EUROPEAN JOURNAL OF EMERGING ARTIFICIAL INTELLIGENCE

QUANTIFYING ALGORITHMIC FAIRNESS: A NOVEL PERSPECTIVE THROUGH UNCERTAINTY ESTIMATION

Dr. Clara Moreau
Department of Computer Science, École Normale Supérieure, France

Prof. Julia Weber
Department of Computer Science, École Normale Supérieure, France

V0LUME01 ISSUE01 (2024)
Published Date: 30 December 2024 // Page no.: - 96-112

ABSTRACT

The increasing deployment of machine learning (ML) systems in high-stakes domains necessitates robust fairness evaluation. Traditional fairness metrics primarily focus on statistical disparities in outcomes, often overlooking the model's confidence in its predictions, particularly for sensitive subgroups. This article proposes a novel framework for assessing algorithmic fairness by integrating uncertainty quantification (UQ) into the evaluation process. We delineate between aleatoric (data-inherent) and epistemic (model-inherent) uncertainties and explore various UQ techniques, including Bayesian Neural Networks, Monte Carlo Dropout, Deep Ensembles, and Deep Deterministic Uncertainty. We argue that disparate levels of uncertainty across demographic groups can serve as a powerful diagnostic tool, indicating issues such as data scarcity, representational bias, or inherent ambiguities within specific populations. By leveraging uncertainty as a fairness measure, we can identify subtle forms of discrimination, enhance model transparency, and enable more proactive and targeted bias mitigation strategies. This approach promises to yield more robust, trustworthy, and equitable ML systems.

Keywords: Algorithmic fairness, Uncertainty quantification, Deep learning, Bias detection, Epistemic uncertainty, Aleatoric uncertainty, Responsible AI, Data imbalance, Model calibration.

INTRODUCTION

The pervasive integration of machine learning (ML) systems into critical societal domains, such as healthcare, finance, employment, and criminal justice, has brought to the forefront urgent concerns regarding their fairness and potential for discriminatory outcomes [2, 17, 20]. As ML models increasingly influence decisions with significant real-world consequences, ensuring their equitable performance across diverse populations is paramount. While these models demonstrate impressive predictive capabilities, their often opaque "black box" nature can obscure inherent biases present in the training data or inadvertently introduced during model development, ultimately leading to disparate and unjust impacts on different demographic groups [54]. This challenge has spurred extensive research into defining and measuring fairness in ML, leading to a rich landscape of various notions and metrics designed to capture different facets of non-discrimination [8, 26, 59, 60].

Historically, the ML community has predominantly addressed fairness by focusing on statistical disparities in prediction outcomes or error rates across protected attributes, such as race, gender, or age [4, 21, 31, 65, 67]. Measures like demographic parity, equal opportunity, and equalized odds compare metrics like positive prediction rates or false negative rates between different groups. While these point-based fairness measures are crucial and have significantly advanced the field, they often provide a limited, static view of a model's performance. They quantify fairness based on aggregate prediction counts, neglecting the underlying confidence or reliability of individual predictions. This oversight can render them susceptible to real-world complexities such as noise in data, missing information, or shifts in data distribution, potentially masking subtle yet significant biases [28, 34, 10]. A model might, for instance, appear statistically "fair" by these metrics, yet consistently exhibit lower confidence in its predictions for a minority group compared to a majority group, indicating a deeper, unaddressed bias related to data representation or model knowledge [53].

This paper posits that a more profound and robust understanding of algorithmic fairness can be achieved by integrating Uncertainty Quantification (UQ) into the fairness assessment framework. UQ in deep learning is a rapidly advancing field dedicated to estimating the confidence or reliability of a model's predictions [1, 24, 27]. It offers methods to disentangle the overall predictive

uncertainty into two distinct components: aleatoric uncertainty, which stems from inherent, irreducible noise and variability in the data itself (e.g., sensor noise, ambiguous labels), and epistemic uncertainty, which reflects the model's lack of knowledge or confidence in its own parameters, often arising from insufficient training data or out-of-distribution inputs [38]. While UQ has traditionally been leveraged to enhance model robustness, reliability, and safety in high-stakes applications where knowing "what the model doesn't know" is critical [42], its potential as a lens through which to evaluate and promote algorithmic fairness remains largely underexplored.

Recent pioneering work has begun to bridge this conceptual gap, suggesting that uncertainty itself can serve as a powerful indicator of unfairness [35, 53, 56]. For example, if a machine learning model consistently exhibits higher predictive uncertainty for individuals belonging to a specific demographic group, even if its point-based accuracy for that group seems comparable to others, this disparity in confidence signals a potential underlying bias. Such a discrepancy could arise from issues like data scarcity for the underserved group, leading to increased epistemic uncertainty, or inherent ambiguities in their data, contributing to higher aleatoric uncertainty [28, 62]. This article aims to develop a comprehensive conceptual framework for integrating uncertainty quantification into the assessment of algorithmic fairness. We argue that by delving into the nuanced aspects of a model's confidence, we can gain a more insightful, comprehensive, and ultimately more equitable measure of its performance, thereby paving the way for more responsible and robust ML systems.

METHODS

This section lays out the foundational concepts and methodologies for quantifying uncertainty in machine learning models and, crucially, for integrating these uncertainty measures into the assessment of algorithmic fairness. We begin by formally defining the two primary types of uncertainty and discussing prominent techniques for their estimation. Subsequently, we introduce our novel uncertainty-based fairness measures, detailing how they complement and extend traditional point-based metrics.

Types of Uncertainty and Their Estimation

Understanding and quantifying uncertainty is paramount for reliable and fair machine learning systems. In deep learning, predictive uncertainty can be decomposed into two fundamental types [1, 27, 38]:

1. Aleatoric Uncertainty (Uα): This form of uncertainty arises from the inherent, irreducible stochasticity or noise present in the data generation process itself. It reflects fundamental ambiguity that cannot be mitigated by simply collecting more data or improving the model. Examples include measurement noise from sensors, inconsistent labeling by annotators, or intrinsic variability in the observed phenomenon. In classification tasks, aleatoric uncertainty is often modeled by having the network predict the parameters of a conditional probability distribution over the outputs, such as the variance for regression problems or a categorical distribution for classification, allowing the model to express its inherent ambiguity for a given input [44]. This uncertainty is captured by the expected data likelihood given a model.

2. Epistemic Uncertainty (Ue): This type of uncertainty reflects the model's own ignorance or lack of knowledge about its parameters. It stems from the limited amount of data available for training, especially in regions of the input space that are sparsely populated or completely unseen. Epistemic uncertainty is reducible; it can theoretically be decreased by acquiring more diverse and representative training data, or by improving the model's capacity to learn complex patterns. This uncertainty is particularly critical for identifying out-of-distribution inputs or highlighting areas where the model is extrapolating rather than interpolating [1, 38]. It is often captured by quantifying the variance in predictions across an ensemble of models or different parameterizations of the same model.

Techniques for Uncertainty Quantification (UQ) in Deep Neural Networks

Quantifying these uncertainties in deep neural networks is an active area of research, with several prominent techniques emerging [1, 27]:

- Bayesian Neural Networks (BNNs): At their core, BNNs fundamentally redefine the parameters of a neural network as probability distributions rather than fixed point values [45, 52]. Instead of learning a single weight for each connection, BNNs learn a distribution (e.g., a Gaussian) over each weight ($\omega i=(\mu i,\sigma i)$). This allows the network to output a distribution over predictions, naturally capturing epistemic uncertainty. The predictive distribution is obtained by integrating over all possible model parameters, weighted by their posterior probability. While theoretically elegant and robust against overfitting [6, 24], performing exact inference (i.e., computing the true posterior distribution over weights) in BNNs is computationally intractable for complex architectures. This intractability necessitates approximation methods like variational inference or Monte Carlo (MC) sampling [6, 25, 55].

- Monte Carlo Dropout (MC Dropout): This method offers a computationally efficient approximation to Bayesian inference in deep learning models [25]. By applying dropout (a regularization technique where neurons are randomly dropped during training) not only during training but also during inference, MC Dropout enables the approximation of sampling from the posterior distribution of a Bayesian neural network. Running multiple forward passes (M forward passes) with dropout enabled at test time yields a distribution of predictions for

a given input. The variance or entropy of these M predictions can then be used to estimate epistemic uncertainty. This method is popular due to its simplicity and ability to provide a practical estimate of model uncertainty without significant architectural changes [25].

