). Recently, by improving the methodology introduced in (Kendall & Gal, 2017), Kwon *et al.* (Kwon et al., 2020) proposed a novel method to quantify these two components without the need to optimize for a separate variance parameter, which we utilize in our work and formally describe in Section 4.1.

2.5 Uncertainty and Fairness

The relationship between the uncertainty and fairness has been subject to numerous recent studies, such as (Mehta, Shui, & Arbel, 2023; Tahir, Cheng, & Liu, 2023; Kaiser, Kern, & Rügamer, 2022). For example, (Mehta et al., 2023) have shown that mitigating bias has an adverse affect on the (predictive) uncertainty of estimations. Furthermore, (Tahir et al., 2023) with (Kaiser et al., 2022) utilize aleatoric uncertainty as part of their bias mitigation strategy. Our work is also distinct from existing work that attempts to quantify the uncertainty of a fairness measure (Roy & Mohapatra, 2023) or the bias present (Ethayarajh, 2020). In addition, it is well-known that ML models are generally under- or over-confident (Guo et al., 2017; Mukhoti et al., 2020) or unreliable under noise (Kendall & Gal, 2017). To illustrate that this may also be the case in datasets which are commonly used in fairness analysis, we plot the prediction confidence of the BNN classifier on the COMPAS dataset

in Fig. 2. The plot shows that the model is under or over-confident about its predictions and that there is a so-called calibration gap.

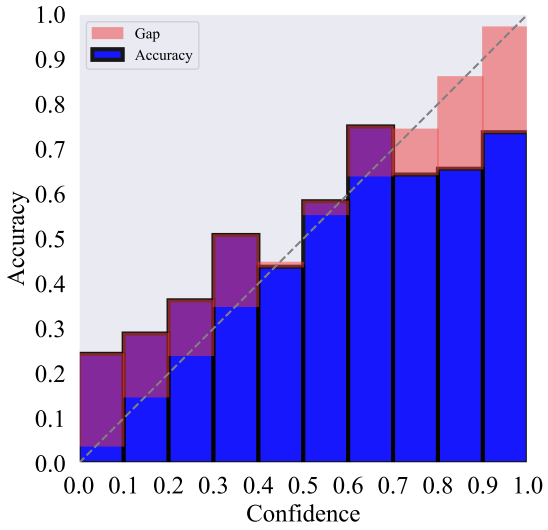


Figure 2: **Illustration of Under- or Over-Confidence:** An example illustrating under- or over-confidence of an ML model's predictions. The diagram is calculated for a Bayesian NN classifier on the COMPAS Dataset, a dataset frequently used in ML fairness research.

2.6 Comparative Summary

As discussed, existing fairness measures have only considered point predictions, which provide an incomplete view about the quality of a model's predictions. To the best of our knowledge, we are the first to address this gap by proposing uncertainty as a fairness measure. We prove that the introduced uncertainty measure is *complementary* to the point-based fairness measures. We showcase that (i) point-based fairness and uncertainty-based fairness can be complementary, and (ii) uncertainty fairness can provide insight about the sources of bias on several datasets.

3. Preliminaries and Background

In this section, we introduce some preliminary notation and background knowledge to aid subsequent understanding of the paper.

3.1 Notation

Following the setting and notation in the literature (Verma & Rubin, 2018b; Castelnovo et al., 2022), we assume a binary classification problem with a dataset D of X, Y and G where X denotes features describing an individual, $Y \in \{0, 1\}$ is the classification target,

and $G \in \{0, 1\}$ is the majority group indicator, with $G = 0$ denoting the minority. Solving the classification problem involves finding a mapping $\hat{Y} = f(X; \theta) \in \{0, 1\}$ with parameters θ. We use $P(Y = y_i|X = \mathbf{x}_i)$ to denote the predicted probability for the correct class y_i for sample $X = \mathbf{x}_i$, and $\hat{Y} = \hat{y}_i \leftarrow \arg\max_c P(Y = c|X = \mathbf{x}_i)$ to denote the predicted class.

3.2 Measuring Group Fairness

A ML model can be considered fair if a chosen performance measure for a specific task is the same across different groups (Garg et al., 2020; Verma & Rubin, 2018a). More formally, for a predictor $\hat{Y} = f(\ \cdot\ ; \theta)$ to be considered fair with respect to a demographic group attribute G, the following equality should be met for a given performance measure $\mathcal{M}$, e.g., true positive rate:

$$\text{Fair}(f; \mathcal{M}, D) \equiv \mathcal{M}(D, f, G = 0) = \mathcal{M}(D, f, G = 1). \tag{1}$$

Existing work exploring different performance measures for $\mathcal{M}$ has shown that each entails a different notion of fairness. For example:

Statistical Parity, or Demographic Parity (Dwork et al., 2012; Mehrabi et al., 2021; Garg et al., 2020; Verma & Rubin, 2018a): Compares model's prediction probabilities for the positive class ($\hat{Y} = 1$) across different groups (with $\mathcal{M}(D, f, G) \equiv P(\hat{Y} = 1|G)$):

$$P(\hat{Y} = 1|G = 0) = P(\hat{Y} = 1|G = 1). \tag{2}$$

Equal Opportunity (Hardt et al., 2016): Compares model's false negative rates, i.e., prediction probabilities for the negative class ($\hat{Y} = 0$) for the known positive class ($Y = 1$):

$$P(\hat{Y} = 0|Y = 1, G = 0) = P(\hat{Y} = 0|Y = 1, G = 1), \tag{3}$$

where $\mathcal{M}(D, f, G) \equiv P(\hat{Y} = 0|Y = 1, G)$.

Equalised Odds (Hardt et al., 2016): Compares model's prediction probabilities for the positive class ($\hat{Y} = 1$) for different ground truth classes ($Y = 1$ and $Y = 0$):

$$P(\hat{Y} = 1|Y = y, G = 0) = P(\hat{Y} = 1|Y = y, G = 1), \tag{4}$$

where $y \in \{0, 1\}$, and we've taken $\mathcal{M}(D, f, G) \equiv P(\hat{Y} = 1|Y = y, G)$.

3.3 Measuring Individual Fairness

Dwork *et al.* (Dwork et al., 2012) defined individual fairness based on a "similar individuals should have similar predictions" principle:

$$d_y(f(\mathbf{x}_1), f(\mathbf{x}_2)) \leq L d_x(\mathbf{x}_1, \mathbf{x}_2), \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}. \tag{5}$$

The above notion assumes suitable distance metrics $d_y(\cdot, \cdot)$ and $d_x(\cdot, \cdot)$ to be available for the predictions and the inputs respectively. The literature has used point predictions to quantify this notion of fairness e.g. by using a consistency measure (Zemel, Wu, Swersky, Pitassi, & Dwork, 2013; Mukherjee, Yurochkin, Banerjee, & Sun, 2020):

$$\mathcal{F}^{indv}_{\hat{y}}(X = \mathbf{x}_i) = 1 - \left| \hat{y}_i - \frac{1}{k} \sum_{\mathbf{x}_j \in k\text{NN}(\mathbf{x}_i)} \hat{y}_j \right|, \tag{6}$$

where $k\text{NN}(\mathbf{x}_i)$ denotes the k-nearest neighbours of $\mathbf{x}_i$.

4. Methodology

We first describe how we quantify uncertainty and then introduce the fairness measures.

4.1 Quantifying Uncertainty

ML models tend to be under- or over-confident about their predictions and unaware of distribution shift, adversarial attacks or noise in data (Abdar, Pourpanah, Hussain, Rezazadegan, Liu, Ghavamzadeh, Fieguth, Cao, Khosravi, Acharya, et al., 2021; Gawlikowski, Tassi, Ali, Lee, Humt, Feng, Kruspe, Triebel, Jung, Roscher, et al., 2021; Cetinkaya, Kalkan, & Akbas, 2024). Quantifying the variance of a model's predictions, i.e., *predictive uncertainty*, facilitates awareness of such hindrances with respect to the data. Predictive uncertainty has two components, reflecting the two different ways to define a variance over predictions:

- **Epistemic or model uncertainty** is measured over different models. Epistemic uncertainty reflects the lack of knowledge about the current input and can be reduced by providing more training data, i.e., more knowledge.
- **Aleatoric or data uncertainty** is measured over classes. Aleatoric uncertainty reflects the irreducible noise in the data.

