Uncertainty Matters: Stable Conclusions Under Unstable Assessment of Fairness Results

Ainhize Barrainkua[1] **Paula Gordaliza**[2] **Jose A. Lozano**[1,3] **Novi Quadrianto**[1,5,6]

[1]Basque Center for Applied Mathematics (BCAM), Spain
[2]Universidad Pública de Navarra (UPNA), Spain
[3]Institute for Advanced Materials and Mathematics (INAMAT[2]), Spain
[4]University of the Basque Country UPV/EHU, Spain
[5]Predictive Analytics Lab, University of Sussex, UK
[6]Monash University, Indonesia

Abstract

Recent studies highlight the effectiveness of Bayesian methods in assessing algorithm performance, particularly in fairness and bias evaluation. We present *Uncertainty Matters*, a multi-objective uncertainty-aware algorithmic comparison framework. In fairness-focused scenarios, it models sensitive group confusion matrices using Bayesian updates and facilitates joint comparison of performance (e.g., accuracy) and fairness metrics (e.g., true positive rate parity). Our approach works seamlessly with common evaluation methods like K-fold cross-validation, effectively addressing dependencies among the K posterior metric distributions. The integration of correlated information is carried out through a procedure tailored to the classifier's complexity. Experiments demonstrate that the insights derived from algorithmic comparisons employing the *Uncertainty Matters* approach are more informative, reliable, and less influenced by particular data partitions. Code for the paper is publicly available at `https://github.com/abarrainkua/UncertaintyMatters`.

1 INTRODUCTION

In the realm of Machine Learning (ML) research, recent studies have raised concerns about conventional methods of evaluating model performance. For instance, Aghbalou et al. (2023) study the bias inherent in traditional K-fold cross-validation (CV). On the other hand, Dwork et al. (2015) propose a general strategy to reuse holdout sets in repeated evaluations while preserving the statistical guarantees of fresh data, thereby ensuring the validity of adaptive analyses. Orthogonal to these contributions, there is a growing trend towards integrating uncertainty into algorithmic evaluation. Notably, researchers such as Benavoli et al. (2017) and Kruschke et al. (2018) have advocated treating performance metrics as random variables, employing Bayesian inference to update their posterior distributions. When considering algorithmic evaluation within a fairness-aware context, it introduces additional complexity due to its inherent multi-objective nature. Furthermore, this complexity is magnified by the necessity to partition the limited test data into sensitive groups for the evaluation of fairness metrics.

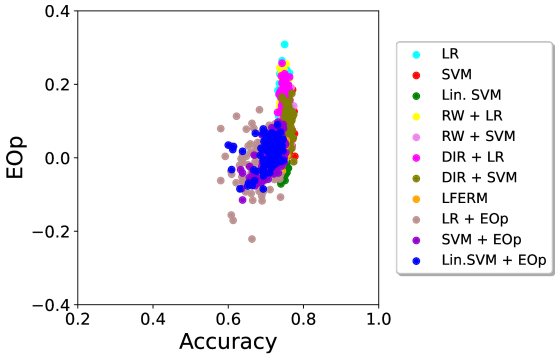
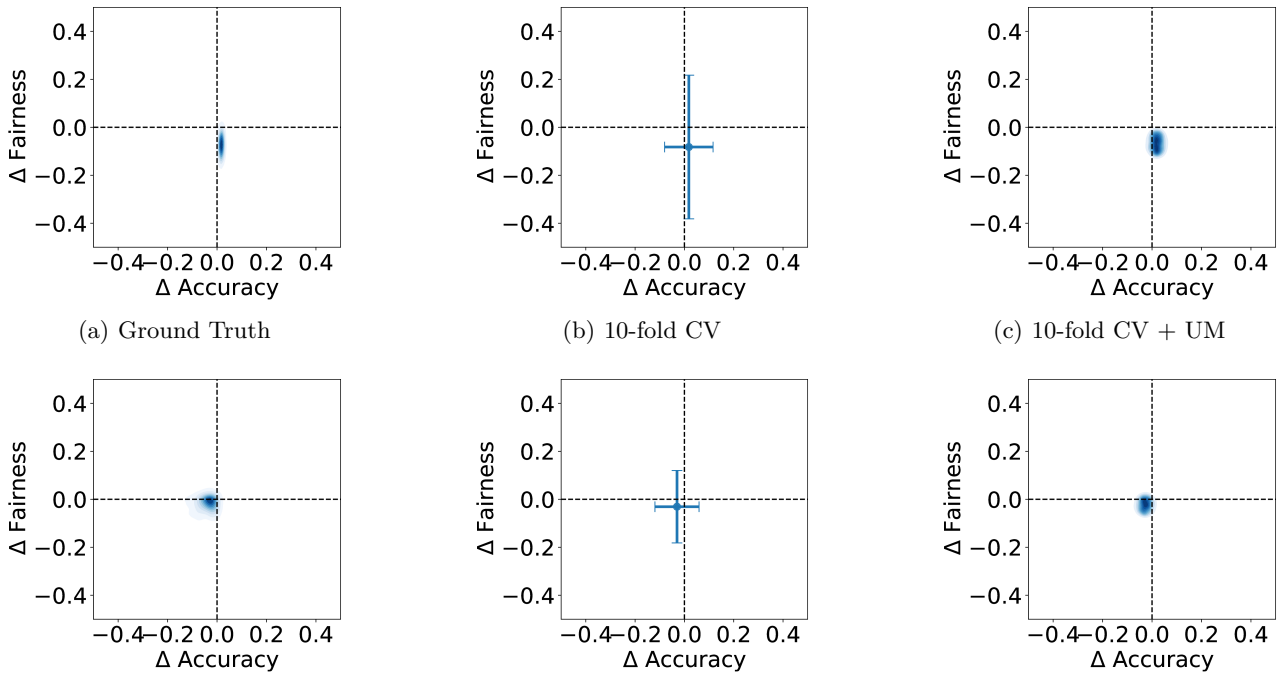


Figure 1: Instability of 10-Fold CV Results. The value of accuracy and fairness (Equality of opportunity-EOp), for 100 different 10-fold CV configurations and different fairness-enhancing methods (color in legend) using the German Credit dataset (Dua et al., 2017) with age as the sensitive attribute.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.



(a) Ground Truth (b) 10-fold CV (c) 10-fold CV + UM

(d) Ground Truth (e) 10-fold CV (f) 10-fold CV + UM

Figure 2: Performance Variability in Algorithmic Comparison. (a,d) The true probability distribution of algorithmic comparison; (b,e) One outcome (average and standard deviation) from traditional 10-fold CV; (c,f) Application of the Uncertainty Matters framework to the results obtained in (b) and (e), respectively. The top row illustrates the comparison between Feldman et al. (2015) + SVM vs. Donini et al. (2018), while the bottom row represents the comparison between LinearSVM + Hardt et al. (2016) and Donini et al. (2018). Fairness is assessed by the EOp metric. Uncertainty Matters significantly improves the reliability of standard 10-fold CV in algorithmic evaluation.

Ji et al. (2020), Friedler et al. (2019), and Qian et al. (2021) demonstrated that fairness metrics exhibit strong variability in hold-out evaluations or under different training-test splits. This trend extends to popular evaluation methods like 10-fold CV, as evidenced in Figure 1, with fairness metrics displaying greater variability than other metrics that evaluate predictive performance on the whole population (e.g. accuracy). This observation aligns with the presence of minority groups that the classifier often struggles to properly learn, resulting in increased instability in its performance within these subgroups. Thus, we are driven by the motivation to introduce uncertainty into fairness assessment, noting that this framework is not limited to fairness but can be applied in any scenario where conducting a multi-objective uncertainty-aware algorithmic comparison is desired.

When algorithmic comparison results display significant fluctuations depending on the selected data split, making a conclusive determination about the superiority of one algorithm over another through a single evaluation becomes inappropriate. Instead, by taking into account all the potential K-fold CV results, the situation suggests the existence of a true probability that one method might excel in certain objectives. The traditional method of K-fold CV, which provides only average and standard deviation values, fails to fully encapsulate this nuanced reality (note the difference between the outcomes in the first and the second columns of Figure 2). Obtaining the precise estimation of such probabilities necessitates consideration of all possible K-fold CV results, each obtained with different data splits, which is practically infeasible. Therefore, our primary aim is to precisely estimate these probabilities through a single evaluation with any random data split, by accounting for the uncertainty inherent in the evaluation process (see Figures 2(c) and 2(f)).

There have been several attempts to address the problem of accurately evaluating the uncertainty of fairness metrics in a supervised scenario. Early approaches focused mainly on a single fairness metric, typically Demographic Parity (DP). In Besse et al. (2018) and Besse et al. (2021), confidence intervals were built using the traditional Delta method, based on the asymptotic distribution of Disparate Impact, one of the most commonly used indexes for quantifying DP. Later, Ji

et al. (2020) considered a Bayesian framework and a calibration procedure to reduce such uncertainty by employing unlabeled data. However, these proposals share two main limitations. Firstly, simultaneous comparison of multiple metrics that account for the predictive performance and fairness guarantees of the learning algorithm is not possible. Secondly, metric uncertainty can only be addressed for already trained predictors or learning algorithms under hold-out evaluation and not for more intricate and popular evaluation frameworks such as K-fold CV, where the K results obtained are not independent due to the overlap of training instances between each of the K training sets.

To address these limitations, we leverage Bayesian inference and propose the *Uncertainty Matters* (UM) framework to quantify the uncertainty inherent in algorithmic evaluations by adopting a probabilistic representation of the confusion matrix (CM). Any metric expressed as a function of the CM is then a random variable. The main advantage of UM is that it enables the computation of any multi-dimensional joint posterior distribution. This allows a comparison of fair learning algorithms across multiple objectives. Importantly, we develop the statistical framework to be employed in the hold-out and K-fold CV evaluation settings. For K-fold CV, that generates correlated results, we introduce an *effective* CM (Wang et al., 2019), which unlike previous approaches, is dependent on the complexity and generalization capability of the algorithm. Assembling all components together, we end up with a multi-objective and uncertainty-aware scheme for algorithmic comparison.

We conducted extensive experiments, encompassing different fairness-enhancing algorithms, to evaluate our UM framework against traditional evaluation settings. These experiments revealed that UM yields more stable and informative conclusions compared to conventional evaluation methods (see, for instance, Figure 2). Specifically, it reduces dependence on the specific partition of the K-fold CV, enhancing result stability, informativeness, and reliability.

The rest of the paper is organized as follows. We describe our statistical framework UM in Section 2 and propose a strategy based on it for algorithmic comparison in Section 3. The related works are outlined in Section 4. The empirical evaluation of our proposal and the obtained results are described in Section 5. Finally, Section 6 provides a concise summary of the key findings, engages in a comprehensive discussion, and outlines potential avenues for future research.

2 THE UM FRAMEWORK

2.1 Notation and Problem Statement

Our approach is built upon a bias-aware supervised learning scenario. Let $\mathcal{A}$ be the predictive algorithm trained using a finite set of examples i.i.d. from an unknown distribution $P(\mathcal{X}\times\mathcal{S}\times\mathcal{Y})$, where every instance is represented by a set of d (non-sensitive) attributes $\mathbf{x}\in\mathcal{X}\subset\mathbb{R}^d$, a sensitive attribute $s\in\mathcal{S}=\{1,\dots,r\}$ (e.g., gender, age, marital status) and a class label $y\in\mathcal{Y}$. The outcome of the classification made by $\mathcal{A}$ will be denoted by $\hat{y}$. For explanatory purposes, we will refer to the binary classification problem where $\mathcal{Y}=\{-1,+1\}$, but UM works for multiclass $\mathcal{Y}$.

If the distribution $P(\mathcal{X}\times\mathcal{S}\times\mathcal{Y})$ were explicitly known, any metric could be evaluated precisely on $\mathcal{A}$ at population level by different combinations of the true probabilities $P(\hat{Y}=i|Y=j)$ or $P(\hat{Y}=i|Y=j,s), s\in\mathcal{S}$, for $i,j\in\{-1,+1\}$. However, only an approximation of these theoretical values can be generally obtained from the relative frequencies of a finite i.i.d. sample from such distribution. Hence, we employ a probabilistic representation of the CM and capture the inherent uncertainty in the evaluation process. This approach allows us to provide reliable information regarding algorithmic comparisons. Under this framework, the metrics will consist of random variables whose distribution would be derived from the probabilistic model of the CM.

Consider a test sample $\{(\mathbf{x}_1,s_1,y_1),\dots,(\mathbf{x}_N,s_N,y_N)\}$, where $N=\sum_{s=1}^{r}N_s$ and N_s is the size of the sensitive group s. Let $\mathcal{C}=\{\boldsymbol{C}_1,...,\boldsymbol{C}_r\}$ be the set of confusion matrices obtained as a result of the prediction given by algorithm $\mathcal{A}$ on subgroup $s\in\mathcal{S}$, namely $\boldsymbol{C}_s=(TP_s,TN_s,FP_s,FN_s)$, consisting of the number of true positives TP_s, true negatives TN_s, false positives FP_s and false negatives FN_s, satisfying $N_s=TP_s+TN_s+FP_s+FN_s$. Then the overall CM of $\mathcal{A}$, say $\boldsymbol{C}$, is obtained as $\boldsymbol{C}=\sum_{s=1}^{r}\boldsymbol{C}_s$. It is worth noting that, even though performance and fairness metrics of $\mathcal{A}$ serve distinct purposes, they both rely on essentially the same information. Indeed, fairness metrics are estimated from the counts of the protected groups in $\mathcal{C}$, while global predictive performance (e.g. accuracy) is generally calculated from the overall $\boldsymbol{C}$. In other words, fairness is generally quantified in terms of the values of certain performance metric θ (could be multi-dimensional $\boldsymbol{\theta}$) restricted to the different sensitive groups. In this sense, a classifier $\mathcal{A}$ is said to be (almost) fair (w.r.t. θ) if θ is similar across the different subgroups $s\in\mathcal{S}$. The quantification of fairness (w.r.t. θ) is obtained by comparing these values, typically by means of differences or ratios. For instance, in the particular case of $\mathcal{S}=\{0,1\}$, a one-

dimensional fairness metric for DP would be the difference in acceptance rate AR across groups, namely $\theta(\boldsymbol{C}_0, \boldsymbol{C}_1) = AR(\boldsymbol{C}_1) - AR(\boldsymbol{C}_0)$. Other multidimensional metrics such as Equalized Odds (EO), that compares the true positive rate (TPR) and the false positive rate (FPR) across groups, would be quantified as $\boldsymbol{\theta}(\boldsymbol{C}_0, \boldsymbol{C}_1) = (\Delta TPR, \Delta FPR)(\boldsymbol{C}_0, \boldsymbol{C}_1) = (TPR(\boldsymbol{C}_1) - TPR(\boldsymbol{C}_0), FPR(\boldsymbol{C}_1) - FPR(\boldsymbol{C}_0))$. Although these examples focus on a binary sensitive attribute $\mathcal{S} = \{0, 1\}$, UM can be applied to multivalue $\mathcal{S}$, since any metric describing the behavior of $\mathcal{A}$ can be written as a function $\boldsymbol{\theta}(\boldsymbol{C}_1, \ldots, \boldsymbol{C}_r)$. See Verma et al. (2018) for an extensive review on fairness metrics.

2.2 Probabilistic Model

To accommodate multiple behavioral guarantees of a learning algorithm $\mathcal{A}$, we introduce UM, a method to derive the joint posterior probability distribution $P(\boldsymbol{\Theta})$ for a vector of user-defined metrics of interest $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$. In general, deriving a closed form expression for $P(\boldsymbol{\Theta})$ is challenging. That being the case, we propose to compute its empirical counterpart on a set of T samples obtained through a posterior hierarchical sampling procedure outlined in (1).

Remark 2.1. Although representing multi-dimensional distributions in a closed form is typically infeasible, there are circumstances where simplification is possible. Specifically, if the metrics in question are defined in terms of distinct counts of the CM, they become independent, allowing for the joint distribution to be expressed as the product of their individual marginal distributions. For example, the posterior distribution of the 2-dimensional metric EO can be computed as the product of the marginal distributions of ΔTPR and ΔFPR.

The UM approach is grounded in the probabilistic modeling of the CM, from which the estimation of performance metrics $\boldsymbol{\Theta}$ could be derived. Following Caelen (2017), we propose to adopt a Bayesian framework on the CM, assuming that its values are drawn from a multinomial distribution, denoted by $\text{MLT}(\cdot;\cdot)$. More precisely, for each group $s \in \mathcal{S}$, we consider $\boldsymbol{C}_s = (TP_s, TN_s, FP_s, FN_s) \sim \text{MLT}(N_s; \boldsymbol{\pi}_s)$. The multinomial parameter $\boldsymbol{\pi}_s = (\pi_{TP_s}, \pi_{TN_s}, \pi_{FP_s}, \pi_{FN_s}) \in [0, 1]^4$, which conforms a simplex, is assumed to follow a Dirichlet distribution, as it is the conjugate of the multinomial distribution. Hence, starting from the *prior* distribution $\boldsymbol{\pi}_s \sim \text{Dir}(\boldsymbol{\alpha}_s) = \text{Dir}((\alpha_1^s, \alpha_2^s, \alpha_3^s, \alpha_4^s))$, by Bayesian inference, the *posterior* distribution of the multinomial parameter will be $\boldsymbol{\pi}_s | \boldsymbol{C}_s \sim \text{Dir}((\alpha_1^s + TP_s, \alpha_2^s + TN_s, \alpha_3^s + FP_s, \alpha_4^s + FN_s))$. In other words, the CM for each sensitive group $\boldsymbol{C}_s$ is assumed to conform to a multinomial distribution, with the parameter $\boldsymbol{\pi}_s$ being adjusted based on $\boldsymbol{C}_s$. This serves as the fundamental technique to generate joint posterior distributions of fairness and performance metrics, which can then be utilized to develop uncertainty-aware algorithmic evaluation procedures.