- Deep Ensembles: This is a surprisingly effective and often state-of-the-art approach for uncertainty estimation [32, 43]. It involves training multiple independent neural networks (an ensemble) with identical architectures but different random initializations and potentially different training data orderings. Since each network converges to a slightly different local minimum in the loss landscape, their individual predictions for the same input will vary. The disagreement (e.g., variance or standard deviation) among the ensemble's predictions serves as a robust measure of epistemic uncertainty. Deep Ensembles are particularly good at capturing epistemic uncertainty, as they directly model the uncertainty across different plausible models [43].

- Deep Deterministic Uncertainty (DDU): More recent advancements have focused on obtaining uncertainty estimates from a single, deterministic deep learning model, aiming to reduce the computational overhead associated with Bayesian methods or ensembles [44, 49, 57, 58]. These methods often incorporate architectural modifications or specific loss functions that encourage the model to be "aware" of its distance to the training data. For instance, some DDU approaches leverage distances in latent space to infer uncertainty, allowing for valuable uncertainty signals without the need for multiple forward passes or complex Bayesian approximations. They offer a promising direction for scalable uncertainty quantification.

- Calibration Techniques: While not directly quantifying uncertainty components, calibration is a crucial prerequisite for reliable uncertainty estimation [29, 50]. A well-calibrated model's predicted probabilities should accurately reflect the true probabilities. For instance, if a model predicts a class with 80% probability, then that class should indeed be the true class for approximately 80% of samples for which it made such a prediction. Miscalibration, where a model is consistently overconfident or underconfident, can mask true uncertainties and lead to misleading fairness assessments. Techniques like Platt scaling or temperature scaling are often applied post-training to improve a model's calibration [29].

Preliminaries and Background for Fairness

Before delving into our uncertainty-based fairness measures, it is essential to establish the foundational notation and traditional definitions of fairness used in the machine learning literature. Following common conventions [8, 59, 60], we consider a binary classification problem operating on a dataset D. This dataset comprises feature vectors X, a binary classification target $Y \in \{0,1\}$, and a sensitive group attribute $G \in \{0,1\}$. In this setup, G=0 typically denotes a minority or historically disadvantaged group, while G=1 represents the majority group.

The objective of a classification task is to learn a mapping $Y=f(X;\theta) \in \{0,1\}$, where θ represents the model's parameters. We denote $P(Y=y_i|X=x_i)$ as the predicted probability for the correct class y_i for a given sample x_i. The final predicted class is $\hat{Y}=\hat{y}_i \leftarrow \text{argmax}_c P(Y=c|X=x_i)$.

MEASURING GROUP FAIRNESS

Group fairness focuses on ensuring that a model's performance or predictions are similar across different demographic groups. Formally, for a predictor $\hat{Y}=f(\cdot;\theta)$ to be considered fair with respect to a demographic group attribute G, a chosen performance measure M (e.g., true positive rate) should satisfy the following equality across groups [26, 59]:

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

Different choices for M lead to distinct notions of group fairness:

- Statistical Parity (or Demographic Parity): This is one of the most fundamental group fairness definitions [21, 26, 59]. It requires that the proportion of positive outcomes (e.g., being granted a loan, being hired) be equal across all demographic groups, regardless of the individual's true label. Mathematically, it compares the model's predicted probabilities for the positive class ($\hat{Y}=1$) across different groups:

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

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

- Equal Opportunity: This definition focuses on ensuring that individuals who truly belong to the positive class (e.g., truly qualified candidates, patients who will truly benefit from a treatment) have an equal chance of being correctly classified into that class, irrespective of their group affiliation [31]. It often compares false negative rates across groups. Specifically, it compares the model's prediction probabilities for the negative class ($\hat{Y}=0$) for the known positive class (Y=1):

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

In this case, $M(D,f,G) \equiv P(\hat{Y}=0|Y=1,G)$. Equivalently, it can be framed as ensuring equal true positive rates (TPRs) across groups.

- Equalized Odds: This is a stronger notion of fairness than Equal Opportunity. It requires that the model's true positive rates and false positive rates are equal across all demographic groups [31]. In other words, the model's predictions should be independent of the sensitive attribute, conditional on the true outcome. This means comparing the model's prediction probabilities for the positive class ($\hat{Y}=1$) for different ground truth classes (Y=1 and Y=0):

$$P(Y^=1|Y=y,G=0)=P(Y^=1|Y=y,G=1) \text{ where } y\in\{0,1\} \quad (4)$$

Here, $M(D,f,G)\equiv P(Y^=1|Y=y,G)$. This definition implies both equal true positive rates and equal false positive rates.

Measuring Individual Fairness

Beyond group-level disparities, fairness can also be assessed at the individual level. Individual fairness, as conceptualized by Dwork et al. [21], posits that "similar individuals should have similar predictions." This principle requires the existence of suitable distance metrics in both the input space ($dx(\cdot,\cdot)$) and the output space ($dy(\cdot,\cdot)$) to quantify similarity. Formally:

$$dy(f(x1),f(x2))\leq Ldx(x1,x2),\forall x1,x2\in X \quad (5)$$

Here, L is a Lipschitz constant, ensuring that small changes in input (similarity) correspond to small changes in output (similar predictions). In practice, individual fairness is often measured using consistency metrics. For point predictions, a common consistency measure involves comparing an individual's prediction to the predictions of their k-nearest neighbors [48, 67]:

$$Fy^indv(X=xi)=1-|y^i-k1xj\in kNN(xi)\sum y^j| \quad (6)$$

where kNN(xi) denotes the k-nearest neighbors of xi in the feature space. This metric quantifies how consistent an individual's prediction is with those of similar individuals, with a score of 1 indicating perfect consistency.

Quantifying Uncertainty for Predictive Models

As discussed in Section 2.1, machine learning models, especially deep neural networks, are often prone to being under- or over-confident in their predictions, and may be unaware of distribution shifts, adversarial attacks, or noise in the data [1, 9, 27]. Quantifying the variance of a model's predictions, commonly referred to as predictive uncertainty, is crucial for developing an awareness of such shortcomings related to the underlying data.

Predictive uncertainty (Up) can be decomposed into its two distinct components: epistemic uncertainty (Ue) and aleatoric uncertainty (Ua). We primarily employ Bayesian Neural Networks (BNNs) for obtaining these uncertainty estimates, as BNNs are known for providing reliable uncertainty quantifications [6]. A BNN defines a distribution over each weight in the network, represented by a mean (μi) and a standard deviation (σi) for each weight ωi. This probabilistic representation allows for sampling different weight configurations, enabling multiple predictions for the same input.

The predictive uncertainty for a given sample x with true label y can be formally quantified as described by Kwon et al. [42] and Shridhar et al. [55]:

$$\text{Up=Epistemic Uncertainty (Ue)} M1m=1\sum M(Pm-P)T(Pm-P)+\text{Aleatoric Uncertainty (Ua)} M1m=1\sum Mdiag(Pm)-PmTPm \quad (7)$$

where:

- M is the number of Monte Carlo samples (i.e., forward passes with different weight samples from the BNN's learned distributions).
- $Pm=P(Y|X=x)$ is the predicted probability distribution for the m-th Monte Carlo sample.
- $P=M1\sum m=1MPm$ represents the average predicted probability distribution across all Monte Carlo samples.

The first term, Ue, captures the variance of predictions across different model configurations (Monte Carlo samples), directly reflecting the model's uncertainty about its own parameters given the data. The second term, Ua, captures the average variance within each individual prediction, representing the inherent noise or ambiguity in the data itself. To obtain group-wise uncertainty estimations, the quantified uncertainty values for all samples belonging to a specific group are aggregated, typically by averaging.

Uncertainty-based Fairness Measures

Building upon the established notions of uncertainty and traditional fairness definitions, we now introduce our novel fairness concepts that leverage prediction uncertainties. Our core premise is that a fair model should not only yield statistically equitable point predictions but also exhibit similar levels of confidence and knowledge (or lack thereof) across different sensitive groups.