We use Bayesian Neural Networks (BNNs) to obtain uncertainty estimates as described in (Blundell et al., 2015) since BNNs provide reliable uncertainty estimations. A BNN defines a distribution over each weight in the model: $\omega_i = (\mu_i, \sigma_i)$, which enables sampling different weights and making multiple predictions for the same input. With such a model, predictive uncertainty for a sample $\mathbf{x}$ with label y can be quantified as follows (Kwon et al., 2020; Shridhar, Laumann, & Liwicki, 2019):

$$\underbrace{\underbrace{\frac{1}{M}\sum_{m=1}^{M}(P_m-\bar{P})^T(P_m-\bar{P})}_{\text{Epistemic unc. } (\mathcal{U}_e)}+\underbrace{\frac{1}{M}\sum_{m=1}^{M}diag(P_m)-P_m^T\cdot P_m}_{\text{Aleatoric unc. } (\mathcal{U}_a)}}_{\text{Predictive uncertainty } (\mathcal{U}_p)}, \tag{7}$$

where $\bar{P} = \frac{1}{M}\sum_{m=1}^{M} P_m$ and $P_m = P(Y|X=\mathbf{x})$ of the m^{th} Monte Carlo sample with M being the number of Monte Carlo samples. To obtain group-wise uncertainty estimations, we aggregate the quantified uncertainty values for the samples of that group by averaging.

4.2 Uncertainty-based Group Fairness Measures

We now introduce our novel fairness notion based on averaged predictive uncertainty over groups where each group is defined by a set of sensitive attributes. For this, we use the uncertainty types and their quantification as outlined in Section 4.1 and extend the definition of fairness in Section 3.2.

Definition 4.1 (Uncertainty-Fairness Measure)**.** *A model is fair if its uncertainties are the same across different groups. More formally, extending the definition in Section 3.2*

$$\mathit{Fair}(f;\mathcal{U},D) \equiv \mathcal{U}(D,f,G=0) = \mathcal{U}(D,f,G=1), \tag{8}$$

where $\mathcal{U}$ is an uncertainty measure, e.g., predictive uncertainty ($\mathcal{U}_p$), epistemic uncertainty ($\mathcal{U}_e$), or aleatoric uncertainty ($\mathcal{U}_a$) as introduced in Section 4.1.

Proposition 4.1 (Independence of Uncertainty Fairness)**.** *Consider a predictor $f(\cdot;\theta)$ with point-predictions $\{\hat{y}_i\}_i$ (and associated probabilities $\{P(\hat{y}_i|\mathbf{x}_i)\}_i$) and uncertainties $\{\mathcal{U}_i\}_i$ (namely, predictive, epistemic and aleatoric). Then, uncertainty fairness Fair$(f;\mathcal{U},D)$ is independent to the conventional point-measure based fairness Fair$(f;\mathcal{M},D)$. More formally:*

- *Fair$(f;\mathcal{M},D)$ $\not\Longrightarrow$ Fair$(f;\mathcal{U},D)$.*
- *Fair$(f;\mathcal{U},D)$ $\not\Longrightarrow$ Fair$(f;\mathcal{M},D)$.*

Fair$(f;\mathcal{U},D)$ does not imply Fair$(f;\mathcal{M},D)$ or vice versa.

Proof. We will prove the two non-implications in the proposition using contradictions:

Proof of Fair$(f;\mathcal{M},D)$ $\not\Longrightarrow$ Fair$(f;\mathcal{U},D)$: we assume that the implication is true, i.e., Fair$(f;\mathcal{M},D)$ $\Longrightarrow$ Fair$(f;\mathcal{U},D)$. That means there may not be a predictor f which is $\mathcal{M}$-wise fair but $\mathcal{U}$-wise unfair. As contradiction, we select as examples the Synthetic Dataset 1 & 2 in Section 5.1(A,B) – see also Fig. 3. In this contradictory example, we see a predictor (namely, a BNN – see Section 5.2 for architecture and training details) which is $\mathcal{M}$-wise fair but $\mathcal{U}$-wise unfair (Table 1). Therefore, Fair$(f;\mathcal{M},D)$ $\Longrightarrow$ Fair$(f;\mathcal{U},D)$ is not necessarily true, and therefore, Fair$(f;\mathcal{M},D)$ $\not\Longrightarrow$ Fair$(f;\mathcal{U},D)$.

Proof of Fair$(f;\mathcal{U},D)$ $\not\Longrightarrow$ Fair$(f;\mathcal{M},D)$. We will follow the same reasoning for this non-implication: we assume that the implication is true, i.e., Fair$(f;\mathcal{U},D)$ $\Longrightarrow$ Fair$(f;\mathcal{M},D)$. That means there may not be a predictor f which is $\mathcal{U}$-wise fair but $\mathcal{M}$-wise unfair. As contradiction, we select the example in Sect. 5.1(C) – see also Fig. 3(c). In this example, we see a predictor (again, a BNN – see Sect. 5.2 for architecture and training details) which is $\mathcal{U}$-wise fair but $\mathcal{M}$-wise unfair (Table 1). Therefore, Fair$(f;\mathcal{U},D)$ $\Longrightarrow$ Fair$(f;\mathcal{M},D)$ is not necessarily true, and therefore, Fair$(f;\mathcal{U},D)$ $\not\Longrightarrow$ Fair$(f;\mathcal{M},D)$. □

4.3 Uncertainty-based Individual Fairness

We extend the definition in Eq. 5 to account for "similar individuals should have similar prediction **uncertainties**":

$$\mathcal{F}_{\mathcal{U}}^{indv}(X=\mathbf{x}_i) = 1 - \left| \mathcal{U}_i - \frac{1}{k} \sum_{\mathbf{x}_j \in k\text{NN}(\mathbf{x}_i)} \mathcal{U}_j \right|, \tag{9}$$

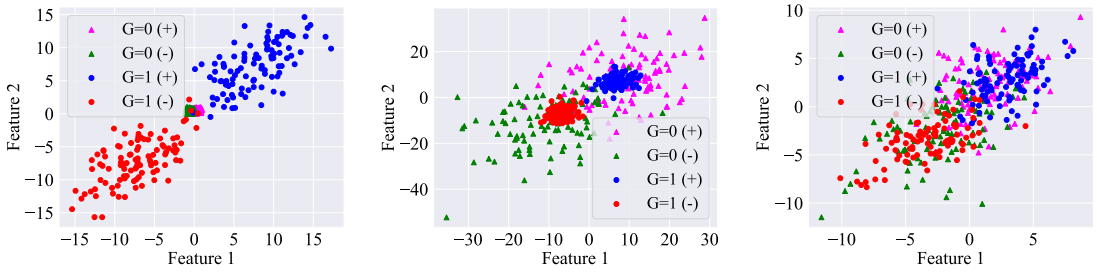
which we aggregate over a group by averaging.

5. Experiments

In this section, we introduce the datasets used, the implementation and training details as well as the evaluation measures used within the experiments.

5.1 Datasets

We introduce three synthetic datasets and utilize three real datasets to evaluate the measures. We adopt the approach of (Zafar, Valera, Gomez Rodriguez, & Gummadi, 2017) for all synthetic dataset curation. Each synthetic dataset has 320 samples with 20% reserved for testing.



(a) Synthetic Dataset 1 (b) Synthetic Dataset 2 (c) Synthetic Dataset 3

Figure 3: Two datasets that appear to be fair with point-based measures but unfair in terms of **(a)** aleatoric uncertainty and **(b)** epistemic uncertainty. In **(c)**, we see a set where the classifier is fair in terms of uncertainties (both epistemic and aleatoric) but unfair in terms of point-based measures.

5.1.1 (A) Synthetic Dataset 1 (SD1) - Case of Aleatoric Uncertainty

This dataset highlights how a classifier may appear to be fair in terms of point-based performance metrics yet unfair *in terms of aleatoric uncertainties*. We obtain 100 samples from each of the following four multivariate distributions, one for each of the respective attribute-label pair that we consider:

$$P(X|G=0,Y=0) = \mathcal{B}eta(\alpha=[0.5,0.5],\beta=[0.5,0.5]), \quad (10)$$
$$P(X|G=0,Y=1) = -\mathcal{B}eta(\alpha=[0.5,0.5],\beta=[0.5,0.5]), \quad (11)$$
$$P(X|G=1,Y=0) = \mathcal{N}([-7,-7],[15,10;10,15]), \quad (12)$$
$$P(X|G=1,Y=1) = \mathcal{N}([7,7],[15,10;10,15]). \quad (13)$$

5.1.2 (B) Synthetic Dataset 2 (SD2) - Case of Epistemic Uncertainty

This dataset highlights how a classifier may be fair with point-based fairness measures but *unfair in terms of epistemic uncertainties*. We obtain 100 samples from each of the following four multivariate distributions, one for each of the respective attribute-label pair that we consider:

$$P(X|G=0,Y=0) = \mathcal{N}([-10,-10],[100,30;30,100]), \quad (14)$$
$$P(X|G=0,Y=1) = \mathcal{N}([10,10],[100,30;30,100]), \quad (15)$$
$$P(X|G=1,Y=0) = \mathcal{N}([-7,-7],[5,1;5,1]), \quad (16)$$
$$P(X|G=1,Y=1) = \mathcal{N}([7,7],[5,1;1,5]). \quad (17)$$