It is worth noting that there is no supporting evidence for elevated values in specific entries of the CM, leading us to adopt *non-informative* priors. Besides, the influence of the chosen prior varies among different protected groups when the number of instances differs significantly across those groups. As the number of instances decreases, the effect of the prior becomes more pronounced. To address this issue, we employed a uniform prior $\boldsymbol{\pi}_s \sim \text{Dir}(\boldsymbol{\alpha}_s) = \text{Dir}((1, 1, 1, 1)),\ \forall s \in \mathcal{S}$. Supplementary experiments in Appendix 15.3 demonstrate that selecting alternative common non-informative priors has minimal impact on the final outcomes.

Once the posterior distribution of the CM is defined, it is possible to assess the uncertainty-aware fairness and performance guarantees of a learning algorithm by means of a hierarchical sampling procedure. Using the test instances for each group s, the trained model makes predictions, which are then used to obtain the empirical CM denoted by $\hat{\boldsymbol{C}}_s$. This matrix is subsequently utilized to update the distribution of $\boldsymbol{\pi}_s$. Then, we calculate the samples $t = 1, \ldots, T$ that approximate the distribution $P(\boldsymbol{\Theta})$ by drawing a sample (i) $\boldsymbol{\pi}_s^t$ of the multinomial parameter from its corresponding posterior Dirichlet distribution, which serves as the basis for the probabilistic model of the CM. Following this, (ii) a sample $\boldsymbol{C}_s^t$ is drawn from the distribution of the CM, enabling us to (iii) compute the values of any metric θ_i^t based on $\boldsymbol{C}_s^t$.

$$\begin{cases} \text{(i)} & \boldsymbol{\pi}_s^t \sim P(\boldsymbol{\pi}_s | \hat{\boldsymbol{C}}_s), \quad \text{for} \quad s = 1, \ldots, r \\ \text{(ii)} & \boldsymbol{C}_s^t \sim P(\boldsymbol{C}_s | \boldsymbol{\pi}_s^t), \quad \text{for} \quad s = 1, \ldots, r \\ \text{(iii)} & \theta_i^t(\boldsymbol{C}_1^t, \ldots, \boldsymbol{C}_r^t), \quad \text{for} \quad i = 1, \ldots, M \end{cases} \tag{1}$$

In hold-out evaluation, where data is randomly separated into two independent sets for training and evaluation, we can straightforwardly derive the posterior distributions by plugging the obtained CM into (1). However, in the case of K-fold CV, a common evaluation framework for limited data scenarios involving resampling, we obtain K CMs to estimate the posterior distributions. When applying (1) to the CM from each fold, we obtain K distinct joint posterior distributions. Nevertheless, the ultimate objective is to derive a single joint posterior distribution that accurately characterizes the learning algorithm's assessments. Merely averaging these posterior distributions would yield inaccurate uncertainty estimates due to correlations among the K results. Therefore, addressing this correlation is crucial to present a more

precise representation of the algorithm's uncertainty-aware performance. In the following section, we detail how we address these correlations to achieve a more accurate representation of the algorithm's uncertainty-aware performance.

2.3 K-Fold Cross-Validation

Inspired by Wang et al. (2019), we propose to address this issue by introducing the concept of *effective* confusion matrix of a sequence of correlated matrices. This matrix is such that the expectation, denoted by $\mathbb{E}$, and variance of any function of it is the same as that obtained from the correlated matrices.

Definition 2.2. Let $\boldsymbol{C}^{(1)}, \ldots, \boldsymbol{C}^{(K)}$ be a sequence of correlated confusion matrices. A matrix $\boldsymbol{C}^e$ is said to be effective if for any function f such that $\psi = \mathrm{Var}(f(C^{(k)}))$, $\forall k$, and $\rho = \mathrm{Corr}(f(C^{(k)}), f(C^{(k')}))$, $\forall k \neq k'$, it holds that

$$\mathrm{Var}(f(\boldsymbol{C}^e)) = \frac{\psi(1 + (K-1)\rho)}{K}.$$

The following proposition provides an *effective* CM for a K-fold CV.

Proposition 2.3. *Let $\boldsymbol{C}_s^{(k)}$ be the CM obtained in the k-th train/test split configuration of the K-fold CV process for subgroup s, for $k \in \{1, ..., K\}$. Then, the following matrix:*

$$\boldsymbol{C}_s^e = \frac{1}{1 + (K-1)\rho_s} \sum_{k=1}^{K} \boldsymbol{C}_s^{(k)}, \tag{2}$$

is a effective confusion matrix for the K-fold CV, where ρ_s denotes the correlation between the results obtained for every train/test configuration in one K-fold cross-validation for subgroup s.

The proof of this proposition can be found in Appendix 9. The true value of the correlation ρ_s is unknown and varies for different problems, thus accurately quantifying this parameter constitutes one of the primary challenges of this approach. Several approaches have already been proposed in the literature. For instance, Nadeau et al. (2003) propose to approximate it as $\rho_s = 1/K$, which is accurate when the Vapnik–Chervonenkis (VC) dimension of the algorithms is not too large compared to the size of the training set, or for algorithms that are robust to perturbations in the training set. Other works, such as the one by Wang et al. (2019) assume that $\rho_s \in [0, 1/K]$ (for $K = 2$, in their case). Even if such approximations are easy to compute, their main limitation is that they assume ρ_s is equal for all algorithms and all the subgroups. In other words, they assume the unlikely case that every algorithm has equal complexity and generalization capability. Indeed, in Figure 1 we clearly observe that different algorithms have different stabilities with respect to data partition. In order to overcome such drawback we propose an alternative complexity and subgroup-aware method to estimate the correlation ρ and its corresponding joint posteriors, described in the following section. Once the correlation is quantified, the joint posterior distributions can be calculated for the K-fold CV procedure using the hierarchical process from (1) for the effective confusion matrix $\boldsymbol{C}_s^e$.

2.3.1 Complexity and Subgroup-Dependent Approximation of the Correlation

Consider a reference algorithm $\mathcal{A}_0$ in the particular setting under consideration, with correlation ρ_s^0 (e.g. SVM, whose ρ_s^0 can be approximated as $1/K$ as studied in Nadeau et al. (2003)). Then, the estimated correlation ρ_s^1 of any algorithm $\mathcal{A}_1$ under a K-fold CV framework is described with respect to the reference ρ_s^0 by:

$$\rho_s^1 = \frac{(r_s - 1) + r_s(K-1)\rho_s^0}{K-1}, \tag{3}$$

where $r_s \equiv r(\mathcal{A}_0, \mathcal{A}_1, s) = \frac{\mathrm{Var}[\bar{\theta}_{K,\mathcal{A}_0,s}]}{\mathrm{Var}[\bar{\theta}_{K,\mathcal{A}_1,s}]}$ is the ratio between the variances of both algorithms on a given metric θ in a K-fold CV, for subgroup s.

We note that those variances cannot be directly calculated and need to be approximated. Nonetheless, Nadeau et al. (2003) propose an ultra-conservative overestimation of the variance from which the ratio $r_{over}(\mathcal{A}_0, \mathcal{A}_1, s) = \frac{\mathrm{Var}_{over}[\bar{\theta}_{K,\mathcal{A}_0,s}]}{\mathrm{Var}_{over}[\bar{\theta}_{K,\mathcal{A}_1,s}]}$ can be concluded. We refer to Appendix 10 for a detailed explanation of Equation (3) and the variance overestimation procedure. Here, we assume such an overestimation is proportional for all the methods, that is, $r(\mathcal{A}_0, \mathcal{A}_1, s) = \frac{\mathrm{Var}[\bar{\theta}_{K,\mathcal{A}_0,s}]}{\mathrm{Var}[\bar{\theta}_{K,\mathcal{A}_1,s}]} \approx r_{over}(\mathcal{A}_0, \mathcal{A}_1, s)$. Furthermore, note that this approximation is given in terms of the reference correlation ρ_0, for which generally a good approximation or a range of possible values is available as mentioned above (typically $\rho_s^0 = 1/K$ or $\rho_s^0 \in [0, 1/K]$). Thus, Equation (3) provides not only a pointwise estimation of ρ_s, but also an upper bound.

3 ALGORITHMIC COMPARISON

In this section we propose a new criterion for the comparison of two different algorithms, say $\mathcal{A}$ and $\mathcal{B}$, in terms of their performance and fairness guarantees, described respectively through the sequence of M different metrics $\boldsymbol{\Theta}_\mathcal{A}$ and $\boldsymbol{\Theta}_\mathcal{B}$. For this, we denote by $\delta(\Delta\boldsymbol{\Theta})$ the density function of $\Delta\boldsymbol{\Theta} = \boldsymbol{\Theta}_\mathcal{A} - \boldsymbol{\Theta}_\mathcal{B}$ the difference of the metrics of the two algorithms.

In order to compare $\mathcal{A}$ and $\mathcal{B}$, we generalize the notion of Region of Practical Equivalence (RoPE; Benavoli

et al., 2017, Kruschke, 2014) as the volume around the origin that will represent the values of $\Delta\boldsymbol{\Theta}$ that are considered negligibly indifferent from $\mathbf{0} \in \mathbb{R}^M$. It is important to note that, in a binary setup, fairness metrics are usually measured by differences in performance in the favored group minus that of the unfavored group, with 0 being the ideal value and 1 the worst. By contrast, the sense for measuring accuracy is the opposite. Therefore, for these differences to be in the same direction, we shall consider $-\theta_j$, for any fairness metric θ_j and some index $j \in \{1, \ldots, M\}$. In this way, for comparison purposes, we can treat fairness and performance metrics interchangeably.

Definition 3.1. Given $\boldsymbol{\varepsilon} \in \mathbb{R}^M_+$, the Region of Practical Equivalence (RoPE) of size $\boldsymbol{\varepsilon}$ is:
$RoPE(\boldsymbol{\varepsilon}) = \{\mathbf{x} \in \mathbb{R}^M : |x_i| \leq \varepsilon_i,\ i = 1, \ldots, M\}$.

The size of the RoPE will depend on the parameter $\boldsymbol{\varepsilon}$, whose values will be set according to the application domain. Then, the comparison between $\mathcal{A}$ and $\mathcal{B}$ is based in the relative position between $\Delta\boldsymbol{\Theta}$, or its highest density region (HDR) (Hyndman, 1996), and the RoPE (see Figure 3). Mainly, we are interested in estimating the posterior probabilities that:

(a) $\mathcal{A}$ and $\mathcal{B}$ are practically equivalent, meaning that they behave similarly in every objective. This refers precisely to the probability that $\Delta\boldsymbol{\Theta}$ is included in the RoPE; that is, $P(\mathcal{A} \approx \mathcal{B}) = P(\Delta\boldsymbol{\Theta} \in RoPE(\boldsymbol{\varepsilon}))$
$= \int_{-\varepsilon_1}^{\varepsilon_1} \cdots \int_{-\varepsilon_M}^{\varepsilon_M} \delta(\Delta\boldsymbol{\Theta})\, d\theta_1 \cdots d\theta_M$.

(b) $\mathcal{A}$ practically outperforms $\mathcal{B}$, meaning that $\mathcal{A}$ is at least better than $\mathcal{B}$ in one of the objectives but they are equivalent in the rest of them; i.e., $P(\mathcal{A} >> \mathcal{B}) =$
$P(\Delta\boldsymbol{\Theta} \in RoPE(\boldsymbol{\varepsilon})^c \cap (-\varepsilon_1, +\infty) \times \cdots \times (-\varepsilon_M, +\infty)) =$
$\int_{-\varepsilon_1}^{\infty} \cdots \int_{-\varepsilon_M}^{\infty} \delta(\Delta\boldsymbol{\Theta})\, d\theta_1 \cdots d\theta_M - P(\mathcal{A} \approx \mathcal{B})$.

(c) $\mathcal{B}$ practically outperforms $\mathcal{A}$; $P(\mathcal{B} >> \mathcal{A}) =$
$P(\Delta\boldsymbol{\Theta} \in RoPE(\boldsymbol{\varepsilon})^c \cap (-\infty, \varepsilon_1) \times \cdots \times (-\infty, \varepsilon_M)) =$
$\int_{-\infty}^{\varepsilon_1} \cdots \int_{-\infty}^{\varepsilon_M} \delta(\Delta\boldsymbol{\Theta})\, d\theta_1 \cdots d\theta_M - P(\mathcal{A} \approx \mathcal{B})$.

Analogously, it is possible to calculate the probabilities of all the other potential events (such as $\mathcal{A}$, resp. $\mathcal{B}$, outperforming $\mathcal{B}$, resp. $\mathcal{A}$, only in a subset of the objectives, while being outperformed by $\mathcal{B}$, resp. $\mathcal{A}$, in the others) as the probability of $\Delta\boldsymbol{\Theta}$ within the corresponding area.

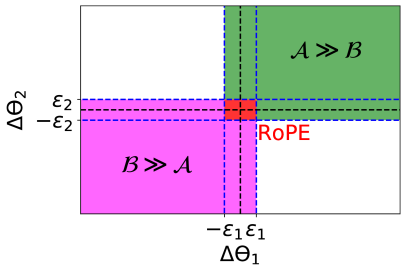


Figure 3: Display of a general bi-objective comparison of methods $\mathcal{A}$ and $\mathcal{B}$ w.r.t. metrics $\boldsymbol{\Theta} = (\Theta_1, \Theta_2)$.

4 RELATED WORK

On the assessment of fairness results. Despite the vast amount of work that has been developed in the field of algorithmic fairness, there is no consensus on the optimal assessment of fairness results. In most cases, the fairness and accuracy of various models are reported *separately* based on average values of repeated experiments (Qian et al., 2021), but the information provided is rather limited and oftentimes unstable (Friedler et al., 2019). Furthermore, experimental results are shown for specific choices of accuracy and fairness trade-offs that are not particularly well justified. A fair sample from the Pareto frontier is really what is needed to be able to confidently make statements about how different approaches compare. Agarwal et al. (2018) produced the convex envelope of the classifiers obtained on training data at various accuracy–fairness trade-offs. Ji et al. (2020) propose to use unlabeled data to get better estimates of fairness metrics, whereas Romano et al. (2020) make use of randomization tests. Some of the assessment schemes are context dependent and are only suitable to draw conclusions in those particular scenarios. Instead, we propose to assess the results in a context-independent manner by means of posterior distributions, similar to the work by Ji et al. (2020), but further extending their approach in two ways: on the one hand, with the possibility of defining a N-dimensional joint posterior distribution of any combination of performance and fairness metrics, and not limiting the analysis to marginal distributions; and on the other hand, by allowing more complex frameworks to evaluate learning algorithms, such as K-fold CV, where K correlated results are obtained.

On Bayesian methods in fairness. The incorporation of uncertainty in the fairness-aware context has mainly been concentrated on model uncertainty (Dimitrakakis et al., 2019; Foulds et al., 2020), considering a probabilistic distribution over model parameters. This concern has most popularly been addressed by means of Bayesian Neural Networks (BNNs) (Bhatt et al., 2021) recently, treating model weights as random variables whose probability distributions are updated by means of Bayesian inference. Several ensemble-based approaches have also been proposed (Foulds et al., 2020). The uncertainty modeled in those cases is related to the confidence of the algorithms on their predictions. However, our work is different in that it instead models the uncertainty inherent in the performance and fairness metrics used to report the behavior of the algorithms.

5 EXPERIMENTS

In this section, we present numerical experiments to validate UM's effectiveness and highlight its primary advantages in fairness-aware contexts over conventional evaluation methods. For these experiments, we opted for the German Credit dataset (Dua et al., 2017), which possesses two beneficial characteristics for our analysis, primarily due to its moderate size: (1) it permits the computation of numerous 10-fold CVs in a manageable time frame, enabling the estimation of a 'ground truth', and (2) the dataset's limited scale naturally introduces increased uncertainty into the metrics under evaluation. Specifically, we establish the ground truth through empirical probabilities obtained from 10,000 distinct 10-fold CV procedures.

We carry out a series of experiments to validate several key aspects. Firstly, we assess how our proposed estimation of the correlation (ρ_s) enables complexity-aware uncertainty estimations (Section 5.1). Secondly, we investigate the most likely uncertainty estimations (Section 5.2) and the worst-case uncertainty estimations (Section 5.3) across a broad range of algorithmic comparison scenarios. We present additional experiments in Appendix 14 that demonstrate how incorporating uncertainty can alter the conclusions drawn from some benchmark published results. A comprehensive description of the specific algorithms and the dataset is available in Appendix 11.

5.1 Best Approximation for ρ_s

First, we compare the existing alternatives to approximate ρ_s to find the most accurate strategy. We examine four alternatives, consisting of two state-of-the-art approaches that are not influenced by model complexity (1-2), along with two complexity-dependent approaches proposed in this study (3-4): (1) $\rho_s = 1/K$ (Nadeau et al., 2003; Benavoli et al., 2017); (2) $\rho_s \in [0, 1/K]$ (Wang et al., 2019); (3) $\rho_s = \rho_{rel}$: the relative ρ_s assuming that $\rho_s^0 = 1/K$ and $\mathcal{M}_0$ = SVM (non-linear); and (4) $\rho_s = \rho_{rel}^{\uparrow}$: the relative ρ_s assuming that $\rho_s^0 \in [0, 1/K]$ and $\mathcal{M}_0$ = SVM (non-linear). When assuming that ρ_s belongs to a specific range $[a, b]$, the estimation of the *effective* CM involves calculating the average contribution of each potential correlation value., i.e., $\boldsymbol{C}_s^e = \frac{1}{b-a}\int_a^b \frac{1}{1+(K-1)\cdot\rho_s} d\rho_s \sum_{k=1}^K \boldsymbol{C}_s^{(k)}$.

The strategy that yields the posterior distribution with the highest alignment w.r.t. the empirical distribution of repeated 10-fold CV results is considered as the optimal approach for approximating ρ_s. This agreement is measured by the proportion of repeated 10-fold CV results that fall within the 95% HDR (%RES) of the posterior distribution. Among the approaches that have equivalent agreement, we will favor the ones that derive the narrowest posterior, measured by the area of the 95% HDR. For each strategy to estimate the correlation, we repeat the experiment over 10,000 initial 10-fold CV configurations to derive the posteriors and report the averaged results. Further details about the experimental setup can be found in Appendix 11.

The results, presented in Table 1, indicate that the methods we suggest for estimating correlation (3-4), which take into account model complexity, yield results similar to those obtained with approaches (1-2) when the classifier's stability closely resembles that of SVM. Notably, our approach delivers the most accurate uncertainty estimations, surpassing the estimations of (1-2), particularly when the classifier's stability significantly deviates from that of SVM. For a more comprehensive version of the table that includes a broader range of methods, please refer to Appendix 12.1.

5.2 Most Probable Case

This section evaluates the effectiveness of UM for the 10-fold CV partition whose value of the performance difference (among the sample of 10,000 CV results) is the closest to the average value within the empirical distribution of differences in terms of Euclidean distance. We explore two distinct algorithmic evaluation scenarios: (a) when two methods exhibit similar stability with respect to data splits, and (b) when the methods display significantly different stability levels. For scenario (a) we utilize the pre-processing method proposed by Feldman et al. (2015) in conjunction with SVM, and the in-processing method by Donini et al. (2018) (see first row of Figure 2). In scenario (b) we employ the post-processing method introduced by Hardt et al. (2016) combined with Linear SVM, along with the in-processing method by Donini et al. (2018) (see second row of Figure 2). We specifically examine how well the framework estimates the probabilities of the following events: $\mathcal{A}$ practically outperforms $\mathcal{B}$ ($P(\mathcal{A} >> \mathcal{B})$), $\mathcal{B}$ practically outperforms $\mathcal{A}$ ($P(\mathcal{A} << \mathcal{B})$), $\mathcal{A}$ and $\mathcal{B}$ are practically equivalent ($P(\mathcal{A} \approx \mathcal{B})$), $\mathcal{A}$ practically outperforms $\mathcal{B}$ in accuracy but is outperformed by $\mathcal{B}$ in fairness ($P(\mathcal{A}_{acc}, \mathcal{B}_{fair})$) and $\mathcal{B}$ practically outperforms $\mathcal{A}$ in accuracy but is outperformed by $\mathcal{A}$ in fairness ($P(\mathcal{B}_{acc}, \mathcal{A}_{fair})$). For these experiments, we consider the correlation approximations discussed in the previous section, employ a RoPE with dimensions (0.01, 0.01), and adopt the fairness notion of EOp. The results are presented in Table 2, showcasing the effectiveness of the UM approach in both scenarios. These findings affirm UM's capacity to deliver accurate event probability estimates, thereby

Table 1: Evaluation of the Different Approximations for ρ_s. The area of the 95% HDR of the posterior distribution and the proportion of 10,000 different 10-fold CV results it encloses (%RES) for different algorithms (rows) and approximations (1-4) of ρ_s (columns). An extended version can be found in Appendix 12.1.

METHOD	$\rho_s = 1/K$ AREA	% RES	$\rho_s \in [0, 1/K]$ AREA	% RES	$\rho_s = \rho_{rel}$ AREA	% RES	$\rho_s = \rho_{rel}^{\uparrow}$ AREA	% RES
SVM	0.0201	100.0	0.0148	99.99	0.0201	100.0	0.0148	99.99
Kamiran et al. (2012) + SVM	0.0201	100.0	0.0149	99.96	0.0196	99.95	0.0148	99.64
FERM (Donini et al., 2018)	0.0107	100.0	0.0079	99.74	0.0110	100.0	0.0080	99.86
LFERM (Donini et al., 2018)	0.0109	99.97	0.0078	99.60	0.0129	100.0	0.0087	99.81
LR + Hardt et al. (2016)	0.0210	90.32	0.0156	82.08	0.0414	99.04	0.0232	92.77
SVM + Hardt et al. (2016)	0.0203	94.09	0.0151	88.67	0.0413	99.48	0.0230	95.84

Table 2: UM's Predictive Performance Applied to the Most Likely 10-fold CV Outcomes. Conventional 10-fold CV provides a single deterministic conclusion ($\mathcal{A}_{acc}, \mathcal{B}_{fair}$) for the first comparison and ($\mathcal{B} >> \mathcal{A}$) for the second, but lacks the ability to estimate probabilities. The probabilities assigned to these events by uncertainty-aware frameworks are shaded in gray. The ground truth probabilities are highlighted in **bold blue**, with the closest probabilities indicated in **bold**. By implementing UM, we can obtain probabilities that closely align with the true probabilities of the evaluation outcomes, thus providing a more accurate representation of the ground truth.

METHODS		$P(\mathcal{A} >> \mathcal{B})$	$P(\mathcal{B} >> \mathcal{A})$	$P(\mathcal{A} \approx \mathcal{B})$	$P(\mathcal{A}_{acc}, \mathcal{B}_{fair})$	$P(\mathcal{B}_{acc}, \mathcal{A}_{fair})$
$\mathcal{A}$: Feldman et al. (2015) + SVM vs. $\mathcal{B}$: Donini et al. (2018)	Ground truth	**0.04**	**0.15**	**0.01**	**0.80**	**0.00**
	SOTA ($\rho_s = 1/K$)	**0.01**	0.27	**0.00**	0.73	**0.00**
	SOTA ($\rho_s \in [0, 1/K]$)	0.00	0.23	0.00	0.77	**0.00**
	UM ($\rho_s = \rho_{rel}$)	0.00	0.20	**0.00**	**0.80**	**0.00**
	UM ($\rho_s = \rho_{rel}^{\uparrow}$)	**0.01**	**0.17**	**0.00**	0.82	**0.00**
$\mathcal{A}$: LinearSVM + Hardt et al. (2016) vs. $\mathcal{B}$: Donini et al. (2018)	Ground truth	**0.02**	**0.83**	**0.04**	**0.00**	**0.11**
	SOTA ($\rho_s = 1/K$)	0.08	0.64	**0.03**	0.03	0.22
	SOTA ($\rho_s \in [0, 1/K]$)	0.06	0.68	**0.03**	0.02	0.21
	UM ($\rho_s = \rho_{rel}$)	0.03	0.78	0.02	**0.01**	0.15
	UM ($\rho_s = \rho_{rel}^{\uparrow}$)	**0.02**	**0.83**	**0.03**	**0.01**	**0.11**

strengthening the trustworthiness of the conclusions drawn from employing UM compared to solely depending on 10-fold CV. Furthermore, these results emphasize the significance of integrating uncertainty quantification that takes classifier complexity into account. Additional results with different HDR and RoPE dimensions can be found in Appendix 12.2.

5.3 Worst-Case: Uncommon Events

This section assesses the performance of UM for the 10-fold CV partition (among the 10,000 CV partitions considered) where the performance difference leads to a conclusion that differs from what is implied by the average value within the empirical distribution and exhibits the highest deviation from that average value. We explore the same algorithmic evaluation scenarios, correlation approximations, RoPE dimensions and fairness notions discussed in the previous section, and the results are detailed in Table 3. Additional results with different HDR and RoPE dimensions, including scenarios without HDR and/or RoPE, can be found in Appendix 12.3. The outcomes reveal that relying solely on conventional 10-fold CV leads to incorrect conclusions. Nonetheless, UM effectively mitigates the issue: although the probability in favor of the 'incorrect' conclusion increases, it remains low and does not significantly surpass the probabilities of other events. Importantly, UM consistently prioritizes the event with the highest true probability. It's important to emphasize that this is not observed in scenario (b) with SOTA approximations to estimate the correlation. The discrepancy in results within scenario (b) can be attributed to the fact that SOTA approaches are generally more suitable for algorithms with stability levels similar to SVM. Consequently, in the case of $\mathcal{A}$, whose stability significantly deviates from that of SVM, SOTA methods prove inadequate for precisely estimating the correlation among $\mathcal{A}$'s outcomes.

6 CONCLUSION AND DISCUSSION

A wealth of literature underscores the effectiveness of Bayesian inference in dealing with non-repeatability issues, as it provides a more robust approach to reasoning in uncertain situations. Within the framework of Bayesian inference, we have introduced *Uncertainty Matters* (UM), a probabilistic approach designed to conduct uncertainty-aware multi-objective algorithmic comparisons. This includes, for instance, estimating uncertainty-aware disparities in performance and fairness guarantees between two algorithms. UM can be effectively applied in various algorithmic evaluation scenarios, such as hold-out and K-fold CV. Our

Table 3: Worst-Case Predictive Performance of UM. Conventional 10-fold CV suggests a single deterministic outcome of algorithmic comparison, whose true probability is negligible (**bold red**). Uncertainty-aware frameworks assign probabilities to these rare outcomes, highlighted in gray, and we identify the event with the highest true probability in **bold blue**. Even when applied to rare 10-fold CV results, UM consistently prioritizes the event with the highest true probability and refrains from assigning a high probability to an incorrect conclusion.

METHODS		$P(\mathcal{A} >> \mathcal{B})$	$P(\mathcal{B} >> \mathcal{A})$	$P(\mathcal{A} \approx \mathcal{B})$	$P(\mathcal{A}_{acc}, \mathcal{B}_{fair})$	$P(\mathcal{B}_{acc}, \mathcal{A}_{fair})$
$\mathcal{A}$: Feldman et al. (2015) + SVM vs. $\mathcal{B}$:Donini et al. (2018)	Ground truth	**0.04**	0.15	0.01	**0.80**	0.00
	SOTA ($\rho_s = 1/K$)	0.32	0.18	0.04	0.44	0.01
	SOTA ($\rho_s \in [0, 1/K]$)	0.30	0.16	0.04	0.50	0.00
	UM ($\rho_s = \rho_{rel}$)	0.31	0.09	0.04	0.56	0.00
	UM ($\rho_s = \rho^{\uparrow}_{rel}$)	**0.22**	0.10	0.04	**0.64**	0.00
$\mathcal{A}$: Linear SVM + Hardt et al. (2016) vs. $\mathcal{B}$:Donini et al. (2018)	Ground truth	**0.02**	**0.83**	0.04	0.00	0.11
	SOTA ($\rho_s = 1/K$)	0.36	0.32	0.05	0.13	0.13
	SOTA ($\rho_s \in [0, 1/K]$)	0.36	0.33	0.07	0.12	0.12
	UM ($\rho_s = \rho_{rel}$)	**0.10**	**0.56**	0.04	0.02	0.28
	UM ($\rho_s = \rho^{\uparrow}_{rel}$)	0.17	0.51	0.07	0.05	0.19

numerical experiments have demonstrated that the UM framework offers enhanced informativeness and a broader perspective on algorithmic evaluation since it captures the inherent instability in results.

Despite the strengths of this study, it is important to acknowledge several limitations. In scenarios where the 10-fold CV outcome suggests a rare event in algorithmic comparison, UM probabilities may not precisely reflect the true probabilities but will effectively prevent the derivation of erroneous conclusions with a high degree of confidence. Additionally, the main source of computational complexity in UM arises from estimating ρ_s, which necessitates performing $2 \times J$ iterations of 10-fold CV on a half-sized dataset (see Appendix 10 for details). Our empirical research has verified that for small datasets, consistently reliable results can be obtained by setting J to 5, whereas for larger datasets, $J = 1$ suffices. Furthermore, while the selection of the RoPE may have a minor influence on the likelihood of events, it does not alter the principal finding of the algorithmic comparison. The underlying concept of RoPE is to allow users to define a range around the null value that includes values considered practically equivalent to the null value. Moreover, with an increasing number of objectives (metrics), the probability of one method being dominated by another method significantly decreases and becomes almost negligible (for experiments with more than 2 objectives refer to Appendix 13). Consequently, with a significant number of objectives, all methods may become Pareto-optimal. Therefore, in reality, employing more than three objectives becomes impractical, even though there is no theoretical constraint in the UM framework.

Future research should prioritize the development of techniques to reduce outcome uncertainty in scenarios with limited sample sizes (see detailed discussions in Appendix 15). One promising direction is to investigate approaches that integrate unlabeled data with test data to enhance the reliability of outcomes. Furthermore, we aim to extend our research to encompass a broader range of real-world scenarios, including those with noisy or corrupted sensitive information in test data, as well as situations where sensitive information may be missing for some instances due to legal restrictions or individual choice. We are also interested in exploring other multi-objective comparison scenarios, such as, privacy-aware evaluations (Yeom et al., 2018).

Acknowledgements

We are sincerely grateful to the reviewers whose insightful comments and constructive feedback greatly enhanced the quality of this work. This research was funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. This work is supported by the European Research Council under the European Union's Horizon 2020 research and innovation programme Grant Agreement no. 851538 - BayesianGDPR, Horizon Europe research and innovation programme Grant Agreement no. 101120763 - TANGO. This work is also supported by the Basque Government under grant IT1504-22 and through the BERC 2022-2025 program; by the Spanish Ministry of Science and Innovation under the grants PID2022-137442NB-I00 and PID2021-128314NB-I00, and through BCAM Severo Ochoa accreditation CEX2021-001142-S / MICIN / AEI / 10.13039/501100011033.

References

Besse, Philippe et al. (2021). "A survey of bias in machine learning through the prism of statistical parity". In: *The American Statistician*, pp. 1–11.

Kruschke, John (2014). "Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan". In.

Agarwal, Alekh et al. (2018). "A reductions approach to fair classification". In: *International Conference on Machine Learning*. PMLR, pp. 60–69.

Aghbalou, Anass, Anne Sabourin, and François Portier (2023). "On the bias of K-fold cross validation with stable learners". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 3775–3794.

Bashtannyk, David M and Rob J Hyndman (2001). "Bandwidth selection for kernel conditional density estimation". In: *Computational Statistics & Data Analysis* 36.3, pp. 279–298.

Benavoli, Alessio et al. (2017). "Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis". In: *The Journal of Machine Learning Research* 18.1, pp. 2653–2688.

Besse, Philippe et al. (2018). "Confidence intervals for testing disparate impact in fair learning". In: *arXiv preprint arXiv:1807.06362*.

Bhatt, Umang et al. (2021). "Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413.

Caelen, Olivier (2017). "A Bayesian interpretation of the confusion matrix". In: *Annals of Mathematics and Artificial Intelligence* 81.3, pp. 429–450.

Dimitrakakis, Christos et al. (2019). "Bayesian fairness". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 509–516.

Donini, Michele et al. (2018). "Empirical risk minimization under fairness constraints". In: *Advances in Neural Information Processing Systems* 31.

Dua, Dheeru and Casey Graff (2017). *UCI Machine Learning Repository*. URL: `http://archive.ics.uci.edu/ml`.

Dwork, Cynthia et al. (2015). "The reusable holdout: Preserving validity in adaptive data analysis". In: *Science* 349.6248, pp. 636–638.

Feldman, Michael et al. (2015). "Certifying and removing disparate impact". In: *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268.

Foulds, James R. et al. (2020). "Bayesian Modeling of Intersectional Fairness: The Variance of Bias". In: *Proceedings of the 2020 SIAM International Conference on Data Mining, SDM 2020, Cincinnati, Ohio, USA, May 7-9, 2020*. Ed. by Carlotta Demeniconi and Nitesh V. Chawla. SIAM, pp. 424–432. DOI: `10.1137/1.9781611976236.48`. URL: `https://doi.org/10.1137/1.9781611976236.48`.

Friedler, Sorelle A et al. (2019). "A comparative study of fairness-enhancing interventions in machine learning". In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 329–338.

Goutte, Cyril and Eric Gaussier (2005). "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation". In: *European conference on information retrieval*. Springer, pp. 345–359.

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

Hyndman, Rob J (1996). "Computing and graphing highest density regions". In: *The American Statistician* 50.2, pp. 120–126.

Hyndman, Rob J, David M Bashtannyk, and Gary K Grunwald (1996). "Estimating and visualizing conditional densities". In: *Journal of Computational and Graphical Statistics* 5.4, pp. 315–336.

Ji, Disi, Padhraic Smyth, and Mark Steyvers (2020). "Can I trust my fairness metric? assessing fairness with unlabeled data and bayesian inference". In: *Advances in Neural Information Processing Systems* 33, pp. 18600–18612.

Kamiran, Faisal and Toon Calders (2012). "Data preprocessing techniques for classification without discrimination". In: *Knowledge and information systems* 33.1, pp. 1–33.

Kruschke, John K and Torrin M Liddell (2018). "The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective". In: *Psychonomic bulletin & review* 25.1, pp. 178–206.

Nadeau, Claude and Yoshua Bengio (2003). "Inference for the Generalization Error". In: *Machine Learning* 52.3, pp. 239–281.