Definition 4.1 (Uncertainty-Fairness Measure)

A model is considered fair if its predicted uncertainties are consistent across different demographic groups. More formally, extending the group fairness definition in Equation (1) from Section 3.2, we propose:

$$Fair(f;U,D)\equiv U(D,f,G=0)=U(D,f,G=1) \quad (8)$$

where U can represent any aggregate measure of uncertainty, such as the average predictive uncertainty (Up), average epistemic uncertainty (Ue), or average aleatoric uncertainty (Ua) for a given group, as introduced in Section 4.1. This definition allows for a nuanced assessment of fairness, considering whether the model is equally uncertain (or certain) for different population segments.

Proposition 4.1 (Independence of Uncertainty Fairness)

A crucial aspect of our proposed uncertainty-based fairness measures is their complementarity to conventional point-based fairness measures. We formally state that uncertainty fairness is independent of point-measure based fairness. Consider a predictor $f(\cdot;\theta)$ that generates point-predictions $\{y^i\}i$ (with associated probabilities $\{P(y^i|xi)\}i$) and corresponding uncertainties $\{Ui\}i$ (namely, predictive, epistemic, and aleatoric). Then, the uncertainty fairness measure $Fair(f;U,D)$ is independent of the conventional point-measure based fairness $Fair(f;M,D)$. More formally:

- ¬(Fair(f;M,D)⟹Fair(f;U,D))
- ¬(Fair(f;U,D)⟹Fair(f;M,D))

This proposition implies that achieving fairness by traditional point-based metrics does not guarantee fairness in terms of uncertainty, and vice versa. This highlights the necessity of considering both aspects for a comprehensive fairness assessment.

Proof of ¬(Fair(f;M,D)⟹Fair(f;U,D)):

We proceed with a proof by contradiction. Assume that the implication is true, i.e., Fair(f;M,D)⟹Fair(f;U,D). This would mean that there cannot exist a predictor f that is M-wise fair (point-based fair) but U-wise unfair (uncertainty-based unfair). However, as demonstrated empirically by Synthetic Dataset 1 and Synthetic Dataset 2 (discussed in Section 5.1(A, B) and illustrated in Figure 3, with results in Table 1), we can construct a scenario where a Bayesian Neural Network (BNN) classifier achieves fairness according to standard point-based measures (e.g., Statistical Parity, Equal Opportunity, Equal Accuracy) but exhibits significant unfairness when evaluated using our uncertainty-based measures (e.g., high disparities in aleatoric uncertainty for SD1 and epistemic uncertainty for SD2). This empirical counterexample contradicts our initial assumption, thus proving that Fair(f;M,D)⟹Fair(f;U,D).

Proof of ¬(Fair(f;U,D)⟹Fair(f;M,D)):

Similarly, we assume for contradiction that Fair(f;U,D)⟹Fair(f;M,D). This implies that there should be no predictor f that is U-wise fair but M-wise unfair. Yet, Synthetic Dataset 3 (discussed in Section 5.1(C) and depicted in Figure 3(c), with results in Table 1) provides a direct counterexample. In this case, a BNN classifier demonstrates fairness in terms of both epistemic and aleatoric uncertainties, while simultaneously exhibiting significant unfairness according to point-based metrics (e.g., Equalized Odds, Equal Accuracy). This empirical finding refutes our assumption, thereby establishing that Fair(f;U,D)⟹Fair(f;M,D).

These proofs underscore that uncertainty-based fairness measures provide a distinct and complementary perspective to traditional point-based metrics, necessitating their combined consideration for a holistic evaluation of algorithmic fairness.

Uncertainty-based Individual Fairness

Extending the individual fairness principle of "similar individuals should have similar predictions" (Equation 5) to incorporate uncertainty, we propose that "similar individuals should also have similar prediction uncertainties." This means that not only should their predicted labels be consistent, but the model's confidence or lack thereof for similar individuals should also be comparable.

We define an uncertainty-based individual fairness measure as follows:

FUindv(X=xi)=1−|Ui−k1xj∈kNN(xi)∑Uj|(9)

where Ui is the uncertainty (predictive, epistemic, or aleatoric) for sample xi, and the sum is over the uncertainties of its k-nearest neighbors (kNN(xi)). This measure quantifies the consistency of uncertainty levels among similar individuals. A score closer to 1 indicates higher uncertainty consistency, implying that similar individuals are treated with similar levels of confidence (or doubt) by the model. These individual uncertainty fairness scores can then be aggregated (e.g., by averaging) over specific groups for group-level insights into individual fairness.

Experiments

This section provides a detailed account of the experimental setup, including the characteristics of the datasets utilized, the implementation and training specifics of our models, and the evaluation measures employed to assess both traditional point-based fairness and our proposed uncertainty-based fairness metrics.

Datasets

To thoroughly evaluate our uncertainty-based fairness measures, we employed a diverse set of synthetic and real-world datasets. The synthetic datasets were meticulously crafted to demonstrate specific scenarios where point-based and uncertainty-based fairness measures yield complementary or contradictory insights, while the real-world datasets allowed for validation in complex, practical contexts.

Synthetic Datasets

We created three distinct synthetic datasets, each designed to highlight a particular aspect of fairness in relation to uncertainty. The approach for curating these datasets largely follows the methodology outlined in Zafar et al. [65]. Each synthetic dataset consists of 320 samples, with 20% reserved for testing. The feature space for these datasets is two-dimensional, allowing for clear visual representation of the data distributions and model behavior.

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

This dataset is specifically designed to illustrate a scenario where a classifier may appear fair according to conventional point-based performance metrics but is inherently unfair in terms of aleatoric uncertainties. This occurs due to differing levels of inherent noise or ambiguity in the data for different groups. We generated 100 samples from each of four multivariate distributions, corresponding to each attribute-label pair (G∈{0,1},Y∈{0,1}):

P(X|G=0,Y=0)=Beta(α=[0.5,0.5],β=[0.5,0.5])(10)P(X|G=0,Y=1)=−Beta(α=[0.5,0.5],β=[0.5,0.5])(11)P(X|G=1,Y=0)=N([−7,−7],[15,10;10,15])(12)P(X|G=1,Y=1)=N([7,7],[15,10;10,15])(13)

The Beta distributions for G=0 introduce inherent variability (noise) that Gaussian distributions for G=1 do not, leading to disparate aleatoric uncertainty even if point predictions are balanced. This dataset is visually represented in Figure 3(a).

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

SD2 is constructed to demonstrate a situation where a classifier might be fair in terms of its point predictions but exhibit unfairness due to disparities in epistemic uncertainties. This can happen when the model has less knowledge or has encountered fewer diverse examples for a particular sensitive group, even if the overall class distributions are balanced. We generated 100 samples from each of four multivariate distributions:

P(X|G=0,Y=0)=N([−10,−10],[100,30;30,100])(14)P(X|G=0,Y=1)=N([10,10],[100,30;30,100])(15)P(X|G=1,Y=0)=N([−7,−7],[5,1;5,1])(16)P(X|G=1,Y=1)=N([7,7],[5,1;1,5])(17)

The distributions for G=1 are significantly tighter, implying less spread and potentially less data density, which can lead to higher epistemic uncertainty for the model when dealing with samples from these regions. This dataset is visually represented in Figure 3(b).

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

This dataset is designed as a crucial counterexample, showcasing a scenario where a classifier can be fair according to our proposed uncertainty-based measures, yet demonstrably unfair in terms of traditional point-based fairness metrics. This highlights the independence of the two fairness notions. We obtained 100 samples from each of four multivariate distributions:

P(X|G=0,Y=0)=N([−2,−2],[7,3;3,7])(18)P(X|G=0,Y=1)=N([2,2],[7,3;3,7])(19)P(X|G=1,Y=0)=N([−3,−3],[5,3;5,3])(20)P(X|G=1,Y=1)=N([3,3],[5,3;3,5])(21)

The distributions are chosen such that the overall uncertainty characteristics might be balanced across groups, but the decision boundary derived by a model might still lead to statistical disparities in predictions. This dataset is visually represented in Figure 3(c).