5.1.3 (C) Synthetic Dataset 3 (SD3) - Fair in Uncertainty, Unfair in Predictions

This dataset aims to highlight how a classifier may be fair according to the proposed uncertainty-based fairness measures but *unfair in terms of point-based fairness measures*. We obtain 100 samples from each of the following four multivariate distributions, one for each of the respective attribute-label pair that we consider:

$$P(X|G=0,Y=0) = \mathcal{N}([-2,-2],[7,3;3,7]), \tag{18}$$
$$P(X|G=0,Y=1) = \mathcal{N}([2,2],[7,3;3,7]), \tag{19}$$
$$P(X|G=1,Y=0) = \mathcal{N}([-3,-3],[5,3;5,3]), \tag{20}$$
$$P(X|G=1,Y=1) = \mathcal{N}([3,3],[5,3;3,5]). \tag{21}$$

5.1.4 (D) COMPAS Recidivism Dataset

The COMPAS Recidivism Dataset is a dataset with criminal offenders' records generally used to predict recidivism (binary classification) (Angwin, Larson, Mattu, & Kirchner, 2022). The data contains 6172 samples with 14 features. We follow (Zafar et al., 2017) in terms of the considered attributes and dataset splits. We assume that the positive label ($Y=1$) stands for the cases where the subject has recidivated and vice versa.

5.1.5 (E) Adult Income Dataset

The Adult Income Dataset contains a $48K+$ samples with 14 features (Becker & Kohavi, 1996). The task is to predict if a person's annual income is greater than $\$50K$ ($Y=1$) or not. We do not consider the samples with missing entries, resulting in a total of $45K+$ samples. We adhere to the training-testing split provided by the authors.

5.1.6 (F) D-Vlog Depression Detection Dataset

The D-Vlog Depression Detection Dataset contains visual and acoustic features from Youtube videos of 555 depressed and 406 non-depressed samples belonging to 639 females and 322 males (Yoon, Kang, Kim, & Han, 2022). The authors truncated the videos with longer than $t=596s$ and zero-pad shorter ones. D-Vlog only provides the gender attribute for its samples. We follow the training and testing splits as provided by the authors. We assume that the positive label ($Y=1$) stands for the depressed class and vice versa.

5.2 Implementation and Training Details

For all of our experiments, except for D-Vlog, we utilize Bayesian Neural Networks (BNNs) for both classification and uncertainty estimation. To address the intractability of $P(Y|X)$, we utilize the well-known *Bayes by Backprop* method (Blundell et al., 2015) which minimizes the following objective consisting of a KL divergence (Kullback & Leibler, 1951) term and a numerically-stable negative log-likelihood term as proposed in (Kendall & Gal, 2017):

$$\mathcal{L}(\theta) = \sum_{m=1}^{M} [\underbrace{\log q_\theta(\omega_m) - \log P(\omega_m)}_{\text{KL divergence}} + \lambda \underbrace{\mathcal{L}_{NLL}(\hat{Y},Y)}_{\text{classification loss}}], \tag{22}$$

with λ being a constant.

For all experiments, we use the Adam optimizer (Kingma & Ba, 2017). Following (Kwon et al., 2020), we set $T = 10$ (the number of Monte Carlo samples for uncertainty quantification as defined in Section 4.1). Furthermore, following one of the settings provided in (Blundell et al., 2015), we use 10 Monte Carlo samples to approximate the variational posterior, $q_\theta(\omega)$, and sample the initial mean of the posterior from a Gaussian with $\mu = 0$ and $\sigma = 1$. The π value, weighting factor for the prior, is set to 0.5 and the two σ_1 and σ_2 values for the scaled mixture of Gaussians is set to 0 and 6 respectively. We consider λ from the BNN training objective to be 2000. We utilize early stopping to determine the number of training iterations for all experiments. In the following, we describe dataset specific details. In all cases, the hyper-parameters are tuned to avoid over-fitting:

Synthetic Datasets: As the datasets are relatively simple, we observe that BNNs with no hidden layers suffice for all three synthetic datasets. We train all of them for 5 epochs with a batch size of 8.

COMPAS Recidivism Dataset: We employ a BNN with a single hidden layer of size 100. We train the model for 10 epochs with a batch size of 256. Similar to (Chouldechova, 2017), we consider fairness with respect to race, gender and age. For the race attribute, we follow (Zafar et al., 2017) and focus on the fairness gap between black and white subgroups, considering African-Americans as the minority group, G_0. For the gender attribute, we designate females as the minority group (G_0) due to the class imbalance in favor of the male group. For the age attribute, we consider individuals younger than 25 to be the minority group and individuals older than 45 as the majority group since our classifier provided the worst result for those younger than 25 and overall best results for those older than 45. To keep the coverage to a binary setting, we do not consider individuals aged between 25 and 45. Extending the measures to such a multi-valued setting is straightforward (Xu et al., 2020) and left as future work.

Adult Income Dataset: We employ a BNN with no hidden layers where the intermediate size is 25. We train the model for 5 epochs with a batch size of 256. Although a deeper analysis could be conducted through considering other variables such as marriage status, highest education level, occupation and nationality, for the sake of consistency with the analysis of the other datasets, we limit the experiments to race, gender and age.

D-Vlog Depression Detection Dataset: D-Vlog samples have significantly larger dimensionality (596s of 136-dim visual and 25-dim acoustic features) compared to COMPAS and Adult, which turned out to be challenging for BNNs. Therefore, we utilize the transformer-based *Depression Detector* architecture proposed by (Yoon et al., 2022). For uncertainty estimation, we follow (Lakshminarayanan, Pritzel, & Blundell, 2017) to obtain T predictions with an ensemble of T models and use the same method (Eq. 4.1) as with uncertainty estimation with BNNS. Specifically, instead of performing MC forward passes, we train T different models on the same training set and consider their predictions in the same testing set during the uncertainty quantification process. We choose $T = 5$ as existing work indicates that performance tends to peak at that number (Havasi, Jenatton, Fort, Liu, Snoek, Lakshminarayanan, Dai, & Tran, 2020).

For all of training configurations, we directly use the setting of (Yoon et al., 2022) with a learning rate of 0.0002 and a batch size of 32, optimized for 50 epochs through the Adam

optimizer (Kingma & Ba, 2017). For the dropout rate, we empirically choose 0.1 though it was not explicitly provided by the authors in the original work. For more details on the architecture and the relevant training details, we refer the reader to (Yoon et al., 2022).

5.3 Evaluation Measures

We evaluate classification performance in terms of accuracy ($\mathcal{M}_{Acc}$), Positive Predictive Value ($\mathcal{M}_{PPV} = TP/(TP + FN)$), Negative Predictive Value ($\mathcal{M}_{NPV} = TN/(TN + FN)$), False Positive Rate ($\mathcal{M}_{FPR}$) and False Negative Rate ($\mathcal{M}_{NPR}$). Fairness measures ($\mathcal{F}$) are defined as follows (similar to e.g. (Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015; Xu et al., 2020; Cheong, Kalkan, & Gunes, 2024)):

$$\text{Statistical Parity:} \quad \mathcal{F}_{SP} = \frac{P(\hat{Y}=1|G=0)}{P(\hat{Y}=1|G=1)}, \tag{23}$$

$$\text{Equal Opportunity:} \quad \mathcal{F}_{EOpp} = \frac{P(\hat{Y}=0|Y=1,G=0)}{P(\hat{Y}=0|Y=1,G=1)}, \tag{24}$$

$$\text{Equalized Odds:} \quad \mathcal{F}_{EOdd} = \frac{P(\hat{Y}=1|Y=y,G=0)}{P(\hat{Y}=1|Y=y,G=1)}, \tag{25}$$

$$\text{Equal Accuracy:} \quad \mathcal{F}_{EAcc} = \frac{\mathcal{M}_{Acc}(D,f,G=0)}{\mathcal{M}_{Acc}(D,f,G=1)}, \tag{26}$$

$$\text{Uncertainty Fairness:} \quad \mathcal{F}_{u} = \frac{\mathcal{U}_{u}(D,f,G=0)}{\mathcal{U}_{u}(D,f,G=1)}, \tag{27}$$

where u can be *Alea* (Aleatoric), *Epis* (Epistemic) or *Pred* (Predictive).

6. Results

In this section, we discuss the results obtained across both the synthetic and real-world datasets.

6.1 Experiment 1: Synthetic Datasets

Here, we analyze the point-based and uncertainty-based fairness measures with SD1, SD2 and SD3.