Qian, Shangshu et al. (2021). "Are My Deep Learning Systems Fair? An Empirical Study of Fixed-Seed Training". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 30211–30227. URL: `https://proceedings.neurips.cc/paper/2021/file/fdda6e957f1e5ee2f3b311fe4f145ae1-Paper.pdf`.

Romano, Yaniv, Stephen Bates, and Emmanuel Candes (2020). "Achieving equalized odds by resampling sensitive attributes". In: *Advances in Neural Information Processing Systems* 33, pp. 361–371.

Samworth, RJ and MP Wand (2010). "Asymptotics and optimal bandwidth selection for highest density region estimation". In: *The Annals of Statistics* 38.3, pp. 1767–1792.

Verma, Sahil and Julia Rubin (2018). "Fairness definitions explained". In: *FairWare'18: IEEE/ACM International Workshop on Software Fairness*. Gothenburg, Sweden: ACM, pp. 1–7. ISBN:

9781450357463. DOI: 10.1145/3194770.3194776. URL: https://doi.org/10.1145/3194770.3194776.

Waltman, Marijn T (2014). "An Algorithm for Approximating the Highest Density Region in d-Space". In.

Wang, Ruibo and Jihong Li (2019). "Bayes test of precision, recall, and F1 measure for comparison of two natural language processing models". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4135–4145.

Yeom, Samuel et al. (2018). "Privacy risk in machine learning: Analyzing the connection to overfitting". In: *IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282.

Zafar, Muhammad Bilal et al. (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.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes, comprehensive descriptions are available in Appendix 11**.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes, we discuss the computational cost of the method in Section 6.**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes, the implementation details can be found in Appendix 11. We also include the code to reproduce the main experimental results as a .py document in the supplementary material.**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes, see Section 2.**
   (b) Complete proofs of all theoretical results. **Yes, the complete proofs can be found in Appendix 9.**
   (c) Clear explanations of any assumptions. **Yes, see Section 2 and Appendix 9.**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes, the code for the paper is publicly available at** `https://github.com/abarrainkua/UncertaintyMatters`.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes, all the details regarding the experimental evaluation can be found in Appendix 11.**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes, we do define the particular statistic under consideration but recommend employing the posterior distribution instead of error bars to measure the uncertainty and variability of the results.**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes, the implementation details can be found in Appendix 11.**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Yes, we cited the authors of all the datasets, methods and approaches mentioned and used.**
   (b) The license information of the assets, if applicable. **Yes, the implementation details can be found in Appendix 11.**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable.**
   (d) Information about consent from data providers/curators. **Not Applicable. We used publicly available datasets and code.**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable.**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable. Our experiments did not require external participants.**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board

(IRB) approvals if applicable. **Not Applicable. Our experiments did not require external participants.**

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable. Our experiments did not require external participants.**

Supplementary Material for Uncertainty Matters: Stable Conclusions Under Unstable Assessment of Fairness Results

7 UNSTABLE 10-FOLD CV RESULTS OF FAIRNESS-ENHANCING INTERVENTIONS

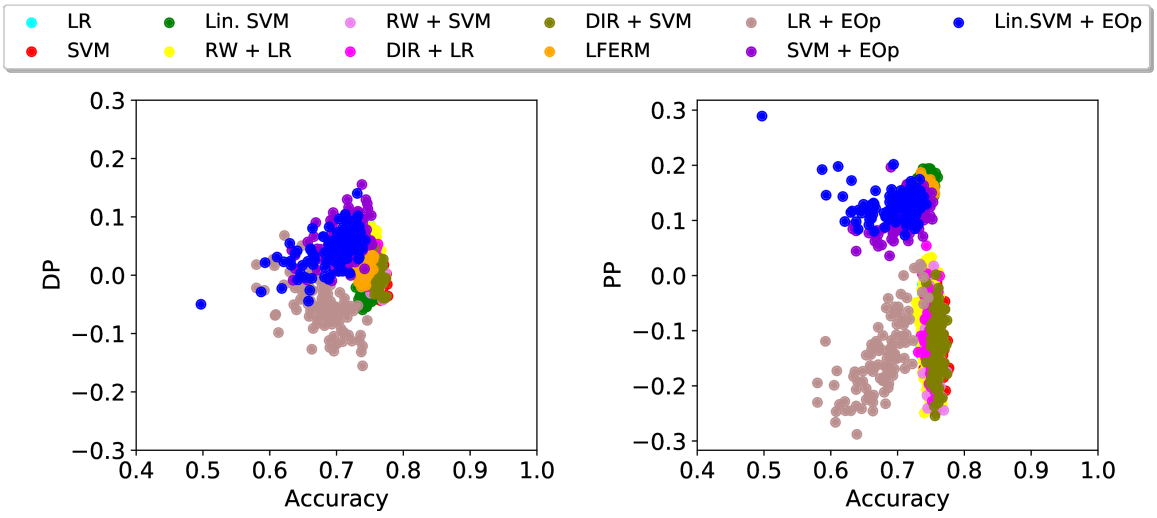


Figure 7.1: Instability of 10-Fold CV Results. The average values resulting from 100 different 10-fold cross-validation for accuracy and fairness, measured by means of (left) *demographic parity* (DP) and (right) *predictive parity* (PP). Each point refers to one averaged results of the 10-fold cross-validation, and the colors refer to different benchmark fairness-enhancing methods. These results were obtained using the German dataset with age as the sensitive attribute. This figure presents equivalent information to Figure 1 in the main text, albeit with distinct fairness metrics.

8 MULTI-CLASS CASE

In this section, we clarify the adaptation of our methodology to encompass the multi-class scenario. Within this context, we encounter two primary challenges: (a) determining the suitable formulation of the confusion matrix for each subgroup when dealing with more than two classes, and (b) devising a framework to quantify fairness metrics within this context. We provide solutions to both challenges in the following discussion.

The CM in Multi-Class Classification. Let us contemplate a scenario where the classification task at hand encompasses a total of J classes, meaning that $|\mathcal{Y}| = J$. In this case, we will consider one 2×2 confusion matrix for each subgroup and class label pair, i.e., $\boldsymbol{\mathcal{C}}_{s,j} = (TP_{s,j}, TN_{s,j}, FP_{s,j}, FN_{s,j})$, where $s = 1, ..., r$ and $j = 1, ..., J$, which will be defined as:

$$\begin{pmatrix} \sum_{i=1}^{N} \mathbb{I}[y_i = j, \hat{y}_i = j, s] & \sum_{i=1}^{N} \mathbb{I}[y_i \neq j, \hat{y}_i = j, s] \\ \sum_{i=1}^{N} \mathbb{I}[y_i = j, \hat{y}_i \neq j, s] & \sum_{i=1}^{N} \mathbb{I}[y_i \neq j, \hat{y}_i \neq j, s] \end{pmatrix}. \tag{4}$$

Fairness Metrics in Multi-Class Classification. Scant attention has been devoted to quantifying fairness infringements in the context of multi-class classification. In this work we propose three alternatives:

(i) Micro-averaging:

$$DP_{micro} = \max_{s,s'} \left| \frac{\sum_{j=1}^{J} TP_{s,j} + FP_{s,j}}{\sum_{j=1}^{J} TP_{s,j} + FP_{s,j} + TN_{s,j} + FN_{s,j}} - \frac{\sum_{j=1}^{J} TP_{s',j} + FP_{s',j}}{\sum_{j=1}^{J} TP_{s',j} + FP_{s',j} + TN_{s',j} + FN_{s',j}} \right| \tag{5}$$

$$DP_{micro} = \max_{s,s'} \left| AR_{micro,s} - AR_{micro,s'} \right| \tag{6}$$

$$EOp_{micro} = \max_{s,s'} \left| \frac{\sum_{j=1}^{J} TP_{s,j}}{\sum_{j=1}^{J} TP_{s,j} + FN_{s,j}} - \frac{\sum_{j=1}^{J} TP_{s',j}}{\sum_{j=1}^{J} TP_{s',j} + FN_{s',j}} \right| \tag{7}$$

$$EOp_{micro} = \max_{s,s'} \left| TPR_{micro,s} - TPR_{micro,s'} \right| \tag{8}$$

(ii) Macro-averaging:

$$DP_{macro} = \sum_{j=1}^{J} p(y=j) \left\{ \max_{s,s'} \left| \frac{TP_{s,j} + FP_{s,j}}{TP_{s,j} + FP_{s,j} + TN_{s,j} + FN_{s,j}} - \frac{TP_{s',j} + FP_{s',j}}{TP_{s',j} + FP_{s',j} + TN_{s',j} + FN_{s',j}} \right| \right\} \tag{9}$$

$$DP_{macro} = \sum_{j=1}^{J} p(y=j) DP_{y=j} \tag{10}$$

$$EOp_{macro} = \sum_{j=1}^{J} p(y=j) \left\{ \max_{s,s'} \left| \frac{TP_{s,j}}{TP_{s,j} + FN_{s,j}} - \frac{TP_{s',j}}{TP_{s',j} + FN_{s',j}} \right| \right\} \tag{11}$$

$$EOp_{macro} = \sum_{j=1}^{J} p(y=j) EOp_{y=j} \tag{12}$$

(iii) Macro-micro averaging:

$$DP = \max_{s,s'} \left| \sum_{j=1}^{J} p(y=j|s) \frac{TP_{s,j} + FP_{s,j}}{TP_{s,j} + FP_{s,j} + TN_{s,j} + FN_{s,j}} - \sum_{j=1}^{J} p(y=j|s') \frac{TP_{s',j} + FP_{s',j}}{TP_{s',j} + FP_{s',j} + TN_{s',j} + FN_{s',j}} \right| \tag{13}$$

$$DP = \max_{s,s'} \left| AR_{macro,s} - AR_{macro,s'} \right| \tag{14}$$

$$EOp = \max_{s,s'} \left| \sum_{j=1}^{J} p(y=j|s) \frac{TP_{s,j}}{TP_{s,j} + FN_{s,j}} - \sum_{j=1}^{J} p(y=j|s') \frac{TP_{s',j}}{TP_{s',j} + FN_{s',j}} \right| \tag{15}$$

$$EOp = \max_{s,s'} \left| TPR_{macro,s} - TPR_{macro,s'} \right| \tag{16}$$

9 PROOFS OF SECTION 2

In order to proof Proposition 2.3 we first introduce the reader to an important lemma by Nadeau et al. (2003):

Lemma 9.1. *Let U^1,..., U^K be random variables with common mean β and the following covariance structure*

$$Var[U^k] = \delta \quad \forall k, \quad Cov[U^k, U^{k'}] = \gamma \quad \forall k \neq k'.$$

Let $\pi = \frac{\gamma}{\delta}$ be the correlation between U^k and $U^{k'}$ $(k \neq k')$. Let $\bar{U} = k^{-1}\sum_{k=1}^{K} U^k$ and $S_U^2 = \frac{1}{K-1}\sum_{k=1}^{K}(U^k - \bar{U})^2$ be the sample mean and the sample variance respectively. Then:

1. $Var[\bar{U}] = \gamma + \frac{(\delta-\gamma)}{K} = \delta\big(\pi + \frac{1-\pi}{K}\big)$.
2. *If the stated covariance structure holds for all K (with γ and δ not depending on K), then:*
   - $\gamma \geq 0$
   - $\lim_{K\to\infty} Var[\bar{U}] = 0 \iff \gamma = 0$
3. $E[S_U^2] = \delta - \gamma$

Proof of Proposition 2.3

Proof. To characterize the form of the effective confusion matrix, our goal is to characterize the elements of the *effective* CM, i.e., (TP^e, TN^e, FP^e, FN^e), that, when assuming our probabilistic approach for the CM, we obtain the same uncertainty-aware description of the classifier's as that derived from the correlated K-fold CV results. Our specific objective is to ensure that the expected value and variance of any function derived from the CM are consistent between the effective procedure and the correlated results.

Let $\theta = f(\boldsymbol{C}^e)$ be a statistic of interest that describes the behavior of the algorithm, $\hat{\theta}$ its mean value resulting from the K-fold cross-validation and θ^k its value for the k-th train/test configuration of the cross-validation. Since the partitioning is arbitrary, $\mathrm{Var}[\theta^k]$ will be similar across all k, and the covariance between the results of two different train/test configurations on the K-fold cross-validation will be similar for any two pairs, i.e. $\mathrm{Corr}[\theta^k, \theta^{k'}] = \rho$ for $k \neq k'$. Then, from Lemma 9.1 we know that:

$$\mathrm{Var}[\hat{\theta}] = \frac{\mathrm{Var}[\theta^k](1 + (K-1)\rho)}{K}$$

Let us assume the case where the metric θ refers to the *true positive rate* (TPR). Then:

$$\mathrm{Var}[\theta^k] = \mathrm{Var}\left[\frac{TP^k}{TP^k + FN^k}\right]$$

where $TP^k + FN^k$ refers to the number of positive labeled instances in the k-th fold. If we assume that we adopt an *stratified* strategy for the CV then, $TP^k + FN^k = n_+/K$, where n_+ refers to the number of positive labeled instances in the whole dataset. Thus,

$$\mathrm{Var}\left[\frac{TP^k}{TP^k + FN^k}\right] = \left(\frac{K}{n_+}\right)^2 \mathrm{Var}[TP^k]$$

Since we have described the probabilistic CM by means of a multinomial distribution, due to marginalization, its elements will follow a binomial distribution. From Goutte et al. (2005) we know that for fixed $TP^k + FN^k$, TP^k follows a binomial distribution with parameters r and n_+/K, that is, $TP^k | TP^k + FN^k \sim \mathrm{Bin}(r, n_+/K)$. Therefore, $\mathrm{Var}[TP^k] = r(r-1)\frac{n_+}{K}$ and:

$$\mathrm{Var}\left[\frac{TP^k}{TP^k + FN^k}\right] = \frac{Kr(1-r)}{n_+}$$

where $r = p(\hat{y} = 1|y = 1)$ represents the expected value of such probability of models that are trained with $\frac{(K-1)N}{K}$ instances. With that, the variance of the estimation resulting from the K-fold cross-validation would be:

$$\text{Var}[\hat{TPR}] = \frac{r(1-r)(1+(K-1)\rho)}{n_+} \tag{17}$$

At the same time, we define the elements of the effective confusion matrix, TP^e, FP^e, TN^e and FN^e, as random variables whose distribution is defined so that for any function of the confusion matrix, the variance of such function is equivalent to the one that would be derived from the actual K correlated CV results. Therefore, based on the effective CM, the variance of the expected value of TPR can be described as:

$$\text{Var}[\hat{TPR}] = \text{Var}\left[\frac{TP^e}{TP^e + FN^e}\right]$$

In this case $TP^e + FN^e$ is constant and $TP^e|TP^e + FN^e \sim \text{Bin}(r, TP^e + FN^e)$. Thus,

$$\text{Var}[\hat{TPR}] = \frac{r(1-r)}{TP^e + FN^e} \tag{18}$$

Comparing equations (17) and (18) we get:

$$\frac{r(1-r)}{TP^e + FN^e} = \frac{r(1-r)(1+(K-1)\rho)}{n_+}$$

and since $n_+ = \sum_{k=1}^{K}(TP^k + FN^k)$:

$$TP^e + FN^e = \frac{1}{1+(K-1)\rho}\sum_{k=1}^{K}(TP^k + FN^k)$$

From which we can define:

$$TP^e = \frac{1}{1+(K-1)\rho}\sum_{k=1}^{K}TP^k$$

$$FN^e = \frac{1}{1+(K-1)\rho}\sum_{k=1}^{K}FN^k$$

If we repeat the procedure with the metric FPR and assuming that the value of the correlation ρ is equal for all the metrics, we can derive equivalent formulations for the effective counts on FP and TN, eventually obtaining:

$$\boldsymbol{C}^e = \frac{1}{1+(K-1)\rho}\sum_{k=1}^{K}\boldsymbol{C}^{(k)} \tag{19}$$

The expected value of a given statistic will be the same across each fold since the partition is randomized. Such expectation refers to the expected value of a classifier trained with $\frac{N(K-1)}{K}$ training samples. Therefore, due to the linearity property of the expectation, the expected value obtained from the effective CM will be identical to that obtained from K-fold CV.

□

10 THE RELATIVE VALUE OF THE CORRELATION ρ

In this section we provide a more detailed explanation about how we can conclude the correlation of the results for a target method $\mathcal{A}_1$ for subgroup s, based on the correlation of a reference method $\mathcal{A}_0$ in that subgroup. Moreover, we explain the method proposed by Nadeau et al. (2003) to obtain the over-estimations of the variance which is used to compute the value of $r_s \equiv r(\mathcal{A}_0, \mathcal{A}_1, s) = \frac{\text{Var}[\bar{\theta}_{K,\mathcal{A}_0,s}]}{\text{Var}[\bar{\theta}_{K,\mathcal{A}_1,s}]}$ of our approximation.

From Lemma 9.1 we know that the variance of the expected value of a metric under a K-fold CV for subgroup s can be written as:

$$\mathrm{Var}[\bar{\theta}_{s,K}] = \frac{\mathrm{Var}[\theta_s^k](1+(K-1)\rho_s)}{K}$$

Thus, for two ML methods ($\mathcal{A}_0$ and $\mathcal{A}_1$) with correlations ρ_s^0 and ρ_s^1, the ratio of their variances is:

$$r(\mathcal{A}_0, \mathcal{A}_1, s) = \frac{\mathrm{Var}[\theta_{s,1}^k]}{\mathrm{Var}[\theta_{s,0}^k]} \cdot \frac{1+(K-1)\rho_s^1}{1+(K-1)\rho_s^0}$$

We assume the variance of the results obtained for each fold will have the same variance for each $k = 1, ..., K$ (Lemma 9.1), and will be similar for different methods (which has been supported by different experiments); that is, $\mathrm{Var}[\theta_{s,1}^k] \approx \mathrm{Var}[\theta_{s,0}^k]$ (note that, starting from the same $\mathrm{Var}[\theta_s^k]$, what makes the variance of the final metric to be different is the correlation). Thus:

$$r_s \equiv r(\mathcal{A}_0, \mathcal{A}_1, s) = \frac{1+(K-1)\rho_s^1}{1+(K-1)\rho_s^0}$$

From which we can conclude:

$$\rho_s^1 = \frac{(r_s - 1) + r_s(K-1)\rho_s^0}{K-1},$$

The value $r(\mathcal{A}_0, \mathcal{A}_1, s) = \frac{\mathrm{Var}[\bar{\theta}_{K,\mathcal{A}_0,s}]}{\mathrm{Var}[\bar{\theta}_{K,\mathcal{A}_1,s}]}$ cannot be estimated exactly. However, Nadeau et al. (2003) propose a method to overestimate the variance $\mathrm{Var}[\bar{\theta}_{s,K}]$. Assuming that such an overestimation is proportional for all the methods, that is, $r(\mathcal{A}_0, \mathcal{A}_1, s) = \frac{\mathrm{Var}[\bar{\theta}_{K,\mathcal{A}_0,s}]}{\mathrm{Var}[\bar{\theta}_{K,\mathcal{A}_1,s}]} \approx r_{over}(\mathcal{A}_0, \mathcal{A}_1, s)\frac{\mathrm{Var}_{over}[\bar{\theta}_{K,\mathcal{A}_0,s}]}{\mathrm{Var}_{over}[\bar{\theta}_{K,\mathcal{A}_1,s}]}$, we can obtain an approximate value of $r(\mathcal{A}_0, \mathcal{A}_1, s)$. With that, starting from a reference value ρ_s^0 we can obtain the value of the correlation for any method. *But, how do we obtain the overestimations of the variances?*

In what follows, we explain the approach by Nadeau et al., 2003 to overestimate the variance of a given performance metric. Particularly, they propose to obtain independent observations of such statistic. To obtain such independent measurements, the dataset must be split into two disjoint datasets D and D^c (where $D \cap D^c = \emptyset$) of size $\lfloor \frac{n}{2} \rfloor$ ($|D| = |D^c| = \lfloor \frac{n}{2} \rfloor$) (being n the size of the complete dataset). Let $\hat{\theta}_s$ and $\hat{\theta}_s^c$ be the values of the statistic of interest in group s when a K-fold CV is performed in D and D^c, respectively. Then, $\frac{1}{2}(\hat{\theta}_s - \hat{\theta}_s^c)^2$ is an unbiased estimate of $\mathrm{Var}_{over}[\bar{\theta}_{K,\mathcal{A},s}]$. The step of splitting the dataset into two disjoint blocks can be repeated J times, yielding the pairs $(\hat{\theta}_{s,j}, \hat{\theta}_{s,j}^c)$ for $j = 1, ..., J$. With that, the following unbiased estimation of $\mathrm{Var}_{over}[\bar{\theta}_{K,\mathcal{A},s}]$ is concluded:

$$\mathrm{Var}_{over}[\bar{\theta}_{K,\mathcal{A},s}] = \frac{1}{2J}\sum_{j=1}^{J}(\hat{\theta}_j - \hat{\theta}_j^c)^2 \tag{20}$$

Thus this approximation requires performing $2J$ additional half-sized 10-fold CV procedures. Nonetheless, low values of J provide stable result with respect to $r_{over}(\mathcal{A}_0, \mathcal{A}_1)$. In fact, Figure 10.1 suggests that, in the case of the German Credit dataset (composed of 1,000 instances), $J = 5$ already constitutes a good approximation. Furthermore, the value of J required to obtain an accurate estimation decreases considerably for increasing dataset size. Thus, for big enough datasets $J = 1$ is sufficient. Nonetheless, in such cases, the conventional approximation $\rho_s = 1/K$ becomes accurate.

11 IMPLEMENTATION DETAILS

11.1 Licenses

In our code base we use packages installed via the Python Package Index (PyPI). A table of the used packages and their associated licenses are available in Table 11.1.

11.2 Dataset

In the experimental section, we explore various algorithmic comparison scenarios for a specific classification task. Now, let's delve into the dataset utilized in these comparative scenarios.

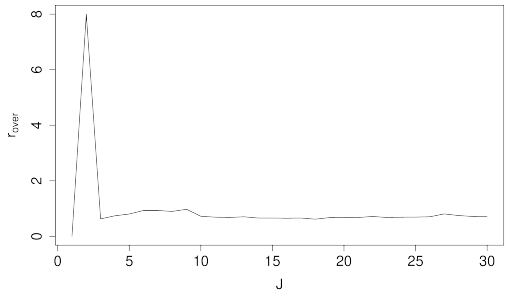


Figure 10.1: The Ratio Between the Overestimation of the Variances ($r_{over}(\mathcal{A}_0, \mathcal{A}_1)$) Between FERM ($\mathcal{A}_1$) of Donini et al. (2018) and Non-Linear SVM ($\mathcal{A}_0$). The value of $r_{over}(\mathcal{A}_0, \mathcal{A}_1)$ stabilizes for very low values for J.

Package	License
EthicML	GPL 3.0
AIF360	Apache License 2.0
Numpy	BSD 3-Clause
Seaborn	BSD 3-Clause
Sklearn	BSD 3-Clause

Table 11.1: Licenses of Packages Installed from the Python Package Index (PyPi).

German Credit. The German Credit dataset collects information about several individuals created from a German bank's data from 1994. It contains details about the socioeconomic situation of individuals: namely its employment, housing, savings, etc. Besides, the set of features includes some sensitive information as well: the gender and age. In this classification task, the objective is to predict whether an individual should obtain a good or bad credit score. This dataset is considerably smaller than the previous, containing only 1,000 instances and 20 features, and it is publicly available in the UCI repository. In our experiments, age has been considered as the sensitive attribute, and we have divided the instances into two groups based on whether their age is greater than 25 or less than or equal to 25. This binarization of age results in two sensitive groups: one group consists of individuals aged 25 or younger, representing 19% of the instances, and the other group consists of individuals older than 25. We pre-process the dataset as in Donini et al., 2018.

We chose the German Credit dataset for its two advantageous characteristics, particularly well-suited to our analysis. These benefits are primarily attributed to its moderate size: (1) it allows for the computation of numerous 10-fold cross-validations within a reasonable timeframe, facilitating the estimation of a 'ground truth', and (2) the dataset's limited scale naturally introduces greater uncertainty into the metrics being evaluated.

In additional experiments presented in this Appendix, we employ various other datasets, which are described below:

Adult Income. It is a dataset based on the data from the 1994 US Census, where the main goal is to predict whether an individual earns more than 50,000$ per year. The 14 features used to describe the instances include occupation, marital status and education. Furthermore, it contains sensitive information such as, age, gender and race. In our experiments, we considered a single sensitive attribute: gender. This dataset is publicly available in the UCI repository[1] and it is already divided into a training and a test set. The former has 32,561 instances and, the latter, 16,281. We pre-process the dataset as in Donini et al., 2018.

[1] `http://archive.ics.uci.edu/ml/index.php`

11.3 Methods

In this section we provide a brief description of the fairness-enhancing interventions that are considered in the experimental section. The considered approaches are grouped, according to the step in which fairness guarantees are enforced within the algorithmic procedure, into pre-processing, in-processing and post-processing methods. Pre-processing mechanisms aim to transform biased datasets so that, when conventional ML classifiers are trained on them, the final outputs are fairer. In-processing interventions, modify existing algorithms to account for fairness guarantees at training time. Lastly, post-processing methods alter algorithmic predictions to obtain fairer final decisions.

11.3.1 Pre-Processing

Reweighing (RW). This pre-processing intervention proposed by Kamiran et al. (2012) aims to transform a biased dataset so that when a conventional ML classifier is fed with such data the effects of it's outcomes are not disproportionate for different sensitive groups. Such transformation constitutes of weighing the instances to achieve equal prevalence across sensitive groups, i.e. to enforce statistical independence between the label and the sensitive attribute.

Disparate Impact Remover (DIR). Feldman et al. (2015) developed a method to pre-process a biased dataset in order to remove the *disparate impact* on the effects of the algorithm across different subgroups of the population when the algorithm is fed with such dataset. The processed dataset is obtained by changing the non-sensitive attributes of the dataset that could be employed to predict the sensitive information.

The code for both methods can be found within the AIF360 Python package, a comprehensive resource comprising benchmark fairness-enhancing interventions and fairness metrics.

11.3.2 In-Processing

Fair Empirical Risk Minimization (FERM). FERM is an in-processing method developed by Donini et al. (2018) which constitutes a modification of the conventional Empirical Risk Minimization (ERM) method. In particular, they propose to introduce additional constraints into the optimization problem to enforce the fulfillment of fairness guarantees by the learning algorithm. These constraints request the learning algorithm to have approximately constant conditional risks with respect to the sensitive attribute. As in their work, we consider SVM as the base learning method, using either a linear kernel (LFERM) or a non-linear kernel (FERM). We have implemented this method using the code provided by the authors[2].

Avoiding Disparate Mistreatment. Zafar et al. (2017) proposed an optimization problem to learn an algorithm, by minimizing a general classification loss subject to fairness constraints. The latter forced the algorithm to achieve similar FNR and FPR performances across the different sensitive groups (which would mean that it does not show disparate mistreatment). In order to avoid tractability issue, they reformulate the problem using a tractable proxy by defining the disparate mistreatment using the covariance between the sensitive attributes of the individuals and the signed distance between the feature vectors of misclassified instances and the classifier decision boundary. We implemented the code provided by the authors[3] with a linear decision boundary.

11.3.3 Post-Processing

Hardt et al. (2016) Post-Processing. Hardt et al. (2016) propose a fairness-enhancing intervention that modifies the outcomes of a given predictor in order to satisfy a given fairness property defined by either *equality of opportunity* or *equality of odds*. In the case of binary predictors, their method flips the decisions with a given probability to satisfy the fairness criteria. On the other hand, for score-based algorithms, they suggest to modify the decision boundary to improve the fairness guarantees of the algorithm: in particular, they modify the decision threshold, assuming different (possibly randomized) thresholds for the distinct sensitive groups. The code for this method can also be found in the AIF360 Python package.

[2] `https://github.com/jmikko/fair_ERM`
[3] `https://github.com/mbilalzafar/fair-classification`

11.4 Experimental Setup to Choose the Optimal Approximation for ρ

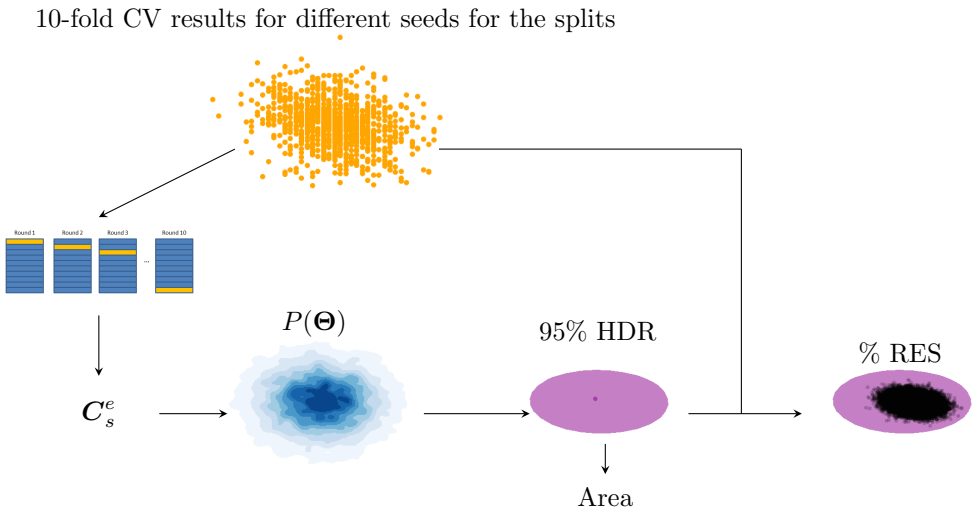


Figure 11.1: The Workflow Representing the Experimental Setup Used to Choose the Best Approximation for ρ_s. Starting from an arbitrary 10-fold CV configuration and compute the effective CM. From the effective CM we conclude the joint posterior distribution of accuracy and a fairness metric. Then, we estimate the 95% HDR of the posterior distribution and calculate its area and proportion of repeated 10-fold CV results that fall within such region (% RES). We repeat this procedure for 10,000 different initial 10-fold CV configurations.

This section presents a more comprehensive overview of the experimental setup discussed in Section 5.1. The primary objective of this experiment is to identify the most accurate approximation for the correlation coefficient ρ. To achieve this goal, we commence by initializing the process with a randomly selected 10-fold CV setup. We then proceed to calculate the posterior distributions using a variety of approximation techniques, including the state-of-the-art (SOTA) methods as well as our proposed approach. For each of these posterior distributions, we calculate the 95% Highest Density Region (HDR) and determine the percentage of repeated 10-fold CV outcomes that fall within this region (% RES). The narrowest posterior distribution capable of encompassing all potential 10-fold CV outcomes will be considered the optimal approach for approximating ρ. Figure 11.1 illustrates a schematic representation of the experimental workflow. We repeat this experiment for 10,000 different initial 10-fold CV configurations and report the results as an average across all iterations.

11.5 Calculating the 95% HDR

In this section, we provide a formal definition of the $(100-\alpha)\%$ HDR and outline existing alternatives to measure it in practice.

The $(100-\alpha)\%$ *highest density region* (HDR) is the $(100-\alpha)\%$ density region with the smallest size, and it constitutes the most compact summary of a probability distribution. Let $f(x)$ be the probability density function of a (possibly multivariate) random variable $X \in \mathbb{R}^d$ and $0 < \alpha < 100$. Following the definition by Hyndman (1996), the $(100-\alpha)\%$ HDR refers to the subset $\{x \,:\, f(x) \geq f_\alpha\}$, where f_α is the highest constant for which $P(X \in \{x \,:\, f(x) \geq f_\alpha\}) = (100-\alpha)\%$:

$$P(X \in \{x \,:\, f(x) \geq f_\alpha\}) = \int_{\{x \,:\, f(x) \geq f_\alpha\}} f(u)\, du \;=\; (100-\alpha)\%$$

There are many alternatives to estimate the $(100-\alpha)\%$ HDR of a distribution. If the probability distribution $f(x)$ is known, the problem can be solved by numerical integration as in Hyndman, 1996. However, this procedure becomes computationally hard with the increasing dimensionality of the sample space. Another popular approach is the so-called 'quantile approach' Hyndman, 1996. In this method, n samples i.i.d. from known $f(x)$ are sorted in descending order, and the $\left\lceil \frac{(100-\alpha)}{100} n \right\rceil$-th element of the sorted sample is considered an approximation of f_α.

The main advantage of this approach is that the computational complexity does not increase when the sample space becomes higher-dimensional.

In the cases where $f(x)$ is unknown and only observable through a set of samples, $f(x)$ can be estimated from the set of samples by, e.g., kernel smoothing Hyndman et al., 1996; Bashtannyk et al., 2001; Samworth et al., 2010. In fact, this is the approach we have adopted in the experiments, based on the implementation of the package `hdrcde`[4] in R. However, there exists other non-parametric alternatives too, such as the heuristic approach by Waltman (2014).

12 ADDITIONAL EXPERIMENTAL RESULTS

12.1 Full Results from Section 5.1

Table 12.1 provides the complete version of Table 1 from the main paper, with a more diverse set of algorithms.

Table 12.1: Evaluation of the Different Approximations for ρ_s. The area of the 95% HDR of the posterior distribution and the proportion of 10,000 different 10-fold CV results it encloses (%RES) for different algorithms (rows) and approximations (1-4) of ρ_s (columns). We use the German Credit dataset with age as the sensitive attribute. It is the extended version of Table 1 from the main paper.

	$\rho_s = 1/K$		$\rho_s \in [0, 1/K]$		$\rho_s = \rho_{rel}$		$\rho_s = \rho^{\uparrow}_{rel}$	
Method	Area	% res	Area	% res	Area	% res	Area	% res
LR	0.0207	100.0	0.0153	100.0	0.0197	100.0	0.0150	99.96
SVM	0.0201	100.0	0.0148	99.99	0.0201	100.0	0.0148	99.99
Lin. SVM	0.0205	99.98	0.0154	99.86	0.0243	100.0	0.0168	99.87
KAMIRAN ET AL. Kamiran et al., 2012 + LR	0.0207	100.0	0.0153	100.0	0.0195	100.0	0.0148	100.0
KAMIRAN ET AL. Kamiran et al., 2012 + SVM	0.0201	100.0	0.0149	99.96	0.0196	99.95	0.0148	99.64
Feldman et al. (2015) + LR	0.0208	100.0	0.0154	100.0	0.0182	99.99	0.0143	99.91
Feldman et al. (2015) + SVM	0.0201	100.0	0.0149	99.99	0.0198	99.99	0.0147	99.93
Zafar et al. (2017)	0.0205	100.0	0.0153	100.0	0.0269	100.0	0.0178	99.99
FERM Donini et al., 2018	0.0107	100.0	0.0079	99.74	0.0110	100.0	0.0080	99.86
LFERM Donini et al., 2018	0.0093	99.99	0.0069	99.93	0.0112	100.0	0.0075	99.97
LR + Hardt et al. (2016)	0.0210	90.32	0.0156	82.08	0.0414	99.04	0.0232	92.77
SVM + Hardt et al. (2016)	0.0203	94.09	0.0151	88.67	0.0413	99.48	0.0230	95.84
Lin. SVM + Hardt et al. (2016)	0.0196	92.91	0.0145	86.94	0.0904	99.99	0.0381	99.00

[4] `https://github.com/robjhyndman/hdrcde`

12.2 The Effect of HDR and RoPE in UM($\rho^{\uparrow}_{rel}$) Results From Section 5.2

12.2.1 Algorithms with Similar Stability

Table 12.2: Most Probable Predictive Performance of UM($\rho^{\uparrow}_{rel}$) in ($\mathcal{A}$) Feldman et al. (2015) + SVM vs. ($\mathcal{B}$) Donini et al. (2018) Algorithmic Comparison for Varying HDR and RoPE. The objectives considered in the algorithmic comparison are accuracy and the fairness notion EOp. Conventional 10-fold CV provides a single deterministic conclusion: $(\mathcal{A}_{acc}, \mathcal{B}_{fair})$. The probabilities assigned to this event by UM($\rho^{\uparrow}_{rel}$) are shaded in gray. The ground truth probabilities are highlighted in **bold blue**. By employing UM, along with commonly used HDR values (e.g., 90% or 95%), we can consistently obtain probabilities that closely resemble the true probabilities of the evaluation outcomes, regardless of the RoPE dimensions chosen.

RoPE		$P(\mathcal{A} >> \mathcal{B})$	$P(\mathcal{B} >> \mathcal{A})$	$P(\mathcal{A} \approx \mathcal{B})$	$P(\mathcal{A}_{acc}, \mathcal{B}_{fair})$	$P(\mathcal{B}_{acc}, \mathcal{A}_{fair})$
(0.01, 0.01)	Ground truth	**0.04**	**0.15**	**0.01**	**0.80**	**0.00**
	UM no HDR	0.01	0.20	0.00	0.79	0.00
	UM 95%HDR	0.01	0.17	0.00	0.82	0.00
	UM 90%HDR	0.00	0.19	0.00	0.81	0.00
	UM 80%HDR	0.00	0.18	0.00	0.82	0.00
(0.02, 0.02)	Ground truth	**0.04**	**0.37**	**0.04**	**0.56**	**0.00**
	UM no HDR	0.02	0.40	0.01	0.57	0.00
	UM 95%HDR	0.02	0.39	0.01	0.58	0.00
	UM 90%HDR	0.02	0.38	0.00	0.60	0.00
	UM 80%HDR	0.00	0.40	0.00	0.60	0.00
None	Ground truth	**0.03**	**0.01**	-	**0.96**	**0.00**
	UM no HDR	0.01	0.06	-	0.93	0.00
	UM 95%HDR	0.00	0.06	-	0.94	0.00
	UM 90%HDR	0.00	0.05	-	0.95	0.00
	UM 80%HDR	0.00	0.04	-	0.96	0.00

12.2.2 Algorithms with Different Stability

Table 12.3: Most Probable Predictive Performance of UM($\rho^{\uparrow}_{rel}$) in ($\mathcal{A}$) Linear SVM + Hardt et al. (2016) vs. ($\mathcal{B}$) Donini et al. (2018) Algorithmic Comparison for Varying HDR and RoPE. The objectives considered in the algorithmic comparison are accuracy and the fairness notion EOp. Conventional 10-fold CV provides a single deterministic conclusion: ($\mathcal{B} >> \mathcal{A}$). The probabilities assigned to this event by UM($\rho^{\uparrow}_{rel}$) are shaded in gray. The ground truth probabilities are highlighted in **bold blue**. By employing UM, along with commonly used HDR values (e.g., 90% or 95%), we can consistently obtain probabilities that closely resemble the true probabilities of the evaluation outcomes, regardless of the RoPE dimensions chosen.

RoPE		$P(\mathcal{A} >> \mathcal{B})$	$P(\mathcal{B} >> \mathcal{A})$	$P(\mathcal{A} \approx \mathcal{B})$	$P(\mathcal{A}_{acc}, \mathcal{B}_{fair})$	$P(\mathcal{B}_{acc}, \mathcal{A}_{fair})$
(0.01, 0.01)	Ground truth	**0.02**	**0.83**	**0.04**	**0.00**	**0.11**
	UM no HDR	0.02	0.82	0.02	0.01	0.13
	UM 95%HDR	0.02	0.83	0.03	0.01	0.11
	UM 90%HDR	0.01	0.87	0.02	0.00	0.10
	UM 80%HDR	0.00	0.88	0.02	0.00	0.09
(0.02, 0.02)	Ground truth	**0.01**	**0.81**	**0.15**	**0.00**	**0.03**
	UM no HDR	0.01	0.85	0.08	0.00	0.06
	UM 95%HDR	0.01	0.85	0.09	0.00	0.05
	UM 90%HDR	0.01	0.86	0.09	0.00	0.04
	UM 80%HDR	0.00	0.91	0.08	0.00	0.01
None	Ground truth	**0.01**	**0.70**	-	**0.03**	**0.26**
	UM no HDR	0.01	0.72	-	0.03	0.24
	UM 95%HDR	0.01	0.74	-	0.03	0.23
	UM 90%HDR	0.00	0.76	-	0.02	0.22
	UM 80%HDR	0.00	0.81	-	0.01	0.18

12.3 The Effect of HDR and RoPE in UM($\rho^{\uparrow}_{rel}$) Results From Section 5.3

12.3.1 Algorithms with Similar Stability

Table 12.4: Worst Case Predictive Performance of UM($\rho^{\uparrow}_{rel}$) in ($\mathcal{A}$) Feldman et al. (2015) + SVM vs. ($\mathcal{B}$) Donini et al. (2018) Algorithmic Comparison for Varying HDR and RoPE. The objectives considered in the algorithmic comparison are accuracy and the fairness notion EOp. Conventional 10-fold CV suggests a single deterministic outcome of algorithmic comparison ($\mathcal{A} >> \mathcal{B}$), whose true probability is negligible (**bold red**). Uncertainty-aware frameworks assign probabilities to these uncommon outcomes, highlighted in gray, and we identify the event with the highest true probability in **bold blue**. Even when applied to rare 10-fold CV results, UM($\rho^{\uparrow}_{rel}$) consistently prioritizes the event with the highest true probability and refrains from assigning a high probability to an incorrect conclusion, regardless of the chosen HDR and RoPE dimension.

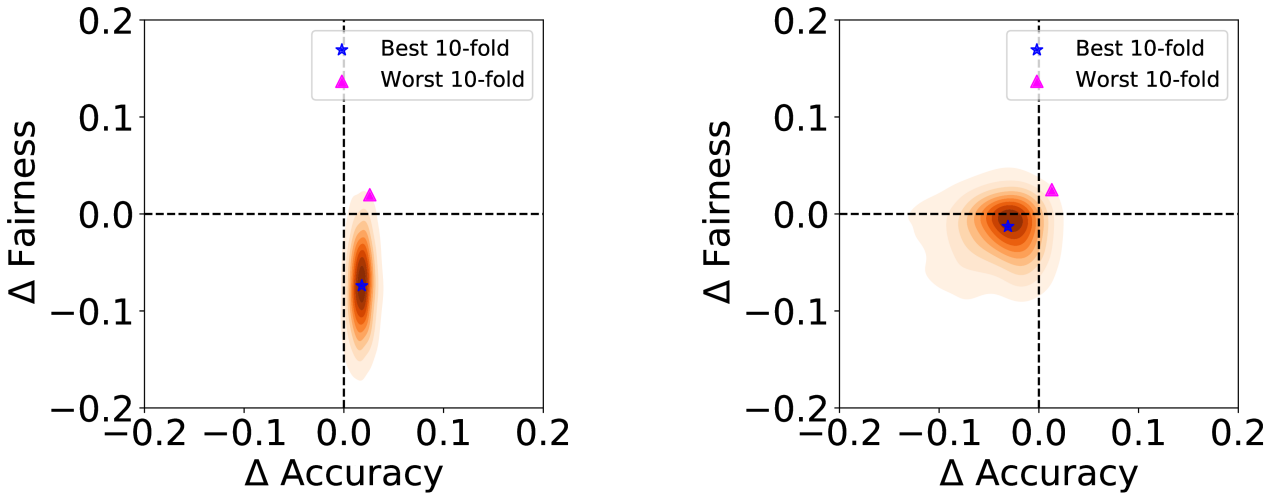
RoPE		$P(\mathcal{A} >> \mathcal{B})$	$P(\mathcal{B} >> \mathcal{A})$	$P(\mathcal{A} \approx \mathcal{B})$	$P(\mathcal{A}_{acc}, \mathcal{B}_{fair})$	$P(\mathcal{B}_{acc}, \mathcal{A}_{fair})$
(0.01, 0.01)	Ground truth	**0.04**	0.15	0.01	**0.80**	0.00
	UM no HDR	0.25	0.13	0.04	0.58	0.00
	UM 95%HDR	0.22	0.10	0.04	0.64	0.00
	UM 90%HDR	0.22	0.09	0.04	0.65	0.00
	UM 80%HDR	0.21	0.11	0.03	0.65	0.00
(0.02, 0.02)	Ground truth	**0.04**	0.37	0.04	**0.56**	0.00
	UM no HDR	0.28	0.21	0.17	0.34	0.00
	UM 95%HDR	0.25	0.22	0.17	0.36	0.00
	UM 90%HDR	0.27	0.21	0.19	0.33	0.00
	UM 80%HDR	0.27	0.21	0.17	0.35	0.00
None	Ground truth	**0.03**	0.01	-	**0.96**	0.00
	UM no HDR	0.14	0.03	-	0.82	0.01
	UM 95%HDR	0.11	0.03	-	0.85	0.01
	UM 90%HDR	0.13	0.02	-	0.85	0.00
	UM 80%HDR	0.08	0.01	-	0.90	0.00

12.3.2 Algorithms with Different Stability

Table 12.5: Worst Case Predictive Performance of UM($\rho^{\uparrow}_{rel}$) in ($\mathcal{A}$) Linear SVM + Hardt et al. (2016) vs. ($\mathcal{B}$) Donini et al. (2018) Algorithmic Comparison for Varying HDR and RoPE. The objectives considered in the algorithmic comparison are accuracy and the fairness notion EOp. Conventional 10-fold CV suggests a single deterministic outcome of algorithmic comparison ($\mathcal{A} >> \mathcal{B}$), whose true probability is negligible (**bold red**). Uncertainty-aware frameworks assign probabilities to these uncommon outcomes, highlighted in gray, and we identify the event with the highest true probability in **bold blue**. Even when applied to rare 10-fold CV results, UM($\rho^{\uparrow}_{rel}$) consistently prioritizes the event with the highest true probability and refrains from assigning a high probability to an incorrect conclusion, regardless of the chosen HDR and RoPE dimension.

RoPE		$P(\mathcal{A} >> \mathcal{B})$	$P(\mathcal{B} >> P(\mathcal{A})$	$P(P(\mathcal{A} \approx \mathcal{B})$	$P(P(\mathcal{A}_{acc}, \mathcal{B}_{fair})$	$P(\mathcal{B}_{acc}, P(\mathcal{A}_{fair})$
(0.01, 0.01)	Ground truth	**0.02**	**0.83**	0.04	0.00	0.11
	UM no HDR	0.18	0.50	0.08	0.04	0.20
	UM 95%HDR	0.17	0.51	0.07	0.05	0.19
	UM 90%HDR	0.18	0.50	0.08	0.04	0.21
	UM 80%HDR	0.16	0.52	0.10	0.03	0.19
(0.02, 0.02)	Ground truth	**0.01**	**0.81**	0.15	0.00	0.03
	UM no HDR	0.17	0.47	0.25	0.01	0.10
	UM 95%HDR	0.16	0.47	0.27	0.01	0.09
	UM 90%HDR	0.15	0.48	0.28	0.01	0.08
	UM 80%HDR	0.13	0.48	0.32	0.00	0.07
None	Ground truth	**0.01**	**0.70**	-	0.03	0.26
	UM no HDR	0.13	0.38	-	0.12	0.37
	UM 95%HDR	0.13	0.38	-	0.12	0.37
	UM 90%HDR	0.12	0.41	-	0.12	0.35
	UM 80%HDR	0.12	0.39	-	0.11	0.38

12.4 Best and Worst 10-fold CV Configurations



(a) (b)

Figure 12.1: The Best and Worst 10-fold CV Partitions from Sections 5.2 and 5.3 for Algorithmic Comparison Scenarios (a) and (b). We examine a sample of 10,000 unique 10-fold partitions and characterize the empirical distribution of the difference based on this sample. The best 10-fold partition is the one where the performance difference is nearest to the average value within the empirical distribution in terms of Euclidean distance. Conversely, the least optimal 10-fold partition is defined by a conclusion that diverges from the one suggested by the average value and demonstrates the most significant deviation from that average. Comparison scenario (a) considers ($\mathcal{A}$) Feldman et al. (2015) + SVM vs. ($\mathcal{B}$) Donini et al. (2018) and scenario (b) refers to ($\mathcal{A}$) Linear SVM + Hardt et al. (2016) vs. ($\mathcal{B}$) Donini et al. (2018).

13 3-OBJECTIVE ALGORITHMIC COMPARISON

In this section, we extend the experiments from Section 5.2 to encompass multi-objective algorithmic comparisons. Specifically, we examine algorithmic comparison between two algorithms, compared with respect to three objectives: overall accuracy and 2 fairness notions (EOp and DP). While with three objectives there are $2^3 + 1$ potential outcomes in the algorithmic comparison, we will primarily focus on the fundamental probabilities: $P(\mathcal{A} \gg \mathcal{B})$, $P(\mathcal{A} \ll \mathcal{B})$, and $P(\mathcal{A} \approx \mathcal{B})$ for the sake of simplicity. The results are shown in Table 13.1.

Analyzing these results, we observe a general trend wherein the probabilities $P(\mathcal{A} \gg \mathcal{B})$, $P(\mathcal{A} \ll \mathcal{B})$, and $P(\mathcal{A} \approx \mathcal{B})$ decrease as the number of objectives considered for comparison increases. This means that as we consider more objectives, it becomes less likely for one algorithm to dominate the other across all objectives, or for their performance to be equivalent. As discussed in Section 6, an increase in the number of objectives also elevates the chances of a method being Pareto-optimal. Consequently, with a substantial number of objectives, it is possible that all methods may achieve Pareto-optimality. In practical terms, utilizing more than three objectives becomes unfeasible, despite the absence of a theoretical constraint within the UM framework.

Table 13.1: UM's Predictive Performance Applied to the Most Likely 10-fold CV Outcomes, in 3-Objective Algorithmic Comparison. The objectives considered in the algorithmic comparison are accuracy and two fairness notions (EOp and DP). We emphasize the true event probabilities using **bold blue**. With regards to the uncertainty-aware framework estimations (SOTA and UM), the **best** result is highlighted, and the second best result is underlined. Through the application of UM, we obtain probabilities that closely match the actual probabilities of the evaluation outcomes.

METHODS		$P(\mathcal{A} >> \mathcal{B})$	$P(\mathcal{B} >> \mathcal{A})$	$P(\mathcal{A} \approx \mathcal{B})$
$\mathcal{A}$: Feldman et al. (2015) + SVM vs. $\mathcal{B}$: Donini et al. (2018)	Ground truth	**0.04**	**0.08**	**0.00**
	SOTA ($\rho = 1/K$)	0.07	0.14	0.01
	SOTA ($\rho \in [0, 1/K]$)	0.03	0.13	**0.00**
	UM (ρ_{rel})	0.07	0.14	0.01
	UM ($\rho^{\uparrow}_{rel}$)	**0.04**	**0.11**	**0.00**
$\mathcal{A}$: LinearSVM + Hardt et al. (2016) vs. $\mathcal{B}$: Donini et al. (2018)	Ground truth	**0.01**	**0.78**	**0.01**
	SOTA ($\rho = 1/K$)	0.02	0.59	**0.01**
	SOTA ($\rho \in [0, 1/K]$)	**0.01**	0.66	**0.01**
	UM (ρ_{rel})	**0.01**	0.71	**0.01**
	UM ($\rho^{\uparrow}_{rel}$)	0.00	**0.77**	0.00

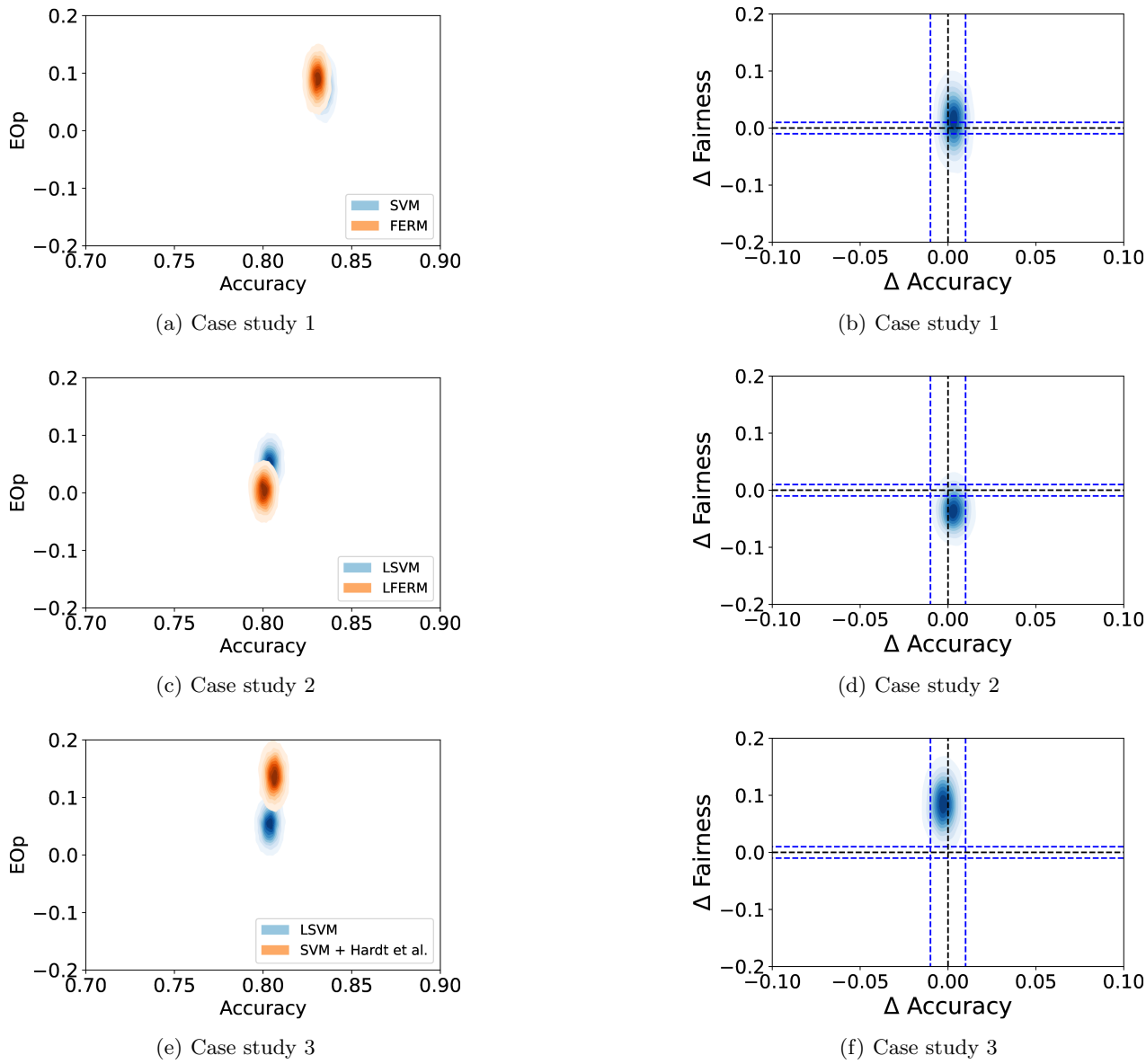
14 INTEGRATING UM INTO EXISTING RESULTS

In this section, we examine the potential impact of integrating UM into the findings of a previously published papers. We compare the outcomes obtained from uncertainty-aware evaluation (UM) against those from the standard result assessment process, where the traditional non-Bayesian evaluation serves as the *baseline* for comparison. Specifically, we employ the study conducted by Donini et al. (2018) as our *baseline*, which evaluates different fairness-enhancing interventions using conventional hold-out and 10-fold CV evaluation techniques.

14.1 Hold-Out Evaluation in Adult Income Dataset

This section shows how the conclusions drawn from a hold-out framework vary when performance metrics are assumed to be random variables. The experiments are performed with the Adult Income dataset which has a given train/test partition, and is the one used by Donini et al. (2018). The algorithms are compared in terms of accuracy and equality of opportunity (EOp) and we consider a RoPE with parameters $\boldsymbol{\epsilon}_{\Theta} = (0.01, 0.01)$. As mentioned in Section 3, we have computed the opposite of EOp for the estimation of $\boldsymbol{\Delta\Theta}$. Specifically, we present two case studies where the inclusion of uncertainty leads to a different conclusion than the deterministic evaluation, and one case study where the uncertainty-aware evaluation aligns with the deterministic conclusion.

Case study 1: ($\mathcal{A}$) Non-linear SVM and ($\mathcal{B}$) FERM (Donini et al., 2018). Under the classic procedure, method ($\mathcal{A}$) obtains an accuracy of 0.84 and its EOP is of 0.07; while method ($\mathcal{B}$) shows a predictive accuracy of 0.83 and the value of EOp is 0.09. Based on these results, method ($\mathcal{A}$) outperforms ($\mathcal{B}$). However, the analysis



(a) Case study 1

(b) Case study 1

(c) Case study 2

(d) Case study 2

(e) Case study 3

(f) Case study 3

Figure 14.1: Visual Representation of the Three Case Studies Associated with the Hold-out Evaluation Method for the Adult Income Dataset. Each row corresponds to a unique case study and displays a comparison between the posterior distributions (left column) and the distribution of $\delta(\Delta\boldsymbol{\Theta})$ (right column).

based on UM (see Figure 1(b)) suggests more fine-grained conclusions: the probability of such dominance is only of 0.54.

Case study 2: ($\mathcal{A}$) Linear SVM and ($\mathcal{B}$) Linear FERM (Donini et al., 2018). Based on the classic hold-out results, method ($\mathcal{A}$) has a better predictive performance (0.805 vs. 0.801), but provides worse fairness guarantees in terms of EOp (0.05 vs. 0.01). Thus, the classical evaluation setting suggests that no algorithm outperforms the other. However, with regards the UM analysis (see Figure 1(d)), there is a probability of 0.77 that algorithm ($\mathcal{B}$) outperforms ($\mathcal{A}$).

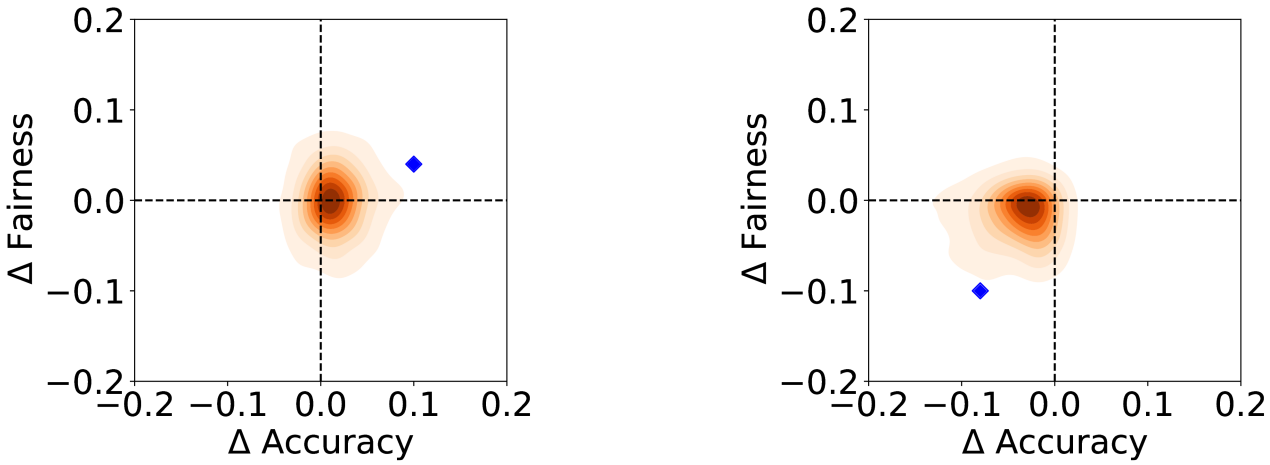
Case study 3: ($\mathcal{A}$) Linear SVM and ($\mathcal{B}$) Non-linear SVM + Hardt et al. (2016) According to the classic result, both methods have an accuracy of 0.805, but ($\mathcal{A}$) provides better fairness guarantees than ($\mathcal{B}$): method ($\mathcal{A}$) obtains an EOp value of 0.05 while ($\mathcal{B}$) achieves 0.14. With that, ($\mathcal{A}$) would be preferred to ($\mathcal{B}$) since, for equivalent accuracy, it provides better fairness guarantees. Furthermore, based on the UM framework

analysis (see Figure 1(f)), the likelihood that $\mathcal{A}$ will outperform $\mathcal{B}$ is 0.93. This outcome serves as evidence that the conclusions drawn from the classic and probabilistic perspectives are consistent with each other.

In addition, we offer a comprehensive visual analysis of the UM approach in the context of the classification task outlined in the Adult Income dataset when using the hold-out evaluation method. Figure 14.1 presents each case study in a separate row, depicting a comparison between the posterior distributions (left column) and the distribution of $\delta(\Delta\boldsymbol{\Theta})$ (right column).

14.2 10-fold CV Evaluation in German Credit Dataset

This section highlights the difference between the conclusions drawn under the classic and UM evaluation frameworks in the case of the 10-fold CV. Different methods were compared in terms of accuracy and EOp, considering a squared RoPE with parameters $\boldsymbol{\epsilon}_\Theta = (0.01, 0.01)$. The posterior distributions will be calculated using $\rho^{\uparrow}_{rel}$ as the approximation for the correlation with $\rho^0_s \in [0, 1/K]$ and $\mathcal{M}_0$ = SVM (non-linear), which has shown to provide the most precise uncertainty estimations.



(a) Case Study 1

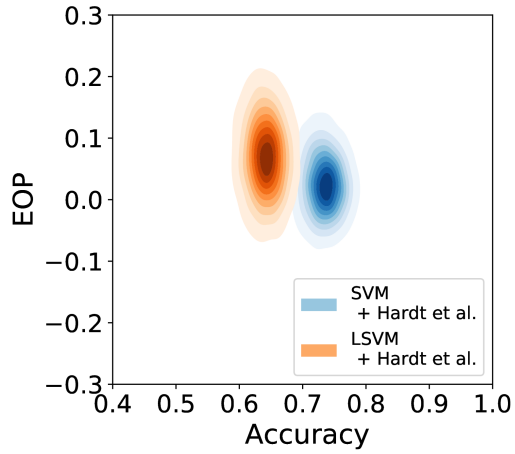
(b) Case Study 2

Figure 14.2: The Result of the 10-fold CV Partition Used in Donini et al. (2018) for German Credit Dataset. The partition employed in Donini et al. (2018) is marked by a blue diamond, while the empirical distribution of 10,000 CV results is represented by the orange area.

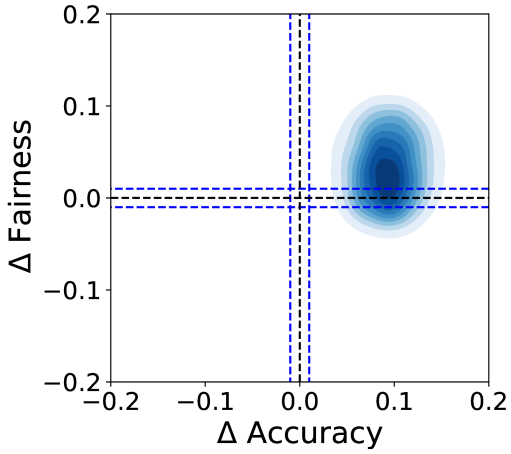
Case study 1: ($\mathcal{A}$) Non-linear SVM + Hardt et al. (2016) and ($\mathcal{B}$) Linear SVM + Hardt et al. (2016). Based on the classical empirical results, it could be argued that utilizing a non-linear kernel for SVM in conjunction with the post-processing method by Hardt et al. (2016) not only yields more accurate predictions (0.71 vs. 0.61) but also results in fairer outcomes in terms of EOp (0.11 vs. 0.15). This suggests that $\mathcal{A}$ significantly outperforms $\mathcal{B}$. However, according to the empirical distribution, $P(\mathcal{A} >> \mathcal{B}) = 0.46$. Even though this event has the highest probability, there are other likely events as well, such as $P(\mathcal{A}_{acc}, \mathcal{B}_{fair}) = 0.32$, which is the second most probable outcome. In fact, it's important to note that this result falls within the tail end of the empirical distribution, as shown in Figure 2(a). If we apply UM to this cross-validation result, we obtain $P(\mathcal{A} >> \mathcal{B}) = 0.78$ and $P(\mathcal{A}_{acc}, \mathcal{B}_{fair}) = 0.21$. Therefore, we would not conclude that $\mathcal{A}$ absolutely dominates $\mathcal{B}$.

Case study 2: ($\mathcal{A}$) Linear SVM + Hardt et al. (2016) and ($\mathcal{B}$) Linear FERM Donini et al., 2018. Classical results suggest that the linear version of the FERM method significantly outperforms the combination between the SVM with linear kernel and the post-processing method by Hardt et al. (2016): the predictive performance and the fairness guarantees are considerably better (0.61 vs. 0.69 in accuracy and 0.15 vs. 0.05 in EOp). While the obtained outcome is relatively unusual, as illustrated in Figure 2(b), it favors the conclusion with the highest true probability ($\mathcal{B} >> \mathcal{A}$), which is precisely 0.83, as detailed in Table 2. When we employ UM on this outcome, it yields $P(\mathcal{B} >> \mathcal{A}) = 0.91$. This implies that although it is the most probable event, there are also other likely outcomes. This added level of detail makes the conclusions more informative and enables

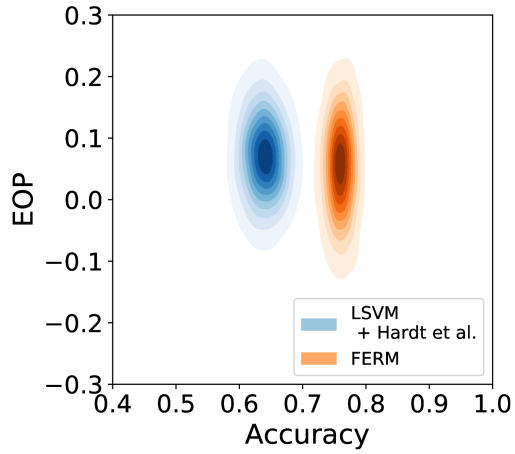
more confident inferences to be drawn.



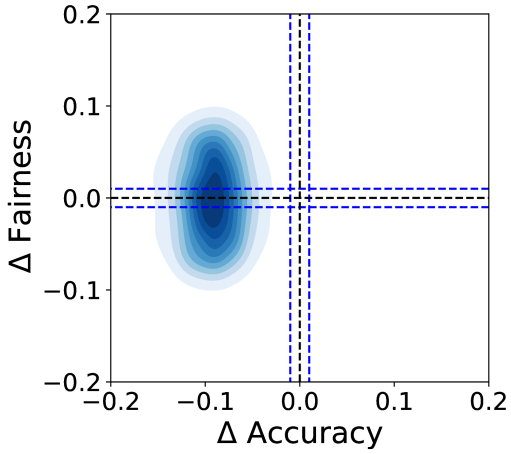
(a) Case study 1



(b) Case study 1



(c) Case study 2



(d) Case study 2

Figure 14.3: Visual Representation of the Case Studies Associated with the 10-fold CV Evaluation Method for the German Credit Dataset. Each row corresponds to a unique case study and displays a comparison between the posterior distributions (left column) and the distribution of $\delta(\Delta\boldsymbol{\Theta})$ (right column).

15 FURTHER DISCUSSION

15.1 Uncertainty of the Majority and the Minority Groups

In the presence of an underrepresented group, the classifier often exhibits a tendency to perform well with the majority group while struggling with the minority population, as it becomes more adept at learning the patterns exhibited by the majority. Additionally, there is less confidence in the algorithm's actual performance on the underrepresented population. Figures 15.1 and 15.2 show that the posterior distributions corresponding to the minority groups are wider that those of the majority group, in the German Credit and Adult Income datasets, respectively. Those graphical results confirm that the uncertainty surrounding the performance of different classifiers is higher for the minority groups.

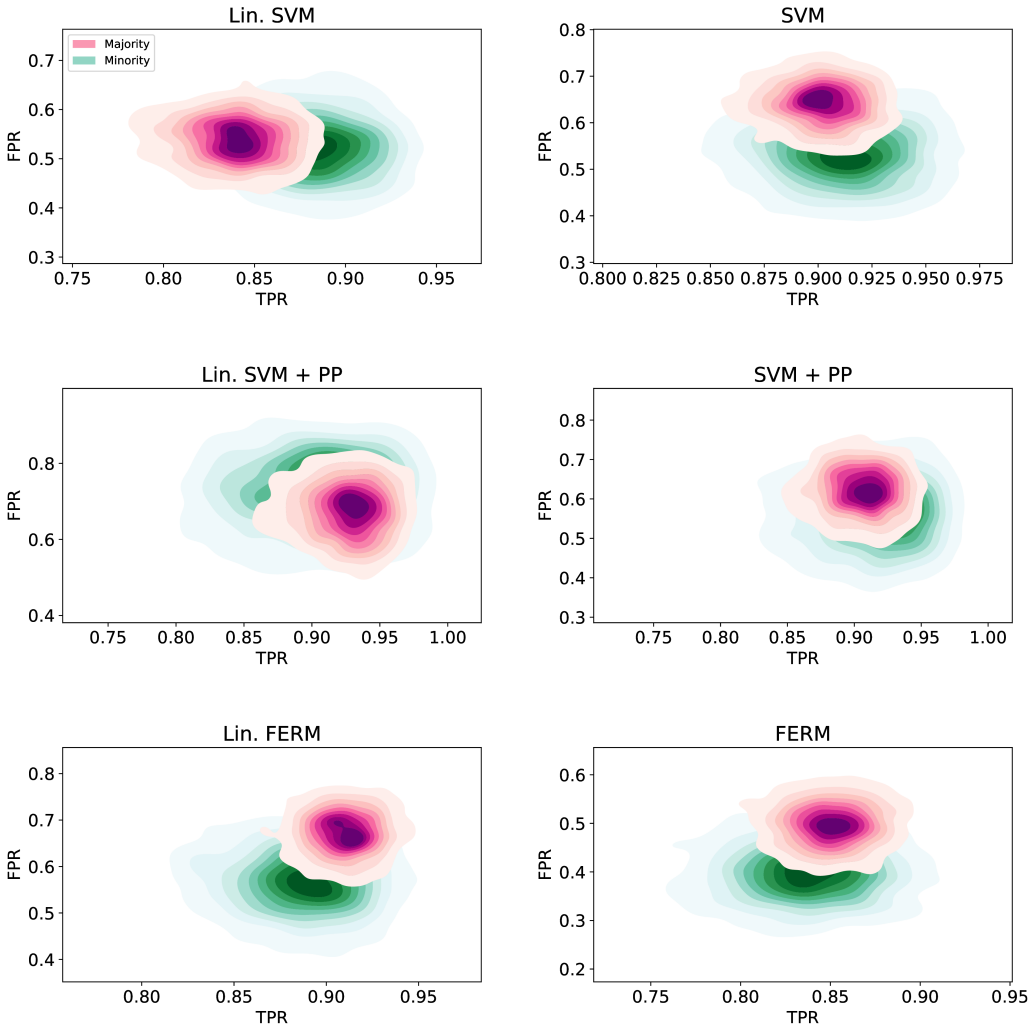


Figure 15.1: Majority vs. Minority in German Credit Dataset. The joint posterior distribution of the true positive rate (TPR) and false positive rate (FPR) for both the majority (purple) and minority (green) groups with the German Credit dataset, using age as the sensitive attribute, where instances corresponding to those aged ≤ 25 account for 19% of the total. The posterior distributions for the minority group are wider than those of the majority group across all methods. This suggests that there is greater uncertainty about how well the model will perform on the minority population.

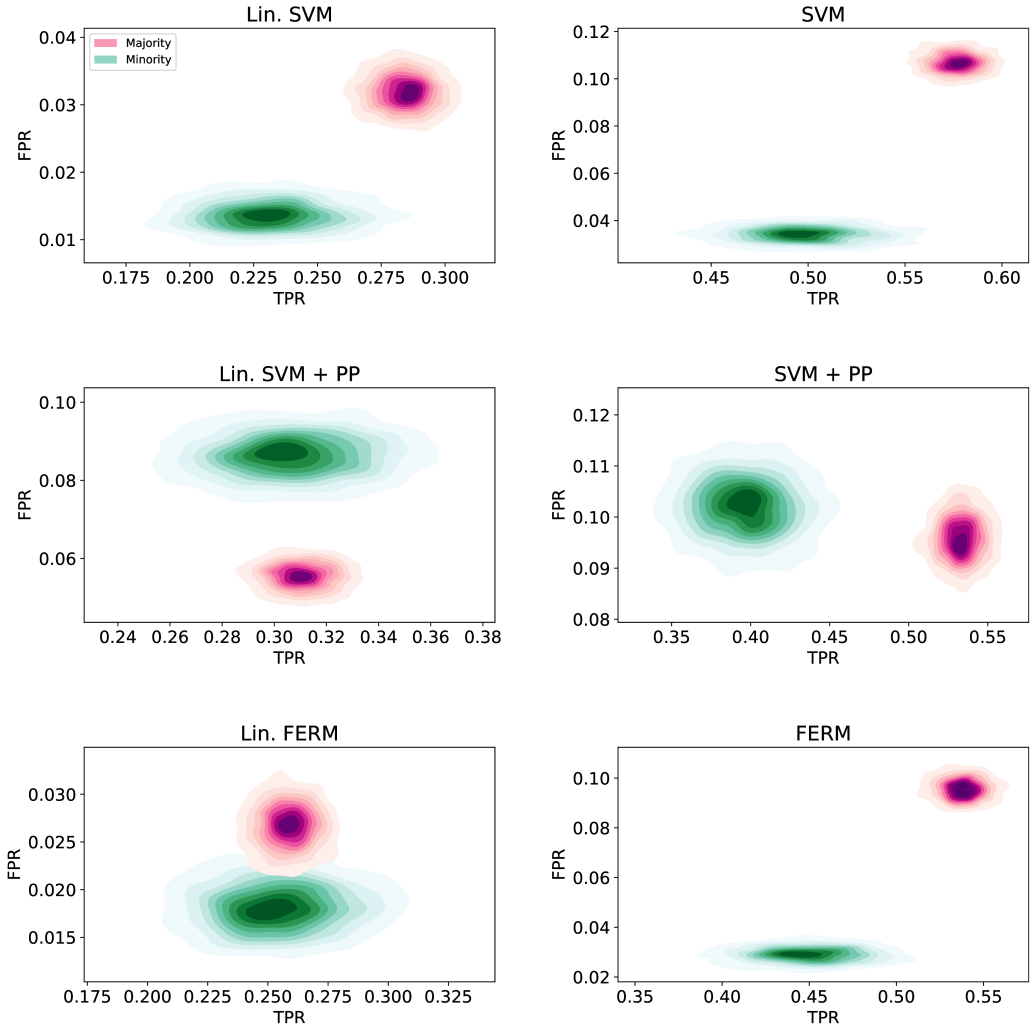


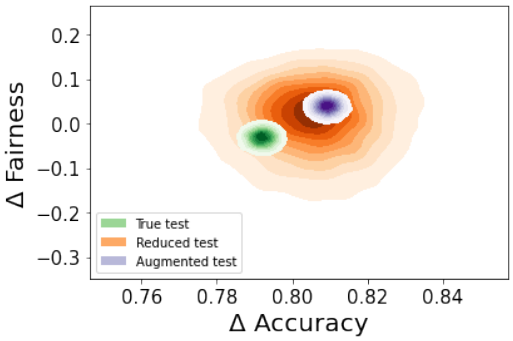
Figure 15.2: Majority vs. Minority in Adult Income Dataset. The joint posterior distribution of the true positive rate (TPR) and false positive rate (FPR) for both the majority (purple) and minority (green) groups with the Adult Income dataset, with gender as the sensitive attribute, where Females represent the 32% of the instances. The posterior distributions for the minority group are wider than those of the majority group across all methods. This suggests that there is greater uncertainty about how well the model will perform on the minority population.

15.2 Can Data-Augmentation Reduce the Uncertainty?

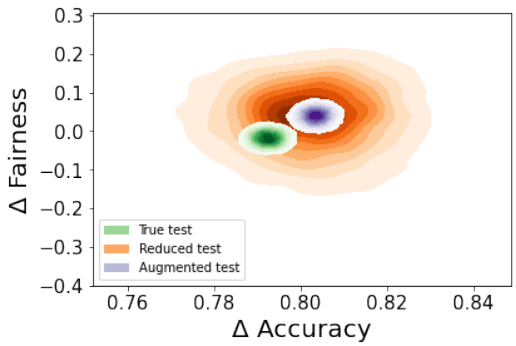
To decrease the uncertainty concerning the model's performance on the target population, it is necessary to increases the information available about the true underlying distribution, by gathering more independent and identically distributed (i.i.d.) samples from the distribution. Note that the number of samples required to achieve a given degree of certainty depends on the model.

One approach to reducing the uncertainty of the posterior distribution is to randomly upsample the number of instances in the test set, but this method has its limitations. Although this can make the uncertainty arbitrarily small, the resulting value may not necessarily converge to that obtained from the true distribution. This is because the instances would be i.i.d. from the empirical distribution, rather than i.i.d. from the true distribution. We empirically verify such statement through an experiment, in which we compare the posterior distribution obtained from the augmented dataset and the posterior distribution derived from the 'true' distribution. We consider the Adult Income dataset, which is randomly divided into a train set with 50% of instances and a test set with the remaining instances, which will be used to represent the 'true' distribution. A reduced test set is constructed by randomly selecting 5% of the instances from this test set. Subsequently, we augment this reduced dataset by randomly upsampling the instances from the reduced test set until the augmented test set contains the same number of instances as the 'true' test set. The posterior distributions obtained from the 'true' test set, the reduced test set, and the augmented test set are shown in Figure 15.3.

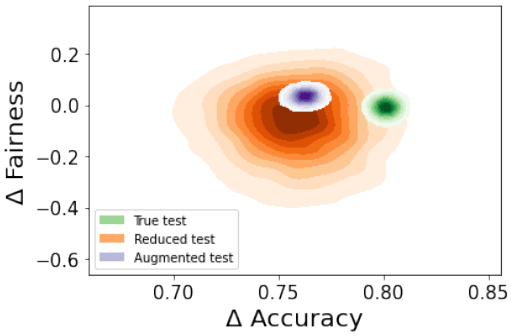
However, a data-augmentation technique that provides guarantees achieving i.i.d samples of the true distribution would be beneficial in decreasing such uncertainty.



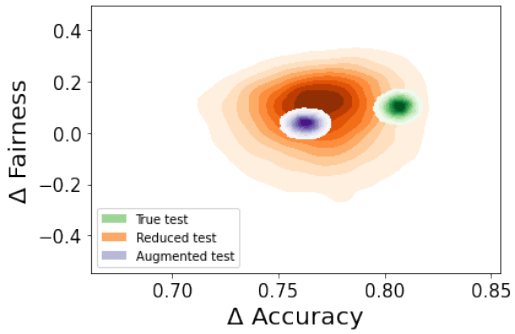
(a) Non-linear SVM + Hardt et al. (2016)



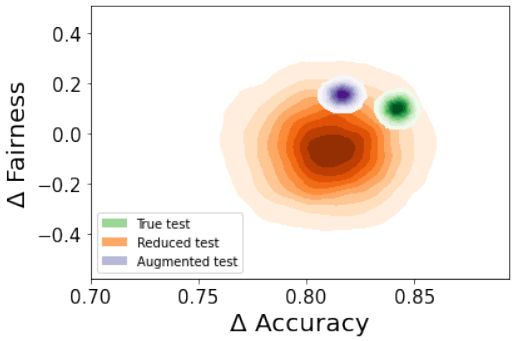
(b) Linear SVM + Hardt et al. (2016)



(c) Linear FERM Donini et al., 2018



(d) Linear SVM



(e) SVM

Figure 15.3: The Effect of Randomly Upsampling the Instances of the Test Set. Even though this strategy can make the uncertainty arbitrarily small, the resulting value may not necessarily converge to that obtained from the true distribution. This is because the instances would be i.i.d. from the empirical distribution, rather than i.i.d. from the true distribution.

15.3 The Effect of the Prior

The experiments in this section are designed to assess the influence of the prior chosen for the multinomial parameter $\boldsymbol{\pi}s$ on the outcomes presented in Section 5.2. We will specifically concentrate on one of the configurations from Table 2, which exhibited strong performance indicators (e.g., UM ($\rho rel^{\uparrow}$) with a 95% HDR), and analyze how various non-informative priors affect the reported probabilities. The results can be found in Tables 15.1 and 15.2. We notice minimal fluctuations, indicating that the impact of the prior (especially when considering typical non-informative priors) on the obtained results and the performance of UM is negligible.

Table 15.1: Impact of Prior Selection on Feldman et al. (2015) + SVM vs. Donini et al. (2018). Predicted probabilities using the UM ($\rho^{\uparrow}_{rel}$) 95%HDR configuration with various non-informative priors and a RoPE of dimensions (0.01,0.01). Minimal variation in probabilities observed.

	$P(A >> B)$	$P(B >> A)$	$P(A \approx B)$	$P(A_{acc}, B_{fair})$	$P(B_{acc}, A_{fair})$
Dir(1,1,1,1)	0.01	0.17	0.00	0.82	0.00
Dir($\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}$)	0.00	0.19	0.00	0.81	0.00
Dir(0,0,0,0)	0.00	0.20	0.00	0.80	0.00

Table 15.2: Impact of Prior Selection on Linear SVM + Hardt et al. (2016) vs. Donini et al. (2018). Predicted probabilities using the UM ($\rho^{\uparrow}_{rel}$) 95%HDR configuration with various non-informative priors and a RoPE of dimensions (0.01,0.01). Minimal variation in probabilities observed.

	$P(A >> B)$	$P(B >> A)$	$P(A \approx B)$	$P(A_{acc}, B_{fair})$	$P(B_{acc}, A_{fair})$
Dir(1,1,1,1)	0.02	0.83	0.03	0.01	0.11
Dir($\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}$)	0.01	0.84	0.03	0.01	0.11
Dir(0,0,0,0)	0.01	0.84	0.03	0.01	0.11

References

Besse, Philippe et al. (2021). "A survey of bias in machine learning through the prism of statistical parity". In: *The American Statistician*, pp. 1–11.

Kruschke, John (2014). "Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan". In.

Agarwal, Alekh et al. (2018). "A reductions approach to fair classification". In: *International Conference on Machine Learning*. PMLR, pp. 60–69.

Aghbalou, Anass, Anne Sabourin, and François Portier (2023). "On the bias of K-fold cross validation with stable learners". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 3775–3794.

Bashtannyk, David M and Rob J Hyndman (2001). "Bandwidth selection for kernel conditional density estimation". In: *Computational Statistics & Data Analysis* 36.3, pp. 279–298.

Benavoli, Alessio et al. (2017). "Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis". In: *The Journal of Machine Learning Research* 18.1, pp. 2653–2688.

Besse, Philippe et al. (2018). "Confidence intervals for testing disparate impact in fair learning". In: *arXiv preprint arXiv:1807.06362*.

Bhatt, Umang et al. (2021). "Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413.

Caelen, Olivier (2017). "A Bayesian interpretation of the confusion matrix". In: *Annals of Mathematics and Artificial Intelligence* 81.3, pp. 429–450.

Dimitrakakis, Christos et al. (2019). "Bayesian fairness". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 509–516.

Donini, Michele et al. (2018). "Empirical risk minimization under fairness constraints". In: *Advances in Neural Information Processing Systems* 31.

Dua, Dheeru and Casey Graff (2017). *UCI Machine Learning Repository*. URL: `http://archive.ics.uci.edu/ml`.

Dwork, Cynthia et al. (2015). "The reusable holdout: Preserving validity in adaptive data analysis". In: *Science* 349.6248, pp. 636–638.

Feldman, Michael et al. (2015). "Certifying and removing disparate impact". In: *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268.

Foulds, James R. et al. (2020). "Bayesian Modeling of Intersectional Fairness: The Variance of Bias". In: *Proceedings of the 2020 SIAM International Conference on Data Mining, SDM 2020, Cincinnati, Ohio, USA, May 7-9, 2020*. Ed. by Carlotta Demeniconi and Nitesh V. Chawla. SIAM, pp. 424–432. DOI: `10.1137/1.9781611976236.48`. URL: `https://doi.org/10.1137/1.9781611976236.48`.

Friedler, Sorelle A et al. (2019). "A comparative study of fairness-enhancing interventions in machine learning". In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 329–338.

Goutte, Cyril and Eric Gaussier (2005). "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation". In: *European conference on information retrieval*. Springer, pp. 345–359.

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

Hyndman, Rob J (1996). "Computing and graphing highest density regions". In: *The American Statistician* 50.2, pp. 120–126.

Hyndman, Rob J, David M Bashtannyk, and Gary K Grunwald (1996). "Estimating and visualizing conditional densities". In: *Journal of Computational and Graphical Statistics* 5.4, pp. 315–336.

Ji, Disi, Padhraic Smyth, and Mark Steyvers (2020). "Can I trust my fairness metric? assessing fairness with unlabeled data and bayesian inference". In: *Advances in Neural Information Processing Systems* 33, pp. 18600–18612.

Kamiran, Faisal and Toon Calders (2012). "Data preprocessing techniques for classification without discrimination". In: *Knowledge and information systems* 33.1, pp. 1–33.

Kruschke, John K and Torrin M Liddell (2018). "The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective". In: *Psychonomic bulletin & review* 25.1, pp. 178–206.

Nadeau, Claude and Yoshua Bengio (2003). "Inference for the Generalization Error". In: *Machine Learning* 52.3, pp. 239–281.

Qian, Shangshu et al. (2021). "Are My Deep Learning Systems Fair? An Empirical Study of Fixed-Seed Training". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 30211–30227. URL: https://proceedings.neurips.cc/paper/2021/file/fdda6e957f1e5ee2f3b311fe4f145ae1-Paper.pdf.

Romano, Yaniv, Stephen Bates, and Emmanuel Candes (2020). "Achieving equalized odds by resampling sensitive attributes". In: *Advances in Neural Information Processing Systems* 33, pp. 361–371.

Samworth, RJ and MP Wand (2010). "Asymptotics and optimal bandwidth selection for highest density region estimation". In: *The Annals of Statistics* 38.3, pp. 1767–1792.

Verma, Sahil and Julia Rubin (2018). "Fairness definitions explained". In: *FairWare'18: IEEE/ACM International Workshop on Software Fairness*. Gothenburg, Sweden: ACM, pp. 1–7. ISBN: 9781450357463. DOI: 10.1145/3194770.3194776. URL: https://doi.org/10.1145/3194770.3194776.

Waltman, Marijn T (2014). "An Algorithm for Approximating the Highest Density Region in d-Space". In.

Wang, Ruibo and Jihong Li (2019). "Bayes test of precision, recall, and F1 measure for comparison of two natural language processing models". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4135–4145.

Yeom, Samuel et al. (2018). "Privacy risk in machine learning: Analyzing the connection to overfitting". In: *IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282.

Zafar, Muhammad Bilal et al. (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.