Real-world Datasets

To validate our framework in more practical and complex settings, we utilized three widely recognized real-world datasets commonly employed in fairness research:

- 5.1.4 (D) COMPAS Recidivism Dataset:

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset contains records of criminal offenders and is frequently used to predict recidivism (re-offense) as a binary classification task [2]. The dataset comprises 6172 samples with 14 features. Following the methodology of Zafar et al. [65], we used a specific subset of attributes and adhered to standard dataset splits. The positive label (Y=1) signifies that an individual has recidivated, while Y=0 indicates no recidivism. We evaluated fairness with respect to race, gender, and age. For race, we focused on the fairness gap between Black (minority group, G=0) and White (majority group, G=1) subgroups, as this disparity has been extensively documented. For gender, females were designated as the minority group (G=0) due to observed class imbalance. For age, individuals younger than 25 were considered the minority group (G=0), and those older than 45 were the majority group (G=1), excluding ages between 25 and 45 to maintain a binary group setting for simplified analysis. The distribution of labels and sensitive attributes for COMPAS is detailed in Table 4.

- 5.1.5 (E) Adult Income Dataset:

The Adult Income dataset, sourced from the UCI Machine Learning Repository [5, 18], contains over 48,000 samples with 14 features. The task is to predict whether an individual's annual income exceeds $50K (Y=1) or not (Y=0). We excluded samples with missing entries, resulting in approximately 45,000 samples. We maintained the original training-testing split provided by the authors. Similar to COMPAS, we limited our fairness analysis to race, gender, and age for consistency, though a deeper exploration could involve other variables like marital status or education level. The class and group distributions for Adult are also shown in Table 4, indicating significant imbalances that are common in real-world demographic data.

- 5.1.6 (F) D-Vlog Depression Detection Dataset:

The D-Vlog dataset is a multimodal dataset designed for depression detection, containing visual and acoustic features extracted from YouTube videos [64]. It includes 555 depressed and 406 non-depressed samples, belonging to 639 females and 322 males. The videos were processed by truncating those longer than 596 seconds and zero-padding shorter ones. For fairness analysis, D-Vlog provides gender as the primary sensitive attribute. We followed the training and testing splits as provided by the dataset authors. The positive label (Y=1) indicates the presence of depression. Table 6 provides detailed statistics on label, duration, and sensitive attribute distributions for D-Vlog, notably highlighting truncation differences across genders.

Implementation and Training Details

For all experiments involving the synthetic datasets, COMPAS, and Adult datasets, we utilized Bayesian Neural Networks (BNNs) for both classification and uncertainty estimation. To overcome the intractability of calculating the true posterior distribution P(Y|X), we employed the widely recognized Bayes by Backprop method [6]. This method minimizes a variational free energy objective, which comprises a Kullback-Leibler (KL) divergence term that encourages the approximate posterior to be close to

the prior, and a numerically stable negative log-likelihood term (classification loss) [38, 40]:

$$L(\theta)=m=1\sum M[\log q\theta(\omega m)-\log P(\omega m)]+\lambda LNLL(Y,Y\hat{})(22)$$

Here, $q\theta(\omega m)$ is the approximate posterior distribution over the weights ωm, $P(\omega m)$ is the prior distribution over the weights, LNLL is the log negative log-likelihood (our classification loss), and λ is a constant scaling factor.

For optimization across all BNN experiments, we used the Adam optimizer [39]. Following the configuration detailed by Kwon et al. [42], we set the number of Monte Carlo samples (M, referred to as T in some literature for test-time sampling) for uncertainty quantification (as defined in Section 4.1) to T=10. Additionally, as per one of the settings in Blundell et al. [6], we used 10 Monte Carlo samples to approximate the variational posterior $q\theta(\omega)$ during training. The initial mean of the posterior was sampled from a Gaussian distribution with $\mu=0$ and $\sigma=1$. The weighting factor for the prior (π in a mixture of Gaussians prior) was set to 0.5, and the $\sigma1$ and $\sigma2$ values for the scaled mixture of Gaussians were set to 0 and 6, respectively. The constant λ from the BNN training objective was set to 2000. Early stopping was employed across all experiments to determine the optimal number of training iterations, preventing overfitting.

Dataset-Specific Hyperparameters and Architectures:

- Synthetic Datasets: Due to their relative simplicity, BNNs with no hidden layers (i.e., a direct input-to-output mapping with a probabilistic interpretation) proved sufficient. These models were trained for 5 epochs with a batch size of 8.
- COMPAS Recidivism Dataset: We used a BNN with a single hidden layer comprising 100 neurons. The model was trained for 10 epochs with a batch size of 256. For fairness evaluation, we considered race, gender, and age. For race, African-Americans were designated as the minority group (G=0) against Whites (G=1), following established literature [65]. For gender, females were set as the minority group (G=0) due to their lower representation. For age, individuals younger than 25 were considered the minority group (G=0), and those older than 45 were the majority group (G=1), excluding the intermediate age group (25-45) for simplification, as extending measures to multi-valued settings is a straightforward extension for future work [63].
- Adult Income Dataset: A BNN with no hidden layers was used, with an intermediate size (number of units in the final layer before the probabilistic output) of 25. The model was trained for 5 epochs with a batch size of 256. The fairness analysis was conducted for race, gender, and age to maintain consistency with the COMPAS dataset.
- D-Vlog Depression Detection Dataset: The high dimensionality of D-Vlog samples (596s of 136-dim visual and 25-dim acoustic features) posed significant challenges for standard BNN training. Therefore, we adopted the transformer-based Depression Detector architecture proposed by Yoon et al. [64]. For uncertainty estimation with this architecture, instead of MC Dropout (which could be applied, but ensembles offer stronger uncertainty estimates), we followed the Deep Ensembles approach [43, 32]. Specifically, we trained an ensemble of T=5 different models on the same training set, and their collective predictions on the test set were used to quantify uncertainty using the same Equation (7) as with BNNs. The choice of T=5 is supported by existing literature suggesting that performance often peaks around this number for ensembles [32]. The training configurations (learning rate of 0.0002, batch size of 32, optimized for 50 epochs using Adam [39]) were directly adopted from Yoon et al. [64], with an empirically chosen dropout rate of 0.1 where not explicitly provided.

In all cases, hyperparameters were carefully tuned to mitigate overfitting and ensure robust model performance.

Evaluation Measures

To provide a comprehensive evaluation, we assessed model performance using standard classification metrics and quantified fairness using both traditional point-based measures and our novel uncertainty-based measures.

Classification Performance Measures

We evaluated the classification performance of our models using the following widely accepted metrics:

- Accuracy (MAcc): The proportion of correctly classified instances.
- Positive Predictive Value (MPPV): Also known as Precision, calculated as TP/(TP+FP), where TP is True Positives and FP is False Positives.
- Negative Predictive Value (MNPV): Calculated as TN/(TN+FN), where TN is True Negatives and FN is False Negatives.
- False Positive Rate (MFPR): Calculated as FP/(FP+TN).
- False Negative Rate (MFNR): Calculated as FN/(FN+TP).

Fairness Measures

Following established conventions in fairness research [23, 63, 15], fairness measures (F) are typically expressed as ratios between the performance of the minority group (G=0) and the majority group (G=1). A value of 1 indicates perfect fairness. Values further from 1 (either much greater or much less than 1) indicate increasing unfairness. For empirical analysis, a deviation of more than 0.2 from 1 (i.e., $|F-1|>0.2$) is often considered a threshold for significant unfairness [23, 66].

- Statistical Parity (FSP):

$$FSP=P(Y\hat{}=1|G=1)P(Y\hat{}=1|G=0)(23)$$

Compares the proportion of positive predictions across groups.

- Equal Opportunity (FEOpp):

FEOpp=P(Y^=0|Y=1,G=1)P(Y^=0|Y=1,G=0)(24)

Compares the false negative rates (or equivalently, true positive rates) across groups for the positive true class.

- Equalized Odds (FEOdd):

FEOdd=P(Y^=1|Y=y,G=1)P(Y^=1|Y=y,G=0)(25)

This is calculated for both y=0 and y=1 (i.e., comparing true positive rates and false positive rates) across groups. In our results, we typically report the ratio for y=1 unless specified.