Analysing Fair$(f;\mathcal{U},D) \not\Longrightarrow$ **Fair**$(f;\mathcal{M},D)$**.** With reference to SD 1 & 2 as introduced in Section 5.1 and in Fig. 3(a) and 3(b), we select the group with higher uncertainty estimations as the minority group, i.e., G_0. From Table 1, we see that the classifier (BNN) can solve the classification task with a good level of performance (with high accuracy and low mis-classification). Moreover, the classifier is fair in terms of the widely-used point-based measures ($\mathcal{F}_{SP}, \mathcal{F}_{Opp}, \mathcal{F}_{Odd}$ and $\mathcal{F}_{EAcc}$) with $|\mathcal{F}-1| \leq 0.2$ – following (Feldman et al., 2015). However, our uncertainty-based measures suggest that the classifier is significantly unfair in terms of aleatoric uncertainty (for SD1 with $\mathcal{F}_{Alea} = 4.68$) and epistemic uncertainty (for SD2 with $\mathcal{F}_{Epis} = 275$).

Analysing Fair$(f;\mathcal{M},D) \not\Longrightarrow$ **Fair**$(f;\mathcal{U},D)$ With reference to SD 3 as introduced in Section 5.1 and and Fig. 3(c), we select the group with the lower classification performance as the minority group, G_0. Results in Table 1 suggest that the classifier provides a good level of performance for the majority group (G_1) and that the classifier is unfair in terms of some

Table 1: Experiment 1: The analysis with SD1, SD2 and SD3 datasets. We see that, for both SD1 and SD2, the classifier is fair in terms of point-based measures ($|\mathcal{F} - 1| \leq 0.2$ – following (Feldman et al., 2015)) whereas it is unfair in terms of aleatoric uncertainty for SD1 and epistemic uncertainty unfair for SD2. We see the inverse for the SD3 dataset. Unfair values are **highlighted**.

	SD1		SD2		SD3	
Measure	G_0	G_1	G_0	G_1	G_0	G_1
Performance Measures						
$\uparrow \mathcal{M}_{Acc}$	0.95	0.95	0.95	0.95	0.74	0.93
$\uparrow \mathcal{M}_{PPV}$	0.95	0.90	0.95	0.95	0.62	0.96
$\uparrow \mathcal{M}_{NPV}$	0.94	0.95	0.94	0.94	0.93	0.91
$\downarrow \mathcal{M}_{FPR}$	0.06	0.05	0.05	0.06	0.38	0.04
$\downarrow \mathcal{M}_{FNR}$	0.05	0.05	0.05	0.05	0.07	0.08
$\downarrow \mathcal{U}_e$	0.0001	0.0001	0.0011	0.0004	0.0002	0.0002
$\downarrow \mathcal{U}_a$	0.4926	0.1053	0.1915	0.2193	0.3349	0.3229
$\downarrow \mathcal{U}_p$	0.4927	0.1054	0.1926	0.2197	0.3351	0.3231
Point-based Fairness Measures						
$\mathcal{F}_{SP}$	1.07		1.00		1.17	
$\mathcal{F}_{Opp}$	1.00		1.00		1.01	
$\mathcal{F}_{Odd}$	1.05		0.95		**7.90**	
$\mathcal{F}_{EAcc}$	1.00		1.00		**0.79**	
Uncertainty-based Fairness Measures (Ours)						
$\mathcal{F}_{Epis}$	1.01		**2.75**		1.05	
$\mathcal{F}_{Alea}$	**4.68**		0.87		1.04	
$\mathcal{F}_{Pred}$	**4.67**		0.88		1.04	

of the point-based fairness measures (namely, $\mathcal{F}_{Odd} = 7.9$ and $\mathcal{F}_{EAcc} = 0.79$). However, the classifier appears to be fair across the uncertainty-based fairness measures.

6.2 Experiment 2: Real-world Datasets

In this section, we analyze the fairness measures across the real-world datasets, i.e., COMPAS, Adult and D-Vlog.

6.2.1 The COMPAS Dataset.

With reference to Table 2, across **race**, results suggest that there is strong bias against African-Americans in terms of recidivism even though there are more samples for African-Americans: The classifier has a clear tendency to suggest a black person to recidivate ($\mathcal{M}_{FPR} = 0.34$ African-Americans vs. 0.10) and vice versa for Whites ($\mathcal{M}_{FNR} = 0.66$ for Whites vs. 0.25). The point-based fairness measures (except for $\mathcal{M}_{Eacc}$) capture this bias strongly, so do the uncertainty-based measures. Despite having more samples, African-Americans have higher $\mathcal{U}_e$ and $\mathcal{U}_a$, leading to significant unfairness in terms of uncertainty ($\mathcal{F}_{Epis}$ and $\mathcal{F}_{Alea}$).

Table 2: Experiment 2: The analysis with COMPAS. Unfair values ($|\mathcal{F}-1| > 0.2$, following (Feldman et al., 2015)) are **highlighted**. B/W: Black/White. F/M: Female/Male.

Measure	**B** (G_0)	**W** (G_1)	**F** (G_0)	**M** (G_1)
↑ Sample Size	3175	2103	1175	4997
Performance Measures				
↑ $\mathcal{M}_{Acc}$	0.70	0.68	0.77	0.68
↑ $\mathcal{M}_{PPV}$	0.69	0.68	0.70	0.68
↑ $\mathcal{M}_{NPV}$	0.72	0.68	0.78	0.68
↓ $\mathcal{M}_{FPR}$	0.34	0.10	0.05	0.25
↓ $\mathcal{M}_{FNR}$	0.25	0.66	0.67	0.39
↓ $\mathcal{U}_e$	0.0006	0.0004	0.0003	0.0006
↓ $\mathcal{U}_a$	0.2299	0.1578	0.1599	0.2053
↓ $\mathcal{U}_p$	0.2305	0.1583	0.1602	0.2059
Point-based Fairness Measures				
$\mathcal{F}_{SP}$	**2.84**		**0.31**	
$\mathcal{F}_{EOpp}$	**2.19**		**0.54**	
$\mathcal{F}_{EOdd}$	**1.57**		**0.40**	
$\mathcal{F}_{EAcc}$	1.03		1.13	
Uncertainty-based Fairness Measures (Ours)				
$\mathcal{F}_{Epis}$	**1.55**		**0.50**	
$\mathcal{F}_{Alea}$	**1.46**		**0.78**	
$\mathcal{F}_{Pred}$	**1.46**		**0.78**	

Across **gender**, females have significantly better prediction performance compared to Males, with the exception of $\mathcal{M}_{FNR}$ ($M_{FNR} = 0.67$ for Females vs. 0.39 for Males). This suggests that the classifier is biased to predict $Y = 0$ ("no recidivism") for Females. Both point-based and uncertainty-based fairness measures capture this bias against Males. Across epistemic uncertainty, we hypothesize that the classifier is less certain for Males. However, fairness gaps in terms of aleatoric uncertainty (0.78) and predictive uncertainty (0.78) are close to the acceptable fairness boundary (0.8), suggesting that the main issue across gender may be the sample imbalance problem across groups (see also Table 4).

In Table 3, we see the complete table for COMPAS, including the age attribute. Across age, we observe that almost all of the performance metrics are better for those age greater than 45 compared to the others, with the exception of $\mathcal{M}_{PPV}$ and $\mathcal{M}_{FNR}$. For $\mathcal{M}_{PPV}$, samples with ages between $25-45$ is the best with $\mathcal{M}_{PPV} = 0.72$ and for $\mathcal{M}_{FNR}$, samples with ages less than 25 is the best with $\mathcal{M}_{FNR} = 0.32$. As explained within Section 6.2, we compute both the point-based and proposed uncertainty-based measures by considering the subgroup age greater than 45 as the majority subgroup (G_1) and the subgroup age less than 25 as the minority subgroup (G_0).

Across the point-based fairness measures, we observe that all but one of the measures point to *unfair* predictions, with $\mathcal{F}_{SP} = 2.44$, $\mathcal{F}_{Opp} = 1.48$ and $\mathcal{F}_{Odd} = 2.37$. Similar to the race attribute, $\mathcal{F}_{EAcc} = 0.85$ claims *fair* predictions, with all of the other point-based

Table 3: Experiment 2: The analysis with the COMPAS Recidivism Dataset for race (Black vs. White), age (younger than 25 vs. older than 45) and gender (male vs. female) attributes. Severe values of fairness values ($|\mathcal{F} - 1| > 0.2$, following (Zanna et al., 2022)) are **highlighted**.

Measure	Race		Age			Gender	
	Black (G_0)	**White** (G_1)	**<25** (G_0)	**25-45** (G_1)	**>45** (G_0)	**Female** (G_1)	**Male**
↑ Sample Size	3175	2103	1347	3532	1293	1175	4997
Performance Measures							
↑ $\mathcal{M}_{Acc}$	0.70	0.68	0.64	0.71	0.75	0.77	0.68
↑ $\mathcal{M}_{PPV}$	0.69	0.68	0.64	0.72	0.64	0.70	0.68
↑ $\mathcal{M}_{NPV}$	0.72	0.68	0.63	0.70	0.78	0.78	0.68
↓ $\mathcal{M}_{FPR}$	0.34	0.10	0.41	0.18	0.12	0.05	0.25
↓ $\mathcal{M}_{FNR}$	0.25	0.66	0.32	0.44	0.54	0.67	0.39
↓ $\mathcal{U}_e$	0.0006	0.0004	0.0010	0.0005	0.0002	0.0003	0.0006
↓ $\mathcal{U}_a$	0.2299	0.1578	0.3459	0.1712	0.1027	0.1599	0.2053
↓ $\mathcal{U}_p$	0.2305	0.1583	0.3469	0.1717	0.1029	0.1602	0.2059
Point-based Fairness Measures							
$\mathcal{F}_{SP}$	**2.84**		**2.44**			**0.31**	
$\mathcal{F}_{EOpp}$	**2.19**		**1.48**			**0.54**	
$\mathcal{F}_{EOdd}$	**1.57**		**2.37**			**0.40**	
$\mathcal{F}_{EAcc}$	1.03		0.85			1.13	
Uncertainty-based Fairness Measures (Ours)							
$\mathcal{F}_{Epis}$	**1.55**		**4.35**			**0.50**	
$\mathcal{F}_{Alea}$	**1.46**		**3.36**			**0.78**	
$\mathcal{F}_{Pred}$	**1.46**		**3.37**			**0.78**	

measures directly implying that the model in question is inclined to *unfairly predict* the subgroup age less than 25 as $y = 1$, i.e recidivating an offense.

Furthermore, the proposed uncertainty-based fairness measures also show similar results with $\mathcal{F}_{Epis} = 4.35$, $\mathcal{F}_{Alea} = 3.36$ and $\mathcal{F}_{Pred} = 3.37$. A similar conclusion with the race attribute could be arrived here with the $\mathcal{F}_{Epis} = 4.35$, i.e the *lack of data* according to the model behavior is higher for the subgroup age less than 25 compared to the the subgroup age greater than 45 even though the dataset actually contains less samples for the latter. Even though we also observe a similar pattern with the race attribute with $\mathcal{F}_{Alea} = 3.36$ and $\mathcal{F}_{Pred} = 3.37$, the fairness gap in this case is significantly higher. These measures also show that the classification hardness, the noise faced by the model according to its own behavior, is drastically higher for the subgroup of individuals with an age of less than 25.

Social Impact: From the results above, we see how existing point-based measures merely highlight the prediction bias present. Our proposed uncertainty-based measures go a step beyond by serving as a fairness evaluation tool which points towards the *potential source of bias*: The persistent social inequity across race (Ding et al., 2021), and towards a potential solution: balancing samples across gender. In addition, data collected from a real-world setting is bound to be implicated or corrupted by group-dependent labelling or annotation

noise (Wang, Liu, & Levy, 2021). For instance, it has been demonstrated that labels for criminal activity generated via crowdsourcing are systematically biased against certain subgroups (Dressel & Farid, 2018). This label class and subgroup dependent heterogeneous systematic bias cannot be quantified by point-based fairness measures. However, we hypothesize that this bias can be captured by our measure which illustrates how the model produced higher $\mathcal{F}_{Epis}$, $\mathcal{F}_{Alea}$ and $\mathcal{F}_{Pred}$ for African-Americans and Males despite having more samples for both demographic groups within the training set. Hence, our measure is a useful diagnostic tool in a real-world setting when clean and accurate labels are not readily available. Future experiments may focus on verifying how unbiased labels may impact the point-based and uncertainty-based fairness measures.

Table 4: Label and sensitive attribute distributions of COMPAS and Adult. B/W: Black/White. F/M: Female/Male.

	COMPAS		**Adult**	
Group	$Y=0$	$Y=1$	$Y=0$	$Y=1$
B	1514 (48%)	1661 (52%)	2451 (87%)	366 (13%)
W	1281 (61%)	822 (39%)	19094 (74%)	6839 (26%)
F	762 (65%)	413 (35%)	8670 (89%)	1112 (11%)
M	2601 (52%)	2396 (48%)	13984 (69%)	6396 (31%)

6.2.2 The Adult Dataset

As with COMPAS, Black and Female are the minority groups (G_0) across race and gender respectively. As listed in Table 4, Adult has severe imbalance across labels and groups. Across **race**, Table 5 shows we only have 2817 samples for African-Americans vs. 25933 for Whites. However, according to the performance and point-based fairness measures, the fairness gap between the African-Americans and Whites is lower compared to that in COMPAS. The uncertainty-based fairness measures provide some interesting insights. Particularly, we observe a surprisingly large fairness gap in terms of epistemic uncertainty, $\mathcal{F}_{Epis} = 151$. This is not surprising since Whites have $10\times$ more samples, yielding very small $\mathcal{U}_e$ value. Aleatoric uncertainty $\mathcal{U}_a$ values for both groups are very small (compared to all other datasets), which suggest that the classifier has more certainty with respect to data noise, yielding $\mathcal{F}_{Alea} \sim 1.00$.

The fairness gap also seems lower across **gender** in Adult compared to COMPAS. There is also class imbalance across gender, with only 9872 samples for females vs. 20380 for males. We observe conflicting outcomes across the point-based fairness measures: $\mathcal{F}_{Opp} = 1.04$ and $\mathcal{F}_{EAcc} = 1.18$ point to *fair* classification whereas $\mathcal{F}_{SP} = 0.62$ and $\mathcal{F}_{Odd} = 0.79$ suggest otherwise. Similar to the race attribute, $\mathcal{F}_{SP} = 0.62$ implies higher salary classification bias in favour of Males. As for epistemic and aleatoric uncertainties, we observe gaps similar to the race attribute: There is significant bias in terms of $\mathcal{F}_{Epis}$ (against Males), despite the dataset containing more Male samples. Moreover, the model appears to have the same level of aleatoric certainty across gender ($\mathcal{F}_{Alea} \sim 1.00$).

Table 5: Experiment 2: The analysis with Adult. Unfair values ($|\mathcal{F} - 1| > 0.2$, following (Zanna et al., 2022)) are **highlighted**. B/W: Black/White. F/M: Female/Male.

Measure	**B** (G_0)	**W** (G_1)	**F** (G_0)	**M** (G_1)
↑ Sample Size	2,817	25,933	9,872	20,380
Performance Measures				
↑ $\mathcal{M}_{Acc}$	0.86	0.77	0.87	0.73
↑ $\mathcal{M}_{PPV}$	0.40	0.60	0.39	0.64
↑ $\mathcal{M}_{NPV}$	0.91	0.79	0.91	0.75
↓ $\mathcal{M}_{FPR}$	0.07	0.07	0.06	0.08
↓ $\mathcal{M}_{FNR}$	0.67	0.69	0.68	0.69
↓ $\mathcal{U}_e$	0.0001	6e-7*	0.0001	6e-8*
↓ $\mathcal{U}_a$	0.01	0.01	0.01	0.01
↓ $\mathcal{U}_p$	0.0007	0.0004	0.0008	0.0003
Point-based Fairness Measures				
$\mathcal{F}_{SP}$	**0.75**		**0.62**	
$\mathcal{F}_{Opp}$	1.08		1.04	
$\mathcal{F}_{Odd}$	0.87		**0.79**	
$\mathcal{F}_{EAcc}$	1.12		1.18	
Uncertainty-based Fairness Measures (Ours)				
$\mathcal{F}_{Epis}$	**151***		**521***	
$\mathcal{F}_{Alea}$	1.00		1.00	
$\mathcal{F}_{Pred}$	**1.49**		**2.65**	

Social Impact: The point-based measures seem to indicate that the outcome is acceptably fair which is non-indicative of the underlying problem, i.e., the model is still unsure of its prediction of the majority class despite having more samples on them. Hypothetically, this could lead to prediction bias when encountering *real-world* issues such as missing data (Goel et al., 2021) and distributional shifts (Chen et al., 2022). Future experiments can be conducted to verify how such real-world challenges, e.g., missing data and distributional shifts, may impact the uncertainty-based fairness measures. Our uncertainty-based fairness measures managed to highlight this discrepancy across both race and gender which could encourage pre-emptive efforts to further investigate the underlying source of bias in the model before deploying them in real-world settings.