- Equal Accuracy (FEAcc):

FEAcc=MAcc(D,f,G=1)MAcc(D,f,G=0)(26)

Compares the overall accuracy between groups.

- Uncertainty Fairness (FU): (Our proposed measure)

FU=U(D,f,G=1)U(D,f,G=0)(27)

Here, U can be specifically:

  - FAlea: Aleatoric Uncertainty Fairness, where U is the average aleatoric uncertainty (Ua).
  - FEpis: Epistemic Uncertainty Fairness, where U is the average epistemic uncertainty (Ue).
  - FPred: Predictive Uncertainty Fairness, where U is the average total predictive uncertainty (Up).

These comprehensive evaluation measures allow us to not only assess the general performance of our models but, more importantly, to analyze the multifaceted nature of fairness, revealing both traditional statistical biases and the subtle disparities in model confidence as captured by our uncertainty-based metrics.

RESULTS/DISCUSSION

This section presents and discusses the experimental results obtained from both synthetic and real-world datasets, highlighting the utility and insights provided by our uncertainty-based fairness measures in conjunction with traditional point-based metrics.

6.1 Experiment 1: Synthetic Datasets

Our experiments with synthetic datasets served as crucial demonstrations of the theoretical independence between point-based and uncertainty-based fairness, as formalized in Proposition 4.1. The controlled nature of these datasets allowed us to isolate specific types of bias and observe how they manifest in different fairness measures.

Analyzing ¬(Fair(f;M,D)⟹Fair(f;U,D))

With reference to Synthetic Dataset 1 (SD1) and Synthetic Dataset 2 (SD2), as introduced in Section 5.1 and visually depicted in Figure 3(a) and Figure 3(b) respectively, we strategically selected the group exhibiting higher uncertainty estimations as the minority group (G=0). The detailed results are presented in Table 1.

For both SD1 and SD2, our Bayesian Neural Network (BNN) classifier demonstrated a commendable level of performance in the classification task, evidenced by high accuracy and low misclassification rates across both groups (e.g., MAcc around 0.95). Crucially, when evaluated using widely-used point-based fairness measures (FSP, FEOpp, FEOdd, and FEAcc), the classifier appeared to be fair, with all fairness ratio values $|F-1| \leq 0.2$, adhering to common thresholds for acceptable fairness [23]. This suggests that, purely based on prediction outcomes, the model's behavior was statistically balanced across the demographic groups.

However, a strikingly different picture emerged when we applied our uncertainty-based fairness measures:

- For SD1, which was designed to have disparate inherent noise, our classifier showed significant unfairness in terms of aleatoric uncertainty (FAlea=4.68). This indicates that despite equivalent point-based performance, the model was substantially more uncertain about the inherent data noise for the minority group (G=0) compared to the majority group (G=1). This disparity correctly reflects the design of SD1, where the minority group's data distribution had higher intrinsic variability.

- For SD2, engineered to induce differences in model knowledge, we observed profound unfairness in terms of epistemic uncertainty (FEpis=2.75). This high ratio implies that the model was considerably less confident in its predictions for the minority group, likely due to insufficient exposure to diverse data points from that specific region of the input space. This aligns with our expectation that epistemic uncertainty would highlight areas where the model "lacks knowledge."

These results emphatically support the first part of Proposition 4.1: a model can indeed be fair according to traditional point-based measures, yet profoundly unfair when assessed through the lens of uncertainty. This demonstrates the critical complementary insights provided by our uncertainty-based metrics.

Analyzing ¬(Fair(f;U,D)⟹Fair(f;M,D))

To illustrate the inverse non-implication, we turned to Synthetic Dataset 3 (SD3), as described in Section 5.1 and visualized in Figure 3(c). For SD3, we designated the group with the lower classification performance as the minority group (G=0). The results, also presented in Table 1, reveal a compelling scenario:

The classifier provided a strong level of performance for the majority group (G=1). However, when examining point-based fairness measures, the classifier exhibited clear unfairness. For instance, FEOdd=7.90 and

FEAcc=0.79, both falling outside the acceptable fairness boundary. This signifies that the model's predictions, particularly in terms of equalized odds and overall accuracy, were significantly biased against the minority group.

In stark contrast, when evaluating the model using our uncertainty-based fairness measures, the classifier appeared to be fair. The ratios for FEpis, FAlea, and FPred were all close to 1.0 (e.g., FEpis=1.05, FAlea=1.04, FPred=1.04). This indicates that, despite its biased point predictions, the model exhibited similar levels of aleatoric and epistemic uncertainty across both the minority and majority groups. That is, the model was equally "certain" or "uncertain" for both fair and unfair predictions across groups.

This result definitively supports the second part of Proposition 4.1: a model can be fair in terms of its uncertainty estimates, yet simultaneously unfair in its conventional point predictions. This underscores the necessity of a multi-faceted approach to fairness assessment, where uncertainty metrics serve as an independent and crucial diagnostic tool that complements traditional measures.

6.2 Experiment 2: Real-world Datasets

Having established the theoretical independence and complementary nature of uncertainty-based fairness on synthetic data, we now present a comprehensive analysis of its utility on complex real-world datasets: COMPAS, Adult, and D-Vlog.

6.2.1 The COMPAS Dataset

The COMPAS dataset, used for recidivism prediction, is a high-stakes application where algorithmic fairness is critically important. Our analysis focused on biases related to race, gender, and age. Table 2 provides results for race (Black vs. White) and gender (Female vs. Male), while Table 3 extends this to include age (younger than 25 vs. older than 45).

Fairness Across Race (Black vs. White):

Most point-based fairness measures (FSP = 2.84, FEOpp = 2.19, FEOdd = 1.57) strongly captured this bias, indicating significant unfairness (well outside the $|F-1| \leq 0.2$ threshold). Interestingly, FEAcc = 1.03 suggested approximate fairness in overall accuracy, demonstrating how individual metrics can provide conflicting insights and underscoring the need for a holistic view.

Our uncertainty-based fairness measures echoed and deepened these findings. African-Americans, despite having more samples, exhibited higher average epistemic uncertainty (Ue=0.0006 vs. 0.0004 for Whites) and aleatoric uncertainty (Ua=0.2299 vs. 0.1578 for Whites). This translated to significant unfairness in terms of uncertainty (FEpis=1.55 and FAlea=1.46), implying that the model was both less knowledgeable (higher Ue) and more susceptible to inherent data noise (higher Ua) for African-Americans. This observation is particularly insightful: even with more data, the model's confidence in its predictions for this group was lower, suggesting a deeper issue potentially related to data quality, representation, or complexity specific to the African-American demographic within the dataset.

Fairness Across Gender (Female vs. Male):

For gender, females generally showed better prediction performance compared to males, except for a high False Negative Rate (MFNR = 0.67 for Females vs. 0.39 for Males). This particular disparity suggests a bias of the classifier towards predicting Y=0 ("no recidivism") for females.

Both point-based and uncertainty-based fairness measures captured this bias, primarily against males in this case. While point-based measures like FSP = 0.31, FEOpp = 0.54, and FEOdd = 0.40 clearly indicated unfairness, FEAcc = 1.13 again suggested near-fairness in overall accuracy. Regarding uncertainty, males experienced higher average epistemic uncertainty (Ue=0.0006 vs. 0.0003 for Females), leading to FEpis=0.50 (meaning males have half the epistemic uncertainty of females, so this is biased against males having high uncertainty, and perhaps the model is too confident in its incorrect predictions for males, or needs more data for males to reduce uncertainty). However, the fairness gaps for aleatoric uncertainty (FAlea=0.78) and predictive uncertainty (FPred=0.78) were closer to the acceptable boundary (0.8), implying that while there are confidence disparities, the primary issue across gender might stem from the sample imbalance problem (females: 1175, males: 4997, as shown in Table 4), leading to the model being less uncertain where it has more data.

Across the point-based fairness measures, most indicated unfair predictions, with FSP = 2.44, FEOpp = 1.48, and FEOdd = 2.37. Similar to the race attribute, FEAcc = 0.85 suggested fair predictions. These figures strongly imply that the model was unfairly inclined to predict Y=1 (recidivating an offense) for the subgroup younger than 25.

Crucially, our uncertainty-based fairness measures corroborated these findings, showing significant disparities: FEpis=4.35, FAlea=3.36, and FPred=3.37. The exceptionally high FEpis (4.35) for the younger-than-25 group suggests a severe lack of model knowledge for this subgroup, despite the dataset actually containing fewer samples for the older-than-45 group. This indicates that the model's behavior is disproportionately uncertain for younger individuals, pointing to a need for more representative data for this age demographic or perhaps a re-evaluation of feature importance for this group. The high FAlea (3.36) also implies that the classification task is inherently harder (more noise/ambiguity) for the younger-than-25 group as perceived by the model.

Social Impact and Insights from COMPAS:

The COMPAS analysis illustrates how traditional point-based measures effectively highlight prediction biases, such as the disproportionate recidivism predictions for African-Americans. However, our uncertainty-based measures go a step further, providing valuable diagnostic insights into the potential sources of these biases. The higher epistemic uncertainty for African-Americans and younger individuals, despite reasonable sample sizes, suggests underlying data-level biases. This could be due to issues like class imbalance within these groups, group-dependent labeling noise, or annotation bias [62, 11]. For instance, it's known that labels for criminal activity generated via crowdsourcing can be systematically biased against certain subgroups [20]. Our measures, by showing higher epistemic, aleatoric, and predictive uncertainties for African-Americans and males, even with larger sample sizes, suggest that the model faces greater difficulty and uncertainty for these groups, possibly due to unseen data characteristics or label quality issues. This makes our measures useful diagnostic tools in real-world settings where clean and accurate labels are often scarce. They highlight where balancing samples (e.g., across gender) or investigating data quality (e.g., label noise) might be more effective interventions than simply adjusting a model to meet point-based fairness criteria. Future research could explicitly verify the impact of unbiased labels on both types of fairness measures.

6.2.2 The Adult Dataset

The Adult Income dataset, aiming to predict income levels, also presents significant challenges due to class and group imbalances (Table 4). We analyzed fairness across race (Black vs. White) and gender (Female vs. Male). Table 5 summarizes the results.

Fairness Across Race (Black vs. White):

The Adult dataset has a severe imbalance, with only 2,817 samples for African-Americans (G=0) compared to 25,933 for Whites (G=1). Despite this, the point-based fairness gaps between African-Americans and Whites appeared lower than in the COMPAS dataset (e.g., FSP = 0.75, FEOpp = 1.08, FEOdd = 0.87, FEAcc = 1.12).

However, the uncertainty-based fairness measures provided crucial, contrasting insights. We observed a surprisingly large fairness gap in terms of epistemic uncertainty (FEpis=151). This extreme value is likely attributable to the massive difference in sample sizes: Whites having approximately 10 times more samples leads to a very small average epistemic uncertainty for the majority group. The model is extremely confident for the well-represented group, but proportionally much less so for the minority group. In contrast, the average aleatoric uncertainty values (Ua) for both groups were very small (e.g., 0.01), resulting in FAlea≈1.00. This suggests that while the model might struggle with knowledge for the minority group, the inherent data noise is perceived to be similar across racial groups.

Fairness Across Gender (Female vs. Male):

Similar to race, there was a class imbalance across gender (9,872 females vs. 20,380 males). The point-based fairness measures presented conflicting outcomes: FEOpp = 1.04 and FEAcc = 1.18 suggested fair classification, while FSP = 0.62 and FEOdd = 0.79 indicated bias. Specifically, FSP = 0.62 implied a higher salary classification bias in favor of males.

For uncertainty-based fairness, we observed gaps similar to the race attribute. There was significant bias in terms of FEpis (521 for males, implying they have less epistemic uncertainty compared to females, who have Ue=0.0001 vs Ue=6e−8 for males), despite males being the more represented group. This could mean the model is overly confident in its male predictions, even if some are incorrect. FAlea was approximately 1.00, suggesting consistent inherent data noise levels across genders.

Social Impact and Insights from Adult:

The Adult dataset results underscore a critical limitation of relying solely on point-based fairness metrics. These measures might indicate acceptable fairness, potentially masking a deeper underlying issue where the model is highly unsure of its predictions for minority groups (high Ue) despite sufficient sample sizes for other groups. This could lead to substantial prediction biases when real-world challenges like missing data or distributional shifts occur [28, 10]. Our uncertainty-based fairness measures successfully highlighted these discrepancies across both race and gender. They encourage a more proactive approach, urging further investigation into the root causes of bias in the model (e.g., unrepresentative features, subgroup complexity) before deployment in real-world settings. A model that is "fair" on aggregate but highly uncertain for specific groups is less trustworthy and potentially more harmful in high-stakes decisions.

6.2.3 D-Vlog Depression Detection Dataset

The D-Vlog dataset, comprising visual and acoustic features for depression detection, introduced a unique challenge related to data preprocessing: video truncation. As detailed in Table 6, female videos were truncated significantly more than male videos, leading to a potential loss of information primarily affecting the female group.

Table 7 presents the experimental results for D-Vlog across multi-modal, audio-only, and visual-only architectures. For the multi-modal model, both point-based and uncertainty-based fairness measures largely indicated fairness (except for FEOdd = 1.68). This was initially surprising, given that the female group size was twice that of the male group, yet no strong bias was detected. We observed high aleatoric uncertainty for both groups (Ua≈0.45), suggesting that the task itself might be inherently ambiguous.

However, the results for uni-modal (audio-only) analysis were particularly insightful. Here, the audio modality exhibited a strong bias against females, with generally

lower performance measures for females compared to males. Crucially, while point-based measures did not coherently capture this bias (e.g., FSP = 0.75, FEOpp = 0.73, FEAcc = 0.84), our uncertainty-based measures consistently highlighted the bias. For audio-only, FEpis=1.38 and FAlea=1.32 indicated higher uncertainty for females. The root cause of this bias appears to be the aforementioned video truncation: female recordings were significantly longer and thus truncated more, leading to a greater reduction in useful information for classification for females, which naturally increased the model's uncertainty for this group. This demonstrates how aleatoric uncertainty, in particular, can reveal biases stemming from data processing or collection issues that directly impact data quality and inherent ambiguity.

Conversely, this effect was less pronounced across the visual modality, where the classifier performed poorly across both males and females, resulting in relatively balanced but poor performance and uncertainty metrics.

6.3 Experiment 3: Individual Fairness

Beyond group-level analyses, understanding individual fairness – the principle that similar individuals should receive similar predictions and, as we propose, similar uncertainties – is vital for ensuring granular equity. We analyzed individual fairness using both point-based and uncertainty-based measures for the COMPAS and Adult datasets.

6.3.1 Individual Fairness Analysis on COMPAS

The results for individual fairness on COMPAS are depicted in Figure 4. This figure shows the consistency scores (Equation 6 for point-based, Equation 9 for uncertainty-based) across different k values for k-nearest neighbors, distinguishing between positive (+) and negative (−) outcomes for Black (B) and White (W), and Female (F) and Male (M) groups.

Point-based Individual Fairness (Fy^indv):

Figures 4(a) and 4(b) illustrate that point-based consistency values (Fy^indv) indeed differ across groups and outcomes. For instance, the consistency for "W(-)" (White, non-recidivating) and "B(-)" (Black, non-recidivating) samples appears similar, suggesting consistent predictions for similar individuals within these non-recidivating groups. However, there are evident differences when comparing "B(+)" (Black, recidivating) to "W(+)" (White, recidivating), indicating that predictions for similar recidivating individuals might be less consistent across racial groups.

Uncertainty-based Individual Fairness (FUindv):

Present the uncertainty-based individual fairness measures for COMPAS. These measures offer a deeper diagnostic layer:

- Aleatoric Consistency (FUaindv): While "W(-)" and "B(-)" showed similar point-based consistencies, their aleatoric consistencies (Figure 4(d)) differ. This suggests that even if the predicted outcome is consistent, the inherent ambiguity or noise in the data for similar individuals might be disparate. Higher aleatoric inconsistencies for "B(+)" samples align with the observation that the classifier faces greater difficulty with this group, potentially due to noisy labels or complex feature interactions specific to them.

- Epistemic Consistency (FUeindv): Figures 4(e) and 4(f) highlight the epistemic consistency. Lower epistemic consistency for "M(+)" (Male, recidivating) and "B(+)" (Black, recidivating) groups implies that the model's knowledge or confidence is less consistent for similar individuals within these groups. This strongly suggests that these specific subgroups would particularly benefit from additional, diverse data to reduce the model's uncertainty and improve its generalizability.

This analysis demonstrates that uncertainty-based individual fairness measures can reveal granular disparities that point-based measures might miss, providing crucial insights into which specific subgroups (defined by group and outcome) suffer from higher model ignorance or data ambiguity.

6.3.2 Individual Fairness Analysis on Adult

Point-based Individual Fairness (Fy^indv):

Uncertainty-based Individual Fairness (FUindv):

The uncertainty-based individual fairness for Adult. The perfect consistency observed in point predictions for positive classes also largely extends to aleatoric and epistemic consistency (≈1.000).

However, we observe slightly lower consistencies for negative classes in both point predictions and aleatoric uncertainty. The high FNR rate for both groups (Table 4) suggests that there are more errors with Y^=0 predictions (i.e., failing to predict a positive outcome), leading to higher inconsistencies for those predictions. More importantly, lower epistemic consistencies for "F-" (Female, income ≤50K) and "B-" (Black, income ≤50K) indicate that more data would be beneficial for these specific subgroups. This observation is strongly supported by the dataset distribution (Table 4), which shows that females and African-Americans are minority groups in the dataset and likely have sparser representation in the negative class.

Social Impact of Individual Fairness with Uncertainty:

Considering uncertainty in individual fairness is crucial for deploying ML models in sensitive applications. For instance, in medical diagnoses, a cancer-free prognosis for a patient should not only be accurate but also exhibit similar uncertainty-based consistency to that of similar healthy individuals. If the model is less confident for certain patient demographics, even if its point predictions are correct, this disparity in confidence could lead to distrust, misallocation of resources (e.g., unnecessary follow-up tests for some groups), or systematic neglect of

subtle warning signs. Our framework provides a tool to audit and improve this granular consistency, ensuring that all individuals, regardless of their sensitive attributes, are treated with comparable levels of confidence by the model.

6.4 Experiment 4: Ablation Analysis

To further understand the behavior of our models and the sensitivity of uncertainty estimations, we conducted an ablation study analyzing the effect of model capacity (specifically, the number of neurons per hidden layer) on both performance and uncertainty estimates, particularly on the COMPAS dataset. The results are visualized in Figure 6.

Figure 6(a) illustrates the accuracy of the model for different hidden layer sizes. It shows that accuracy tends to saturate after approximately 100 neurons per hidden layer for all demographic groups (Black, White, Male, Female). Beyond this point, increasing model capacity does not yield substantial performance gains.

Figures 6(b) and 6(c) display the average aleatoric uncertainty and epistemic uncertainty, respectively, across different hidden layer sizes for the various groups.

- For aleatoric uncertainty (Figure 6(b)), we observe that increasing the hidden layer size generally leads to a reduction in average aleatoric uncertainty, suggesting that more complex models might better capture and account for the inherent noise in the data. However, the relative ordering and the differences in aleatoric uncertainty between the demographic groups (e.g., Black vs. White, Male vs. Female) largely persist across varying model capacities.

- Similarly, for epistemic uncertainty (Figure 6(c)), as the model capacity increases, the average epistemic uncertainty generally decreases. This is expected, as a larger model might have more capacity to learn from the available data, thereby reducing its "ignorance." Crucially, akin to aleatoric uncertainty, the relative disparities in epistemic uncertainty across different demographic groups remain consistent regardless of the number of neurons.

These findings are significant: while model capacity does influence the absolute levels of uncertainty, the relative fairness gaps in terms of uncertainty between demographic groups appear to be robust to changes in model architecture within a reasonable range. This stability reinforces the idea that disparate uncertainty is an intrinsic property reflecting data characteristics or the model's fundamental limitations, rather than a mere artifact of model complexity. Based on these observations, and to balance performance with computational cost, we chose a hidden layer size of 100 neurons for our BNNs in all the main experiments. It is also important to note that adding more hidden layers (i.e., increasing depth) beyond a single layer led to significant overfitting problems for the synthetic, COMPAS, and Adult datasets, further justifying our choice of a single hidden layer (or no hidden layer for simpler datasets).

DISCUSSION

In this paper, we have critically re-evaluated existing point-based fairness measures, arguing that their sole reliance on discrete prediction outcomes—while valuable—provides an incomplete and potentially misleading view of a model's true fairness. By neglecting the underlying uncertainty or confidence of these predictions, traditional measures become susceptible to nuances stemming from data quality, representation, and model knowledge. To address this fundamental limitation, we introduced and rigorously explored the use of various types of uncertainty, particularly aleatoric and epistemic uncertainty, as novel and complementary fairness measures. Through both theoretical proofs and extensive empirical validation on a range of synthetic and real-world datasets, we have demonstrated that these uncertainty-based fairness measures are independent of point-based measures, and, more importantly, they offer profound insights into the presence and underlying sources of bias in machine learning predictions.

7.1 Main Insights

Our comprehensive experimental analysis has yielded several key insights regarding the power of uncertainty-based fairness:

- Insights through the Epistemic Fairness Measure (FEpis):

The epistemic fairness measure (FEpis), by its very definition, quantifies the disparity in the model's "lack of knowledge" across different groups. A high FEpis ratio indicates that the model is significantly more uncertain (or less knowledgeable) about predictions for one group compared to another. What is particularly beneficial about this measure is that it is not solely driven by the sheer number of samples in a group, which can often be misleading. For instance, in the COMPAS dataset, both Black and Male groups have substantially more samples than their counterparts. Yet, our results consistently showed higher average epistemic uncertainty (Ue) for these groups (Tables 2 and 3). This suggests that despite larger raw sample sizes, the dataset for these groups may still contain data-level biases, such as subtle class-imbalance problems within specific subgroups or a lack of diversity in their feature distributions that the model struggles to generalize from. The high epistemic uncertainty for African-Americans and individuals younger than 25 in COMPAS, for example, directly points to an actionable need for more representative and diverse data for these specific segments, rather than simply more raw data.

- Insights through the Aleatoric Fairness Measure (FAlea):

Aleatoric uncertainty fundamentally reflects the inherent

"hardness" of a prediction task due to irreducible noise, ambiguity, or variability within the data itself [38]. Our use of this informative measure revealed that the classification task can be intrinsically harder for some groups compared to others. A prime example is the D-Vlog dataset. The analysis showed that the truncation of female video recordings during data collection led to a significant increase in aleatoric uncertainty for females (Table 7). This is a critical insight: even if a model makes statistically "fair" point predictions, if the underlying data for one group is inherently noisier or less complete (due to collection artifacts or real-world variability), the model will reflect this as higher aleatoric uncertainty, which our FAlea metric effectively captures.

Another prominent illustration comes from the COMPAS dataset. Despite being a widely used benchmark for fairness, COMPAS is known to suffer from issues like typographical errors and data inaccuracies [54]. Our uncertainty-based measures, by quantifying the perceived data noise, can inherently capture some of these data quality issues that would be completely missed by traditional point-based measures. As evidenced in Table 3, the higher FAlea for the younger-than-25 group suggests that the data points for this demographic are inherently more ambiguous or noisy from the model's perspective, thus providing valuable diagnostic information about the roots of bias beyond simple classification accuracy.

7.2 Social Impact

The rapid proliferation of machine learning models has been paralleled by an escalating urgency to ensure their fairness and mitigate bias. While the past few years have witnessed a surge in proposed bias mitigation methods [33], there often remains a critical lack of clarity regarding which source of bias each method is designed to address. This ambiguity can lead to suboptimal or even counterproductive interventions. For instance, recent research has demonstrated that when bias is attributed to missing values, some existing mitigation methods might inadvertently reduce point-based performance disparities at the cost of overall accuracy [61]. Our contribution lies in leveraging existing, well-understood uncertainty measures to quantify an alternative aspect of fairness, thereby providing a more precise diagnostic tool.

Our framework allows for a clearer distinction between biases that arise from the model's lack of knowledge (epistemic uncertainty, addressable by more data) and those stemming from irreducible noise or ambiguity in the data itself (aleatoric uncertainty, which might require re-evaluating the task or data collection processes). This disaggregation is crucial. If the bias is due to epistemic uncertainty, the solution might involve targeted data augmentation or collection for underrepresented groups. If it's aleatoric, the problem might lie in the inherent ambiguity of the labels or features for a specific group, demanding more robust labeling protocols or even a redefinition of the prediction task. The inherent properties of epistemic and aleatoric uncertainties—that they are meant to be irreducible given a dataset and model—mean that simply forcing point-based parity in such scenarios will likely be suboptimal or misleading. Our proposed uncertainty-based measures directly highlight these underlying problems, serving as a critical caution against blindly pursuing "fair" outcomes solely based on point-based metrics.

Furthermore, many existing bias mitigation solutions are predicated on optimistic machine learning assumptions, such as access to perfectly clean, noise-free labels and the assurance that the model will be deployed in an environment that perfectly mirrors its training setting [36]. This incongruence between theoretical formulations and real-world settings represents a significant handicap for the machine learning fairness research community. Our work directly addresses this gap by highlighting the need to develop methods capable of addressing both epistemic and aleatoric sources of discrimination. By providing a quantifiable metric for these inherent uncertainties, we hope that our proposed uncertainty-based fairness measures serve as a foundational step towards designing more robust, practical, and ethically informed bias mitigation strategies for real-world deployments. They offer transparency into why a model might be struggling with a particular group, guiding developers towards more effective and targeted interventions for developing truly responsible AI systems.

7.3 Limitations

Despite the significant merits and insights offered by uncertainty-based fairness measures, it is important to acknowledge certain limitations that also present promising avenues for future research.

First, the primary constraint is the requirement for models that inherently provide or can be modified to provide uncertainty estimates. Our current framework relies on techniques like Bayesian Neural Networks (BNNs), Deep Ensembles, and Monte Carlo Dropout. While powerful, these methods are not universally applicable to all state-of-the-art deep learning architectures without significant modifications to their training procedures or inference pipelines. This can sometimes hinder the direct application of our fairness analysis to cutting-edge models that prioritize predictive accuracy over explicit uncertainty quantification.

Second, quantifying uncertainty in a reliable and computationally efficient manner remains an active research area. While we have demonstrated consistent outcomes using both BNNs and Deep Ensembles, we encountered challenges with the overall ranges and absolute values of estimated uncertainties across different datasets. Newer approaches to uncertainty quantification, such as advanced Deep Deterministic Uncertainty methods [49, 44, 57], are continuously emerging and could potentially offer more reliable or computationally lighter

alternatives. A thorough comparative evaluation of different uncertainty quantification methods within the context of fairness remains an important direction for future work. The computational overhead involved in obtaining multiple predictions (e.g., Monte Carlo samples or ensemble predictions) can also be substantial, though one-pass uncertainty estimation approaches are being developed, albeit often with trade-offs in reliability [1].

Third, our current framework utilizes average uncertainty values across groups. While informative, taking the average across a group can potentially obscure important characteristics of the uncertainty distribution within that group. Future work could explore alternative metrics for measuring the dispersion or higher moments of uncertainty values within a group, such as the variance of uncertainties or the proportion of highly uncertain predictions, to capture a more granular view of fairness.

Finally, a crucial conceptual point is that while our uncertainty-fairness measures are differentiable and could theoretically be converted into loss functions for model training, directly forcing epistemic and aleatoric uncertainties to be similar across groups or individuals will not necessarily change the "real uncertainties." These measures fundamentally reflect issues inherent in the data (noise, ambiguity) or the model's learning (lack of knowledge). Simply optimizing for fairness in uncertainty without addressing the root cause (e.g., collecting more diverse data, improving data quality, or refining model architecture) might lead to artificially reduced uncertainty disparities without genuine improvement in fairness or model reliability. This highlights that our proposed measures are primarily diagnostic tools that point towards underlying problems, rather than direct mitigation strategies in themselves. However, this diagnosis is a critical first step towards effective and targeted interventions.

Despite these limitations, we sincerely hope that our work provides a significant stepping stone towards a more robust and nuanced understanding of algorithmic fairness. By integrating uncertainty quantification, we can move beyond superficial statistical equalities to address the deeper, more complex challenges that impede equitable outcomes in real-world machine learning deployments.

CONCLUSION

The pervasive influence of artificial intelligence in contemporary society underscores the paramount importance of ensuring that machine learning systems operate fairly and equitably. While the field has made significant strides in defining and measuring algorithmic fairness through a multitude of point-based metrics, this article has argued that such conventional approaches, by focusing solely on prediction outcomes, often overlook critical dimensions of a model's performance—namely, its confidence and inherent knowledge about its predictions. This oversight can mask subtle yet profound disparities that disproportionately affect different demographic groups.

This paper introduced a compelling framework that leverages uncertainty quantification (UQ) as a critical and complementary measure of algorithmic fairness. By carefully disentangling predictive uncertainty into its constituent parts—aleatoric uncertainty (stemming from inherent data noise) and epistemic uncertainty (reflecting the model's lack of knowledge)—and employing robust UQ techniques such as Bayesian Neural Networks, Monte Carlo Dropout, and Deep Ensembles, we demonstrated a more granular approach to fairness assessment. Our theoretical propositions and extensive empirical evaluations on both synthetic and real-world datasets (COMPAS, Adult, D-Vlog) unequivocally show that uncertainty-based fairness measures are independent of traditional point-based metrics. This independence signifies that a model can appear statistically "fair" by conventional means yet be profoundly "unfair" in its confidence levels, or vice versa.

The insights gleaned from our uncertainty-based analysis are profoundly actionable. Disparities in epistemic uncertainty across groups serve as powerful indicators of data scarcity or representational bias, signaling the urgent need for more diverse and comprehensive data collection for underserved populations. Conversely, disparities in aleatoric uncertainty can reveal intrinsic ambiguities or noise within the data pertaining to specific subgroups, guiding interventions towards data cleaning, improved labeling protocols, or even a re-evaluation of the problem's definition. Beyond diagnosis, integrating uncertainty fosters enhanced transparency and trustworthiness in AI systems. By revealing not only what a model predicts but also how confident it is in that prediction for different individuals, we empower users with a clearer understanding of algorithmic decisions, fostering greater accountability and informed human-in-the-loop interventions. This allows for proactive bias mitigation, where uncertainty acts as an early warning signal, enabling targeted interventions before models are widely deployed.

Ultimately, incorporating uncertainty into the fairness discourse moves us beyond superficial statistical averages to consider the reliability and consistency of individual predictions. This holistic view provides a deeper, more nuanced understanding of algorithmic equity, revealing hidden biases and guiding the development of more robust, transparent, and genuinely fair machine learning solutions that better serve all members of society. Future research should prioritize developing standardized methodologies for measuring and comparing uncertainty across diverse sensitive groups, rigorously exploring the causal links between uncertainty disparities and various sources of bias (such as data missingness, label noise, and annotation bias), and creating comprehensive frameworks for designing uncertainty-aware fairness interventions. Through these concerted efforts, we can truly advance the

frontier of ethical and responsible artificial intelligence.

REFERENCES

[1] 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 uncertainty quantification in deep learning: Techniques, applications and challenges. Information fusion, 76, 243–297.

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

[3] 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.

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

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

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

[16] 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).

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

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

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

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

[26] 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.

[27] 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..

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] Kaiser, P., Kern, C., & R¨ugamer, D. (2022). Uncertainty-aware predictive modeling for fair data-driven decisions..

[36] 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.

[37] 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.

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

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

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

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

[42] 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.

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

[44] 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.

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

[46] 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.

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

[48] 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.

[49] 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.

[50] 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.

[51] 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).

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

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

[54] 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.

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

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

[57] 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.

[58] 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.

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

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

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