6.2.3 D-Vlog Dataset

D-Vlog Truncation Statistics Table 6 shows that Female videos are truncated significantly, which leads to loss of information and an increase of uncertainty in predictions (Cheong et al., 2023).

As the dataset owners have explored both multi-modal and uni-modal architectures, we analyze D-Vlog both in a multi-modal and in a uni-modal manner. Table 7 provides the experimental results. Both point-based and uncertainty-based fairness measures deem the classifier to be fair (except for $\mathcal{F}_{Odd}$). Uncertainty-based fairness results are especially surprising since the Female group size is twice the size of the Male group. However, we observe that the classifier has high aleatoric uncertainty for both groups.

Table 6: Label, duration and sensitive attribute distributions of D-Vlog. Both average duration and average truncated amount are given in seconds. Absolute value of the entries with negative value in the last row shows the amount of zero padding whereas the positive values directly state the amount of truncation.

	Male		**Female**	
	$Y=0$	$Y=1$	$Y=0$	$Y=1$
# Samples	140 (0%)	182 (0%)	2666 (0%)	373 (0%)
Avg. Duration(s)	483	583	587	667
Avg. Truncation(s)	-158	-13	-9	+71

Table 7: Experiment 2: The analysis with D-Vlog. Unfair values are **highlighted**. F/M: Females/Males. G_0: Females.

	Multi-Modal		Audio Only		Visual Only	
Measure	**F**	**M**	**F**	**M**	**F**	**M**
↑ Sample Size	639	322	639	322	639	322
Performance Measures						
↑ $\mathcal{M}_{Acc}$	0.59	0.73	0.63	0.75	0.63	0.66
↑ $\mathcal{M}_{PPV}$	0.62	0.78	0.72	0.82	0.65	0.69
↑ $\mathcal{M}_{NPV}$	0.54	0.66	0.56	0.66	0.61	0.57
↓ $\mathcal{M}_{FPR}$	0.57	0.38	0.29	0.28	0.54	0.61
↓ $\mathcal{M}_{FNR}$	0.28	0.20	0.43	0.22	0.23	0.18
↓ $\mathcal{U}_e$	0.006	0.006	0.022	0.016	0.034	0.035
↓ $\mathcal{U}_a$	0.45	0.45	0.28	0.22	0.10	0.09
↓ $\mathcal{U}_p$	0.46	0.46	0.31	0.24	0.14	0.13
Point-based Fairness Measures						
$\mathcal{F}_{SP}$	1.01		**0.75**		0.91	
$\mathcal{F}_{Opp}$	0.89		**0.73**		0.94	
$\mathcal{F}_{Odd}$	**1.68**		**1.40**		0.94	
$\mathcal{F}_{EAcc}$	0.81		0.84		0.96	
Uncertainty-based Fairness Measures (Ours)						
$\mathcal{F}_{Epis}$	1.00		**1.38**		0.96	
$\mathcal{F}_{Alea}$	1.00		**1.32**		1.11	
$\mathcal{F}_{Pred}$	1.00		**1.32**		1.07	

The results per modality suggest that the audio modality has strong bias against Females since the performance measures are generally lower for Females. This, however, is not coherently captured by point-based measures whereas our uncertainty-based measures consistently highlight the bias. The cause of this bias appears to be the truncation of the videos by the dataset owners: Recordings of Females are significantly longer and therefore, truncated more. This naturally results in more reduction in information useful for the classification task for females, thus increasing the uncertainty for females. However, this effect

is not observed across the visual modality as the classifier performs poorly across both males and females.

6.3 Experiment 3: Individual Fairness

We now analyze individual fairness with point-based and uncertainty measures for COMPAS. The results in Fig. 4 suggest that $\mathcal{F}_{\hat{y}}^{indv}$ values differ across different groups of race and gender as well as outcomes. This is also evident with the uncertainty-based individual fairness measures ($\mathcal{F}_{\mathcal{U}}^{indv}$). However, although "W(-)" and "B(-)" have similar point-based consistencies, they are very different across both aleatoric and epistemic consistencies. Aleatoric consistencies align with $\mathcal{F}_{\hat{y}}^{indv}$ that the classifier is having difficulty with "B(+)" samples. $\mathcal{F}_{\mathcal{U}_e}^{indv}$ values highlight that "M(+)" and "B(+)" might especially benefit from additional data.

6.3.1 Experiment 3: Individual Fairness Analysis on Adult

With reference to Figure 5, we see that $\mathcal{F}_{\hat{y}}^{indv}$ for both race and gender are largely similar. This is also true for the the Uncertainty-based individual fairness measures $\mathcal{F}_{\mathcal{U}_a}^{indv}$ and $\mathcal{F}_{\mathcal{U}_e}^{indv}$. A noteworthy point is that all measures indicate an interesting insight about the positive classes, ("B+", "W+", "F+" and "M+"). All of them point towards a perfect consistency score of ≈ 1. We hypothesize that this might be due to the severe class imbalance within the Adult dataset where there is a very small subgroup that belongs to the positive class $Y = 1$ thus causing the classifier to memorize and be highly confident about the $\hat{Y} = 1$ predictions. This is also supported by the small (0.07) FPR reported in Table 4 (the main manuscript) for all groups. We see slightly lower consistency for negative classes in point predictions and aleatoric uncertainty. The high FNR rate for both groups (Table 4 in the main text) suggests that there are more errors with $\hat{Y} = 0$ predictions, causing higher inconsistencies for those predictions. Lower consistencies for "F-" and "B-" for epistemic uncertainty suggest more data can be helpful for Female and Black groups, which is supported by the dataset distribution highlighted in Table 3.

Social Impact: Considering uncertainty in individual fairness can be crucial in many applications. For instance, cancer-free prognosis of a patient should take into account uncertainty-based consistency for similar individuals.

6.4 Experiment 4: Ablation Analysis

We analyze the effect of model capacity on performance and uncertainty estimations. The results in Fig. 6 show that the uncertainty estimations are affected by the change in the number of neurons per layer. However, the relative ordering between the different demographic groups do not appear to be affected. Since accuracy appears to saturate after 100 neurons and to lower the computational cost, we have chosen the hidden layer sizes as 100 in all experiments. Adding more layers led to significant over-fitting problems for SD1, SD2, SD3, COMPAS, and Adult datasets. Therefore, we performed the rest of the experiments with a single hidden-layer for COMPAS and no hidden-layer for SD1, SD2, SD3 and Adult.

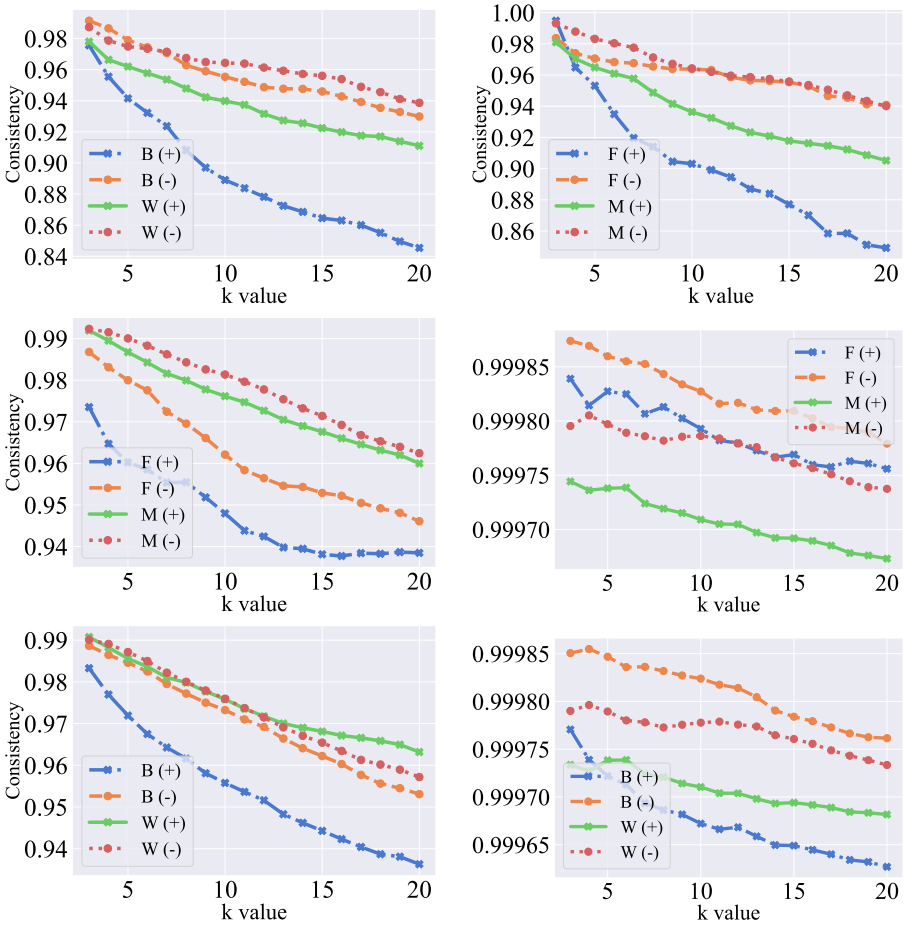


Figure 4: Experiment 3: Point-based **(a,b)** and uncertainty-based individual fairness **(c-f)** scores for COMPAS.

7. Discussion

In this paper, we have argued that existing point-based fairness measures may not be reliable as they depend on point predictions of the ML models and ignore their uncertainties. To address this limitation, we introduce the use of different types of uncertainty as fairness measures. We prove that the proposed fairness measures are independent of point-based fairness measures and empirically show that uncertainty-based fairness measures provide more insights about the presence and the source of bias in predictions.

7.1 Main Insights

In the following, we summarize the main insights:

Insights through the Epistemic Fairness Measure ($\mathcal{F}_{Epis}$) Measuring the fairness gap in terms of epistemic uncertainty, by definition, highlights the lack of data for one group. What is beneficial is that this is not affected by the mere number of samples, which

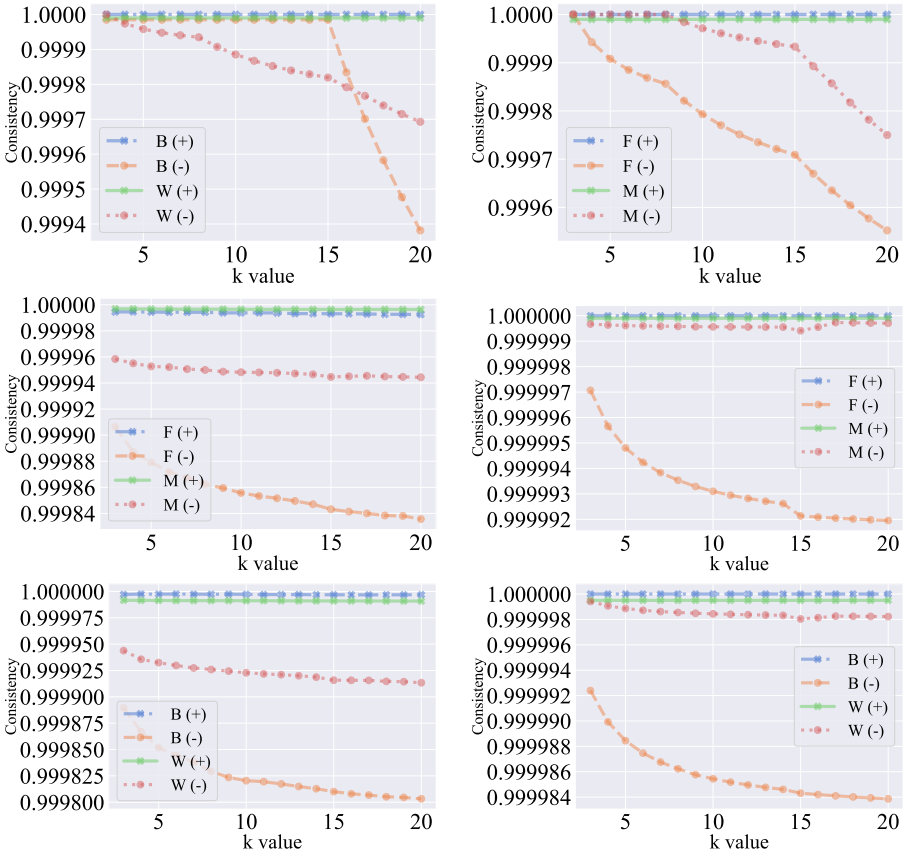
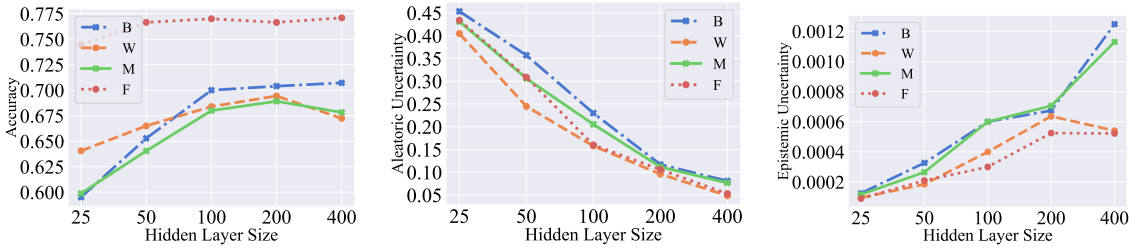


Figure 5: Experiment 3: Point-based **(a,b)** and uncertainty-based individual fairness **(c-f)** scores for Adult.



(a) Accuracy (b) Aleatoric Uncertainty (c) Epistemic Uncertainty

Figure 6: Experiment 4: Ablation analysis on the effect of model capacity on COMPAS. **(a)** Accuracy. **(b)** Aleatoric uncertainty. **(c)** epistemic uncertainty. B/W: Black/White. F/M: Female/Male. As model performance starts saturating at 100 neurons, we have used 100 neurons in BNNs.

can be misleading. For example, in COMPAS, Black and Male groups have significantly

more samples. However, both groups still witness higher $\mathcal{U}_e$ values. This suggests that the dataset may contain data-level bias. The dataset distribution confirms that there is a class-imbalance problem for these groups, which can be remedied with more data.

Insights through the Aleatoric Fairness Measure ($\mathcal{F}_{Alea}$) Aleatoric uncertainty reflects the hardness of a problem owing to label or data noise (e.g., occlusion) (Kendall & Gal, 2017). The use of this informative measure has shown that the classification task is harder for some groups. For example, in D-Vlog, we see that truncating videos has increased aleatoric uncertainty for females.

Another prominent example is that of COMPAS. For instance, despite being a frequently used benchmark for fairness evaluation, an oft-cited key limitation of COMPAS is that errors in typography is a major flaw in this dataset (Rudin, 2019). Uncertainty can, by definition, capture some of the typography and data issues, which would be missed by point-based measures. As evidenced in Table 3, uncertainty-based fairness measures provided some insights about the roots of bias and can be used in conjunction with point-based measures.

7.2 Social Impact

There is a rapid increase in bias mitigation methods for the past years (Hort, Chen, Zhang, Harman, & Sarro, 2023). However, it is unclear which source of bias each method is intended to address. In fact, recent work has demonstrated that if bias is due to missing values, existing bias mitigation methods often reduce (point-based) performance disparities at the cost of accuracy (Wang et al., 2023). Our contribution lies in leveraging existing uncertainty measures to quantify an alternative aspect of fairness. That said, probing a model by adding noise or perturbations to its inputs is useful in analyzing model robustness or increasing model robustness if noise or perturbations are added during training. Epistemic and aleatoric uncertainties, on the other hand, pertain to how well the model captures the lack of data and the absence of noise (ambiguity) respectively. Given a dataset and a model, both types of uncertainties are supposed to be irreducible. In such an instance, using point-based measures will likely be sub-optimal. Our proposed uncertainty-based measure highlights this underlying problem and cautions against foisting a "fair" outcome using point-based fairness measures.

Moreover, many of the existing bias mitigation solutions rest on strict machine learning assumptions such as having access to clean or noise-free labels and requiring the model to be deployed in a fair environment that does not deviate from the training setting (Kang, Li, Weber, Liu, Zhang, & Li, 2022). This is optimistic at best and harmful at worst. This incongruence between theoretical formulation and real-world settings is one of the handicaps that the machiner learning fairness research community needs to overcome. Our work also highlights the need to develop methods which are able to address epistemic and aleaoteric sources of discrimination. We hope that the proposed uncertainty-based fairness measures present a step towards that direction.

7.3 Limitations

Despite its merits, uncertainty-based fairness measures require working with models which provide or can be modified to provide uncertainties. Moreover, quantifying uncertainty is

an active research area, and in this work we have not been able to undertake a thorough evaluation of different uncertainty quantification methods. The above provides opportunities for future work. A key point to note is that the uncertainty-fairness measures in our paper are differentiable and can be converted to a loss function. However, forcing epistemic and aleatoric uncertainties to be similar across groups or individuals will not necessarily change the "real uncertainties" as these measures simply reflect issues inherent in the data and noise (or ambiguity).

Although prediction uncertainty can be helpful in analyzing fairness, this approach has certain limitations which we view as opportunities for future work. For example, uncertainty estimation requires either using models that directly provide multiple predictions (e.g., BNNs, Deep Ensembles) or modifying models (and their training procedure) to do so (e.g., Monte Carlo Dropout (Gal & Ghahramani, 2016)). This hinders the use of state-of-the-art architectures (or their trained versions) in fairness analysis. Moreover, there is also the overhead involved with obtaining multiple predictions to quantify uncertainty. This can be alleviated with one-pass uncertainty estimation approaches, though they tend to be less reliable than the approaches considered in this paper (Abdar et al., 2021).

Quantifying uncertainty in a reliable manner is a challenging and an active research topic (Mukhoti et al., 2023; Liu, Lin, Padhy, Tran, Bedrax-Weiss, & Lakshminarayanan, 2020; van Amersfoort, Smith, Teh, & Gal, 2020a). Although we have obtained similar outcomes with two different methods (BNNs and Deep Ensembles), we have encountered difficulties with the ranges of estimated uncertainties. It would be beneficial to perform our analyses with newer approaches. Another promising research direction is to consider alternative metrics for measuring the dispersion of uncertainty values in a group as taking the average across a group can miss important characteristics of the distribution. Despite the aforementioned limitations, we sincerely hope that our work can provide a stepping stone towards investigating and addressing these challenges.

Acknowledgments

Both first authors, Selim Kuzucu and Jiaee Cheong, contributed equally to this work. This work was undertaken while Jiaee Cheong was a visiting PhD student at the Middle East Technical University (METU). **Open access:** The authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising. **Data access:** This study involved secondary analyses of existing datasets. All datasets are described and cited accordingly. **Funding:** J. Cheong is supported by the Alan Turing Institute Doctoral Studentship, the Cambridge Trust and the Leverhulme Trust, and further acknowledges resource support from METU during her visiting studentship. H. Gunes is supported by the EPSRC/UKRI project ARoEQ under grant ref. EP/R030782/1. We gratefully acknowledge the computational resources provided by METU Center for Robotics and Artificial Intelligence (METU-ROMER) and METU Image Processing Laboratory.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. (2021). A review of un-

certainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, *76*, 243–297.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications.

Baltaci, Z. S., Oksuz, K., Kuzucu, S., Tezoren, K., Konar, B. K., Ozkan, A., Akbas, E., & Kalkan, S. (2023). Class uncertainty: A measure to mitigate class imbalance. In *arXiv preprint arXiv:2311.14090*.

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. *NeurIPS Tutorial*, *1*, 2.

Becker, B., & Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. In *International conference on machine learning*, pp. 1613–1622. PMLR.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, *12*(1), 4209.

Cetinkaya, B., Kalkan, S., & Akbas, E. (2024). Ranked: Addressing imbalance and uncertainty in edge detection using ranking-based losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3239–3249.

Chen, Y., Raab, R., Wang, J., & Liu, Y. (2022). Fairness transferability subject to bounded distribution shift. *Advances in Neural Information Processing Systems*, *35*, 11266–11278.

Chen, Y., & Joo, J. (2021). Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14980–14991.

Cheong, J., Kalkan, S., & Gunes, H. (2021). The hitchhiker's guide to bias and fairness in facial affective signal processing: Overview and techniques. *IEEE Signal Processing Magazine*, *38*(6), 39–49.

Cheong, J., Kalkan, S., & Gunes, H. (2022). Counterfactual fairness for facial expression recognition. In *European Conference on Computer Vision*, pp. 245–261. Springer.

Cheong, J., Kalkan, S., & Gunes, H. (2023). Causal structure learning of bias for fair affect recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 340–349.

Cheong, J., Kalkan, S., & Gunes, H. (2024). Fairrefuse: Referee-guided fusion for multimodal causal fairness in depression detection. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Cheong, J., Kuzucu, S., Kalkan, S., & Gunes, H. (2023). Towards gender fairness for mental health prediction. In *32nd Int. Joint Conf. on Artificial Intelligence (IJCAI)*.

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

Ding, F., Hardt, M., Miller, J., & Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, *34*, 6478–6490.

Domnich, A., & Anbarjafari, G. (2021). Responsible ai: Gender bias assessment in emotion recognition..

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1), eaao5580.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226.

Ethayarajh, K. (2020). Is your classifier actually biased? measuring fairness under uncertainty with bernstein bounds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2914–2919.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268.

Gal, Y., et al. (2016). *Uncertainty in deep learning*. Ph.D. thesis, University of Cambridge.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning..

Garg, P., Villasenor, J., & Foggo, V. (2020). Fairness metrics: A comparative analysis. In *IEEE International Conference on Big Data (Big Data)*, pp. 3662–3666. IEEE.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. (2021). A survey of uncertainty in deep neural networks..

Goel, N., Amayuelas, A., Deshpande, A., & Sharma, A. (2021). The importance of modeling data missingness in algorithmic fairness: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7564–7573.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR.

Han, M., Canli, I., Shah, J., Zhang, X., Dino, I. G., & Kalkan, S. (2024). Perspectives of machine learning and natural language processing on characterizing positive energy districts. *Buildings*, *14*(2), 371.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

Havasi, M., Jenatton, R., Fort, S., Liu, J. Z., Snoek, J., Lakshminarayanan, B., Dai, A. M., & Tran, D. (2020). Training independent subnetworks for robust prediction. In *International Conference on Learning Representations*.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2023). Bias mitigation for machine learning classifiers: A comprehensive survey. In *ACM J. Responsib. Comput.*, New York, NY, USA. Association for Computing Machinery.

Jiang, H., & Nachum, O. (2020). Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR.

Kaiser, P., Kern, C., & Rügamer, D. (2022). Uncertainty-aware predictive modeling for fair data-driven decisions..

Kang, M., Li, L., Weber, M., Liu, Y., Zhang, C., & Li, B. (2022). Certifying some distributional fairness with subpopulation decomposition. *Advances in Neural Information Processing Systems*, *35*, 31045–31058.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision?. *CoRR*, *abs/1703.04977*.

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization..

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, *30*.

Kwon, Y., Won, J.-H., Kim, B. J., & Paik, M. C. (2020). Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, *142*, 106816.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles..

Liu, J. Z., Lin, Z., Padhy, S., Tran, D., Bedrax-Weiss, T., & Lakshminarayanan, B. (2020). Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *CoRR*, *abs/2006.10108*.

MacKay, D. J. (1992). A practical bayesian framework for backpropagation networks. *Neural computation*, *4*(3), 448–472.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, *54*(6), 1–35.

Mehta, R., Shui, C., & Arbel, T. (2023). Evaluating the fairness of deep learning uncertainty estimates in medical image analysis..

Mukherjee, D., Yurochkin, M., Banerjee, M., & Sun, Y. (2020). Two simple ways to learn individual fairness metrics from data. In *International Conference on Machine Learning*, pp. 7097–7107. PMLR.

Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H., & Gal, Y. (2023). Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24384–24394.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., & Dokania, P. (2020). Calibrating deep neural networks using focal loss. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., & Lin, H. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 33, pp. 15288–15299. Curran Associates, Inc.

Naik, L., Kalkan, S., & Kruger, N. (2024). Pre-grasp approaching on mobile robots: A pre-active layered approach. *IEEE Robotics and Automation Letters*, *9*(3).

Neal, R. M. (1995). *Bayesian Learning for Neural Networks*. Ph.D. thesis, University of Toronto.

Roy, A., & Mohapatra, P. (2023). Fairness uncertainty quantification: How certain are you that the model is fair?..

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206–215.

Shridhar, K., Laumann, F., & Liwicki, M. (2019). A comprehensive guide to bayesian convolutional neural network with variational inference. *CoRR*, *abs/1901.02731*.

Tahir, A., Cheng, L., & Liu, H. (2023). Fairness through aleatoric uncertainty..

van Amersfoort, J., Smith, L., Teh, Y. W., & Gal, Y. (2020a). Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network. *CoRR*, *abs/2003.02037*.

Van Amersfoort, J., Smith, L., Teh, Y. W., & Gal, Y. (2020b). Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pp. 9690–9700. PMLR.

Verma, S., & Rubin, J. (2018a). Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pp. 1–7.

Verma, S., & Rubin, J. (2018b). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, p. 1–7, New York, NY, USA. Association for Computing Machinery.

Wang, H., He, L., Gao, R., & Calmon, F. (2023). Aleatoric and epistemic discrimination: Fundamental limits of fairness interventions. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wang, J., Liu, Y., & Levy, C. (2021). Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 526–536.

Xu, T., White, J., Kalkan, S., & Gunes, H. (2020). Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pp. 506–523. Springer.

Yoon, J., Kang, C., Kim, S., & Han, J. (2022). D-vlog: Multimodal vlog dataset for depression detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), 12226–12234.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180.

Zanna, K., Sridhar, K., Yu, H., & Sano, A. (2022). Bias reducing multitask learning on mental health prediction..

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR.