Reasonable Doubt in the Face of Bias: Fair Flagging with Dirichlet-Based Models

Charmaine Barker
University of York
York, UK
charmaine.barker@york.ac.uk

Dimitar Kazakov
University of York
York, UK
dimitar.kazakov@york.ac.uk

Abstract

Ensuring safety in machine learning requires not only robustness to adversarial or distributional uncertainty but also protection against systematic bias. Models that produce unfair or group-dependent predictions pose critical risks when deployed in socially sensitive domains such as credit, justice, or healthcare. This work introduces EviFair, a fairness-aware safety monitor that flags predictions exhibiting excessive dependence on protected attributes. EviFair combines evidential uncertainty modelling with sensitivity analysis to detect biased decision paths, flagging predictions where fairness cannot be guaranteed. We also show that EviFair's bias scores can effectively guide post-processing fairness methods. Results on standard fairness benchmark datasets show that EviFair achieves substantial reductions in group disparities with minimal impact on predictive performance, demonstrating its promise as a practical, inference-time mechanism for bias-sensitive, safety-aware model oversight.

CCS Concepts

• **Computing methodologies** → **Supervised learning by classification**; *Uncertainty quantification*; Neural networks.

Keywords

Safe Artificial Intelligence, Fairness in Machine Learning, Dirichlet-Based Modelling, Fair Abstention, Bias-Aware Decision Making, Safety Monitors

ACM Reference Format:
Charmaine Barker and Dimitar Kazakov. 2026. Reasonable Doubt in the Face of Bias: Fair Flagging with Dirichlet-Based Models. In *Proceedings of The 41st ACM/SIGAPP Symposium on Applied Computing (SAC'26).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Ensuring the safe deployment of machine learning (ML) systems necessitates addressing not only robustness to adversarial perturbations and distributional shifts, but also fairness in decision making. In sensitive application areas such as credit scoring, criminal justice, and healthcare, disparities in model error rates across protected groups can perpetuate social inequities and undermine the reliability of automated systems [15]. Consequently, fairness failures should be viewed as safety failures, as biased outcomes constitute risks of comparable severity to those arising from adversarial or distributional threats [2, 16].

Existing fairness interventions typically fall into three categories, pre-processing (e.g., data reweighting/representation), in-processing (objective or constraint modifications), and post hoc adjustments to a trained model's decisions, aimed at reducing group disparities [11, 1, 8]. Representative post hoc methods include threshold adjustment and reject-option classification [12, 17], which can improve parity metrics without retraining but operate purely on the model's outputs. Consequently, they do not assess *why* certain predictions are unfair or identify which individual predictions carry high fairness risk. In practice, these techniques adjust or flip decisions globally but lack the capacity to detect fairness violations arising from sensitive feature dependence or epistemic uncertainty at inference time.

We propose EviFair, a fairness-aware abstention wrapper that acts as a safety monitor around any probabilistic classifier. It augments a base model with an evidential uncertainty head that quantifies both predictive confidence and the contribution of sensitive features. At inference, EviFair computes a fairness sensitivity score that reflects how much a prediction's confidence depends on protected attributes, flagging those that exceed a tunable threshold. To prevent disproportionate flagging across groups, we introduce an equalised coverage policy that enforces balanced retention rates. This formulation transforms fairness control into an explicit coverage–fairness trade-off that can be adjusted according to operational safety requirements. Our main contributions are as follows:

- EviFair introduces a fairness-aware safety monitor that flags predictions exhibiting excessive dependence on protected attributes, providing a principled mechanism for bias-sensitive decision oversight.
- It formulates an evidential sensitivity framework that combines uncertainty modelling with feature-path analysis to detect fairness-related confidence shifts at inference time.
- An equalised coverage policy is proposed to ensure group-balanced abstention behaviour, transforming fairness control into a tunable coverage–fairness trade-off aligned with safety requirements.
- Extensive evaluation on multiple benchmark datasets demonstrates substantial fairness improvements with minimal impact on predictive performance.

The paper is laid out as follows. Section 2 reviews relevant research on algorithmic fairness, abstention, and evidential uncertainty. Section 3 outlines necessary background on evidential deep learning. Section 4 presents EviFair, describing the evidential and classifier paths as well as the equalised abstention policy. Section 5

SAC'26, Thessaloniki, Greece

ACM ISBN 979-X-XXXX-XXXX-X/26/03
https://doi.org/XXXXXXX.XXXXXXX

details the datasets and evaluation metrics, and Section 6 reports the experimental results. Finally Section 7 summarises the main findings and discusses avenues for future work.

2 Related Work

Research into algorithmic fairness has developed along three main pathways: pre-processing, in-processing, and post-processing mitigation strategies. Pre-processing techniques aim to reduce bias in the training data itself, typically through reweighing or representation learning that removes correlations between sensitive attributes and labels [11, 19]. In-processing approaches incorporate fairness directly into the optimisation objective by constraining or regularising the model during training [1, 20]. Post-processing methods adjust the outputs of an already-trained classifier to improve fairness metrics without retraining, for example by threshold adjustment, probabilistic relabelling, or reject-option classification [8, 17].

Beyond direct fairness interventions, selective prediction and abstention frameworks offer a complementary means of promoting reliability by allowing models to defer uncertain decisions [7]. Recent work has extended this paradigm to fairness-aware abstention, constraining rejection behaviour to prevent disproportionate deferrals across protected groups [12, 13]. Such methods, however, typically rely on softmax confidence or calibrated probabilities, which may not faithfully capture epistemic uncertainty or the influence of sensitive attributes on model confidence. In contrast, our approach derives abstention signals from evidential uncertainty and fairness-sensitive feature contributions, providing a principled and interpretable basis for flagging predictions that are likely to be biased.

Evidential approaches extend probabilistic classification by predicting the parameters of a Dirichlet distribution over class probabilities rather than point estimates [14, 18]; the underlying formulation is detailed in Section 3. This formulation encodes both the predicted class mean and its epistemic uncertainty through total evidence, enabling principled confidence estimation without sampling or ensembles. Although evidential models have been widely applied for out-of-distribution detection, calibration, and uncertainty-aware decision-making, their potential for bias detection and fairness monitoring remains largely unexplored. Our work bridges this gap by leveraging evidential uncertainty to expose fairness-sensitive confidence shifts and integrate them into a safety-oriented monitoring framework.

3 Preliminaries

Evidential deep learning extends probabilistic classification by modelling uncertainty directly over the probability simplex rather than producing a single-point estimate [18]. Instead of outputting logits passed through a softmax layer, a model predicts the parameters $\alpha = (\alpha_1, \ldots, \alpha_K)$ of a Dirichlet distribution over class probabilities $p = (p_1, \ldots, p_K)$:

$$\mathrm{Dir}(p \mid \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \qquad B(\alpha) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}.$$

where $\Gamma(\cdot)$ denotes the gamma function. The model outputs non-negative evidence values $e_k \geq 0$ for each class, from which the Dirichlet parameters are obtained as $\alpha_k = e_k + 1$. The Dirichlet mean $\mathbb{E}[p_k] = \alpha_k / \sum_j \alpha_j$ yields the expected class probability, while the total evidence $E = \sum_k e_k = \sum_k \alpha_k - K$ quantifies epistemic confidence. High evidence corresponds to strong support for the prediction, whereas low evidence indicates uncertainty arising from insufficient or conflicting data.

Training typically minimises a Dirichlet negative log-likelihood coupled with a regularisation term that encourages proximity to a uniform prior:

$$\mathcal{L}_{\mathrm{EVD}} = -\sum_{k=1}^{K} y_k \big(\psi(\alpha_k) - \psi(\sum_j \alpha_j)\big) + \lambda_{\mathrm{KL}} D_{\mathrm{KL}}\big(\mathrm{Dir}(\alpha) \,\|\, \mathrm{Dir}(\mathbf{1})\big), \tag{1}$$

where y_k is the one-hot label and $\psi(\cdot)$ denotes the digamma function. The first term aligns the predicted Dirichlet mean with the ground-truth label, while the KL term prevents overconfident evidence accumulation under ambiguous data.

4 Methodology

We build on the evidential framework introduced in Section 3 to ground fairness-sensitive safety monitoring in a probabilistic representation of uncertainty. The evidential formulation provides a structured measure of confidence and uncertainty that can be analysed and constrained under fairness considerations, forming the basis for the approach developed below.

4.1 Overview

We propose **EviFair**, a safety wrapper that enables fairness-aware abstention based on evidential uncertainty and feature-path sensitivity. EviFair can be integrated on top of any probabilistic classifier, acting as an external safety layer that suppresses or abstains from predictions where the model's confidence or fairness cannot be guaranteed. It does so by jointly modelling predictive uncertainty, capturing evidential confidence (*Evi*), and the contribution of sensitive attributes, enforcing fairness-aware constraints (*Fair*), to equalise abstention across groups.

Formally, let (x_i, y_i, p_i) denote a sample with features $x_i \in \mathbb{R}^d$, label $y_i \in \{0, \ldots, K-1\}$, and protected attribute vector $p_i \in \{0, 1\}^m$. The base classifier $f_\theta(x)$ outputs logits $z \in \mathbb{R}^K$ and predicted class probabilities $\pi_\theta(x) = \mathrm{softmax}(z)$. Alongside it, an evidential bias head $g_\phi(x)$ produces class-wise Dirichlet parameters $\alpha = g_\phi(x) \in \mathbb{R}_+^K$, as introduced in Section 3. The resulting Dirichlet distribution models uncertainty over class probabilities, with total evidence $E = \sum_k \alpha_k - K$ quantifying epistemic confidence. Low evidence indicates high uncertainty and is used to guide abstention.

4.2 Feature-wise Evidential Parameterisation

The evidential head g_ϕ assigns *feature-wise, class-conditional* evidence. For input $x \in \mathbb{R}^d$, it produces an evidence tensor $e \in \mathbb{R}_{\geq 0}^{d \times K}$ via

$$e_{i,k} \;=\; \mathrm{softplus}\big(w_{i,k}\, x_i + b_{i,k}\big), \qquad i = 1, \ldots, d, \;\; k = 1, \ldots, K, \tag{2}$$

and aggregates feature contributions into Dirichlet parameters

$$\alpha_k \;=\; 1 + \sum_{i=1}^{d} e_{i,k}, \qquad S \equiv \sum_{j=1}^{K} \alpha_j, \qquad E \equiv S - K. \tag{3}$$

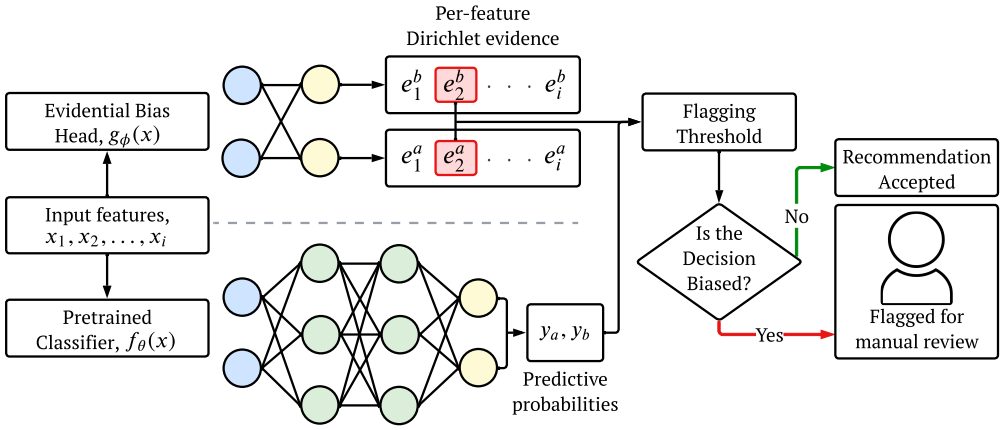


Figure 1: Overview of the safety wrapper. Inputs $x \in \mathbb{R}^d$ are fed to a classifier $f_\theta(x)$ and an evidential head $g_\phi(x)$, trained in parallel on the same data. The evidential head outputs per-feature, per-class Dirichlet evidence $e_{i,k}$ with parameters $\alpha_k = 1 + \sum_i e_{i,k}$, from which the total evidence $E = \sum_k \alpha_k - K$ encodes prediction confidence. At inference, evidence corresponding to sensitive features $\mathcal{S}$ is isolated to evaluate two sensitivity paths: (i) the *evidential path* $\Delta_{\mathbf{EVD}}(x)$, the drop in Dirichlet belief when masking sensitive-feature evidence, and (ii) the *classifier path* $\Delta_{\mathbf{CLS}}(x)$, the drop in predicted class probability when sensitive features are counterfactually zeroed. The overall fairness sensitivity score is $s(x) = \max(\Delta_{\mathbf{EVD}}, \Delta_{\mathbf{CLS}})$, which implicitly reflects both fairness dependence and confidence. Predictions are accepted when $s(x) \le \tau_g$ and *flagged* for review otherwise.

This factorisation makes the contribution of each feature explicit for every class, enabling sensitivity analysis to protected attributes. Let $\mathcal{S} \subseteq \{1, \ldots, d\}$ index sensitive features and define a masking operator that zeroes their evidence:

$$\alpha_k^{(-\mathcal{S})} \;=\; 1 + \sum_{i \notin \mathcal{S}} e_{i,k}. \tag{4}$$

The resulting change in Dirichlet mean probabilities (Section 4.3) yields the evidential-path score $\Delta_{\mathrm{EVD}}(x)$.

To encourage identifiability and parsimonious attributions, we apply (i) random feature masking during training to decorrelate evidence channels, and (ii) a small ℓ_1 penalty on $\sum_{i,k} e_{i,k}$ (with weight 10^{-5}) to discourage diffuse allocations. This stabilises per-feature evidence and improves sensitivity isolation for protected attributes.

4.3 Evidential and Classifier Path Scores

We train the evidential head and the base classifier in parallel on the same inputs. For an input x, let $\pi_{\mathrm{EVD}}(x)$ denote the Dirichlet mean with components $\pi_{\mathrm{EVD},k}(x) = \alpha_k / S$, and let $\pi_{\mathrm{EVD}}^{(-\mathcal{S})}(x)$ be the corresponding mean computed from the masked parameters $\alpha^{(-\mathcal{S})}$. The classifier outputs $\pi_\theta(x) = \mathrm{softmax}(f_\theta(x))$.

To compare the effect of sensitive information, we anchor both paths to the model's own predicted class under the relevant representation. Specifically,

$$y^*_{\mathrm{EVD}} = \arg\max_k \pi_{\mathrm{EVD},k}(x), \qquad y^*_{\mathrm{CLS}} = \arg\max_k \pi_{\theta,k}(x),$$

so that each path measures the change in the probability of the class the model actually intends to predict.

Evidential path. The evidential path quantifies how much the Dirichlet mean of the predicted class degrades when we remove the contribution of sensitive features from the *evidence* itself. Formally,

$$\Delta_{\mathrm{EVD}}(x) = \pi_{\mathrm{EVD},\, y^*_{\mathrm{EVD}}}(x) - \pi^{(-\mathcal{S})}_{\mathrm{EVD},\, y^*_{\mathrm{EVD}}}(x). \tag{5}$$

because masking reduces the available evidence, $\Delta_{\mathrm{EVD}}(x) \in [0, 1]$ and is monotone in the strength of sensitive-feature evidence for the predicted class: large values indicate that a substantial fraction of the belief in the prediction is attributable to sensitive attributes.

Classifier path. The classifier path performs the analogous comparison at the *input* level: we counterfactually zero the sensitive coordinates and recompute the predictive probability of the classifier's own argmax. Let $x^{(-\mathcal{S})}$ denote this masked input. Then

$$\Delta_{\mathrm{CLS}}(x) = \pi_{\theta,\, y^*_{\mathrm{CLS}}}(x) - \pi_{\theta,\, y^*_{\mathrm{CLS}}}\big(x^{(-\mathcal{S})}\big). \tag{6}$$

This score captures direct reliance of the classifier on sensitive features, independent of how the evidential head apportions evidence.

Masking and calibration. We construct $x^{(-\mathcal{S})}$ by zero-imputation (appropriate after z-scoring so that 0 approximates the empirical mean); alternative counterfactuals—mean-imputation, group-conditional means, or model-based generators—are also compatible with our definitions. To place both path scores on a common $[0, 1]$ scale and make them comparable across datasets, we apply a calibration map $\nu(\cdot)$ using percentile scaling on a held-out split, and define the unified sensitivity

$$s(x) = \max\big\{\nu(\Delta_{\mathrm{EVD}}(x)),\ \nu(\Delta_{\mathrm{CLS}}(x))\big\}. \tag{7}$$

By construction, larger $s(x)$ indicates greater dependence of the decision at x on sensitive features, either through the evidential mechanism or through the classifier's direct use of those features.

4.4 Equalised Abstention Policy

Given a sensitivity score $s(x)$ for all test instances, we seek an abstention policy that maintains fairness in coverage across sensitive groups. For a target coverage $\eta \in [0, 1]$, we define per-group thresholds such that the proportion of retained samples (non-abstained) is equal within each group:

$$\mathcal{K}_g(\eta) \;=\; \{i : s_i < \tau_g(\eta)\}, \qquad \tau_g(\eta) \;=\; Q_\eta\big(s_i \,\big|\, p_i = g\big), \quad (8)$$

and Q_q denotes the empirical quantile of group g. The final retained set is $\mathcal{K}(\eta) = \bigcup_g \mathcal{K}_g(\eta)$. Abstentions thus scale per-group, preserving equalised coverage and preventing the model from disproportionately rejecting one group.

For multi-attribute fairness, we extend this scheme to intersectional groups $\{(p_1, \ldots, p_m)\}$ by enumerating their combinations and applying the same per-intersection quantile calibration.

The resulting abstention wrapper functions as a *safety monitor* that controls exposure to unfair or unstable predictions. By rejecting samples where either the evidential contribution of sensitive features is excessively high or confidence drops sharply when these features are masked, the model enforces bounded harm at the cost of reduced coverage. This trade-off can be tuned by policymakers or system integrators, aligning with safety standards in risk-sensitive domains such as justice, finance, or healthcare. Flagged inputs can be subsequently subjected to manual review or processed under alternative fairness criteria, enabling adaptive remediation and further improving fairness beyond abstention alone. In addition, the unified sensitivity score naturally extends to post-processing, where it can prioritise instance flips under Equalized Odds to target the most fairness-sensitive cases (see Sec. 6.4).

5 Experimental Setup

We evaluate EviFair across three widely studied fairness benchmarks–*German Credit*, *Adult Income*, and *COMPAS Recidivism*–each capturing distinct sources of social bias and risk-sensitive decision contexts. The German Credit dataset [9] contains 1,000 credit applications with 20 financial and demographic attributes. We binarise the *age* attribute into "young" (≤ 25) and "old" (> 25) groups [10], and use *sex* as an additional protected feature in the multi-attribute setting. The prediction task is to classify credit risk as good (favourable) or bad (unfavourable). The Adult Income dataset [4] consists of 48,842 records from the 1994 US Census, where the task is to predict whether an individual's income exceeds \$50K per year. We encode categorical variables, remove missing entries, and follow standard binarisation for *sex* (Male/Female) and *race* (White/Non-White). The COMPAS dataset [3] contains criminal justice data used to predict recidivism within two years of release. We retain defendants identified as African-American or Caucasian. The protected attributes are *race* and *sex*, enabling both single- and intersectional-group evaluations. Each dataset is randomly split into 70% training and 30% testing sets with stratification by outcome.

The safety wrapper of EviFair comprises two parallel components: (i) a probabilistic classifier $f_\theta(x)$, and (ii) an evidential bias head $g_\phi(x)$ that models per-feature, per-class Dirichlet evidence. The classifier is implemented as a multilayer perceptron with hidden layers of sizes (128, 64), ReLU activations, and a dropout rate of 0.2. The evidential head mirrors the input dimensionality d and the number of classes K, producing positive evidence values $e_{i,k} = \text{softplus}(w_{i,k}x_i + b_{i,k})$ for each feature–class pair. Both networks are trained concurrently on the same minibatches using Adam optimisers with learning rates 10^{-3}. The classifier minimises the standard cross-entropy loss, while the evidential head minimises the Dirichlet negative log-likelihood with a KL regularisation term weighted by $\lambda_{\text{KL}} = 10^{-2}$. Early stopping with patience 10 and a maximum of 80 epochs is employed to prevent overfitting. During training, random feature masking with probability 0.1 is applied to the evidential head to encourage disentangled evidence allocation, thereby improving its sensitivity to protected attributes.

Evaluation covers three axes: predictive performance, group fairness, and abstention coverage. Performance is reported as Accuracy and F1 averaged over retained (non-abstained) samples. Fairness is measured using Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), Average Odds Difference (AOD), and Disparate Impact (DI) [8, 6, 5]; in the multi-attribute setting we report metrics per attribute and, where relevant, across intersectional groups. In deployment, EviFair acts as a flagger where predictions with high fairness sensitivity are withheld. In our experiments we emulate this by *abstention*: let coverage $\eta \in [0, 1]$ denote the retained fraction (abstention rate $1 - \eta$). For each target η, we enforce equalised coverage by selecting per-group thresholds (Section 4.4) and form the retained set $\mathcal{K}(\eta)$; performance and fairness are computed on $\mathcal{K}(\eta)$. Varying η traces the coverage–performance and coverage–fairness trade-off curves used in our analysis.

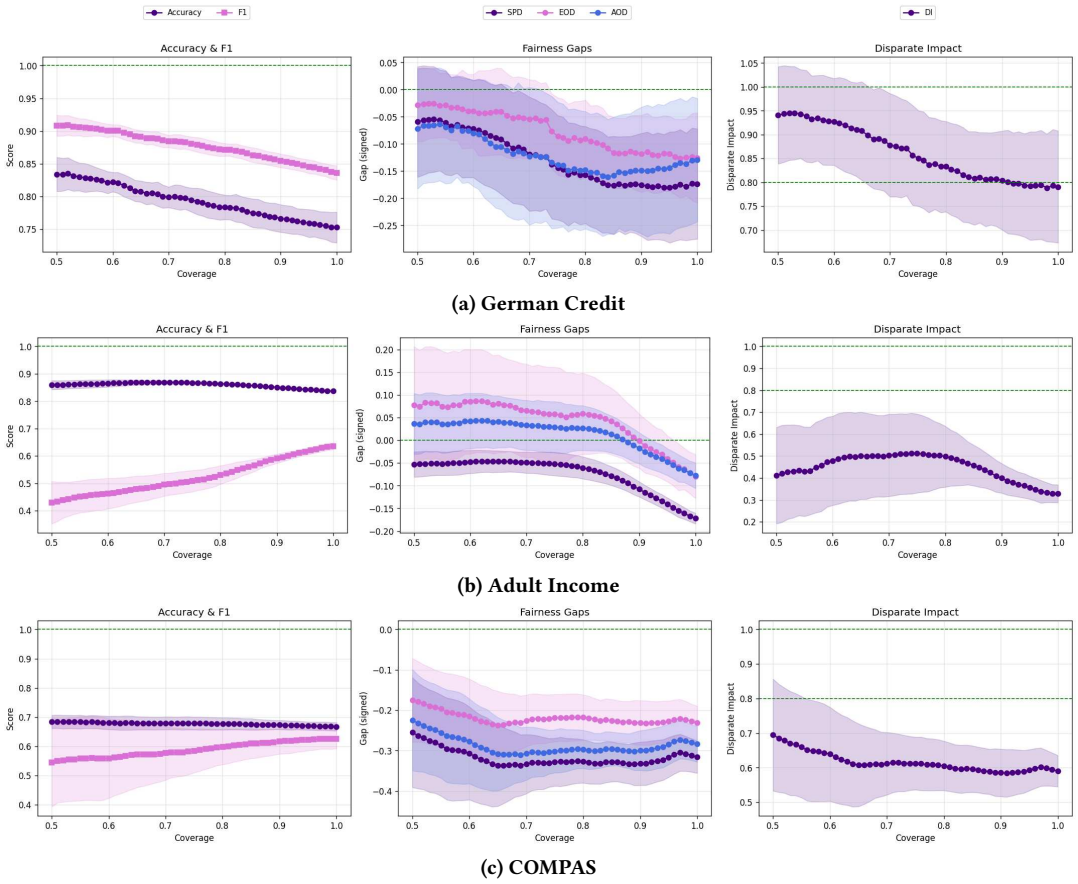
6 Evaluation

6.1 Single-Attribute Fairness Evaluation

Figure 2 reports the coverage–fairness–performance trade-offs obtained by EviFair on the German Credit, Adult Income, and COMPAS datasets. Each curve quantifies the change in accuracy, F1, and fairness metrics as the abstention threshold is relaxed, revealing how fairness and performance trade off across coverage levels.

Across all datasets, stricter abstention consistently enhances fairness without substantially compromising predictive performance. In the German Credit dataset, fairness gaps (SPD, EOD, AOD) shrink sharply as coverage decreases, with only a moderate reduction in accuracy. The Adult Income task exhibits a similar trend, though with less pronounced disparities at full coverage, reflecting that the model already exhibits comparatively balanced group performance. For COMPAS, the fairness improvements are strongest, with equal opportunity and disparate impact showing substantial gains at reduced coverage, demonstrating that EviFair can effectively flag and remove bias-sensitive predictions in highly imbalanced social domains. Empirically, both evidential and classifier sensitivity paths are triggered, so taking the max over paths ensures EviFair responds whenever either source flags fairness risk.

These results confirm that evidential sensitivity to protected attributes provides a reliable proxy for fairness risk. By flagging samples with large sensitivity scores, EviFair functions as a safety monitor that constrains exposure to potentially discriminatory decisions without altering the base classifier. Unlike conventional fairness regularisers, which modify the decision boundary and may induce overcorrection, this post hoc mechanism preserves the model's predictive integrity while enabling explicit control over the



(a) German Credit

(b) Adult Income

(c) COMPAS

Figure 2: Coverage–fairness–performance trade-offs across datasets. Left: predictive performance (Accuracy, F1) as a function of coverage η, where lower coverage corresponds to stricter abstention thresholds. Middle: group fairness gaps (SPD, EOD, AOD); values closer to 0 indicate improved fairness. Right: Disparate Impact (DI); values approaching 1 reflect demographic parity. Shaded regions denote the standard deviation over ten random seeds.

fairness–coverage balance. In deployment, flagged inputs could be re-examined, corrected, or reweighted according to domain-specific fairness criteria, allowing adaptive compliance with different regulatory or ethical standards. This flexibility supports integration into human-in-the-loop pipelines, where fairness interventions can be selectively applied to critical or high-risk cases. Together, these findings demonstrate that EviFair provides measurable fairness gains alongside a configurable safeguard for responsible, bias-aware model operation.

Table 1 reports results for each dataset at four representative coverage levels: full coverage ($\eta = 1.0$), flight flagging ($\eta = 0.9$), moderate flagging ($\eta = 0.75$), and strong flagging ($\eta = 0.5$). This selection spans practical operating ranges for deployed systems while illustrating the full fairness-coverage trade-off observed in Figure 2. The results in Table 1 highlight the controllable trade-off between fairness and predictive performance under different flagging regimes. As coverage decreases, parity gaps shrink consistently across datasets, confirming that EviFair effectively identifies bias-sensitive predictions. While $\eta = 0.5$ represents an intentionally conservative setting included for completeness, practical deployments would typically target $\eta \geq 0.9$, where fairness gains are already achieved with negligible accuracy loss. It is important to note that these results correspond to the abstention-based evaluation protocol, in which flagged samples are withheld to quantify the achievable fairness–coverage frontier. In operational use, flagged instances would instead be re-examined, corrected, or reweighed according to domain-specific fairness objectives. Such adaptive re-evaluation could further enhance fairness at higher coverage levels, effectively reducing the proportion of cases requiring abstention to achieve the same parity improvements observed at harsher coverage thresholds.

6.2 Multi-Attribute Fairness Evaluation

To assess whether EviFair generalises beyond single protected attributes, we extend the analysis to intersectional fairness settings in which multiple sensitive features are considered jointly (e.g., *age* and *sex* in German Credit, *race* and *sex* in Adult and COMPAS). The evidential and classifier components remain unchanged;

Table 1: Performance and fairness metrics across coverage levels. Results are reported as mean ± standard deviation across ten random seeds, and are shown for full coverage ($\eta = 1.0$), light (0.9), moderate (0.75), and strong (0.5) flagging regimes. Bold values indicate improved fairness relative to the baseline.

Dataset	Coverage (η)	Accuracy	F1	SPD	EOD	AOD	DI
German Credit	0.50	**0.834 ± 0.026**	**0.908 ± 0.016**	**-0.059 ± 0.102**	**-0.028 ± 0.068**	**-0.072 ± 0.111**	**0.941 ± 0.102**
	0.75	**0.793 ± 0.020**	**0.879 ± 0.011**	**-0.144 ± 0.096**	**-0.086 ± 0.071**	−0.139 ± 0.116	**0.850 ± 0.100**
	0.90	**0.767 ± 0.021**	**0.855 ± 0.010**	−0.175 ± 0.095	**-0.118 ± 0.071**	−0.149 ± 0.117	**0.804 ± 0.104**
	1.00	0.753 ± 0.024	0.836 ± 0.011	−0.173 ± 0.101	−0.125 ± 0.083	−0.129 ± 0.114	0.791 ± 0.117
Adult Income	0.50	**0.859 ± 0.016**	0.429 ± 0.077	**-0.053 ± 0.028**	**0.078 ± 0.129**	**0.037 ± 0.067**	**0.411 ± 0.220**
	0.75	**0.867 ± 0.004**	0.508 ± 0.039	**-0.052 ± 0.026**	**0.058 ± 0.104**	**0.029 ± 0.056**	**0.512 ± 0.180**
	0.90	**0.851 ± 0.003**	0.590 ± 0.017	**-0.108 ± 0.017**	**-0.001 ± 0.060**	**-0.017 ± 0.034**	**0.399 ± 0.068**
	1.00	0.836 ± 0.004	0.636 ± 0.012	−0.172 ± 0.012	−0.080 ± 0.049	−0.078 ± 0.028	0.329 ± 0.041
COMPAS	0.50	**0.681 ± 0.019**	0.568 ± 0.109	**-0.303 ± 0.119**	**-0.205 ± 0.111**	**-0.268 ± 0.096**	**0.640 ± 0.151**
	0.75	**0.681 ± 0.014**	0.606 ± 0.060	−0.369 ± 0.066	−0.251 ± 0.075	−0.334 ± 0.052	0.566 ± 0.087
	0.90	**0.675 ± 0.013**	0.623 ± 0.036	−0.336 ± 0.033	−0.235 ± 0.043	−0.299 ± 0.029	0.577 ± 0.052
	1.00	0.669 ± 0.012	0.626 ± 0.025	−0.311 ± 0.033	−0.228 ± 0.045	−0.276 ± 0.036	0.596 ± 0.045

the same trained models are evaluated under an extended masking and calibration procedure. In this multi-attribute setting, both sensitivity paths are computed by jointly masking all protected features, thereby capturing aggregate dependence on their combined effect rather than treating each attribute in isolation. Equalised coverage is then enforced over the Cartesian product of protected attributes, ensuring that intersectional subgroups (e.g., "young female" or "non-White male") receive balanced abstention rates. This formulation provides a stricter test of fairness robustness, revealing whether EviFair can maintain bias-aware abstention when sensitive attributes interact or overlap.

The intersectional evaluation reveals that the fairness–coverage trade-off observed in single-attribute experiments persists under multi-attribute conditions. As coverage decreases, parity gaps across combined groups (e.g., age–sex or race–sex) are substantially reduced, confirming that EviFair remains effective when multiple protected dimensions interact. Although the compound attribute structure introduces additional imbalance—particularly in the Adult and COMPAS datasets—EviFair achieves stable fairness improvements without significant degradation in accuracy. This demonstrates that evidential sensitivity generalises beyond single protected features, capturing fairness dependencies distributed across intersecting demographic factors.

6.3 Random Equalised Abstention

To test whether the observed fairness improvements arise merely from the act of abstaining on a fixed proportion of samples per group, rather than from the model's targeted selection of biased instances, we conduct a control experiment. In this setting, the same per-group coverage proportions as in our method are enforced, but the specific samples removed at each coverage level are chosen uniformly at random within each subgroup. This random equalised abstention preserves the coverage structure while eliminating any bias-aware selection signal, allowing us to isolate the causal effect of which samples are abstained upon. The resulting fairness and accuracy curves thus serve as a baseline to demonstrate that improvements in parity metrics stem from informed abstention decisions rather than arbitrary data exclusion.

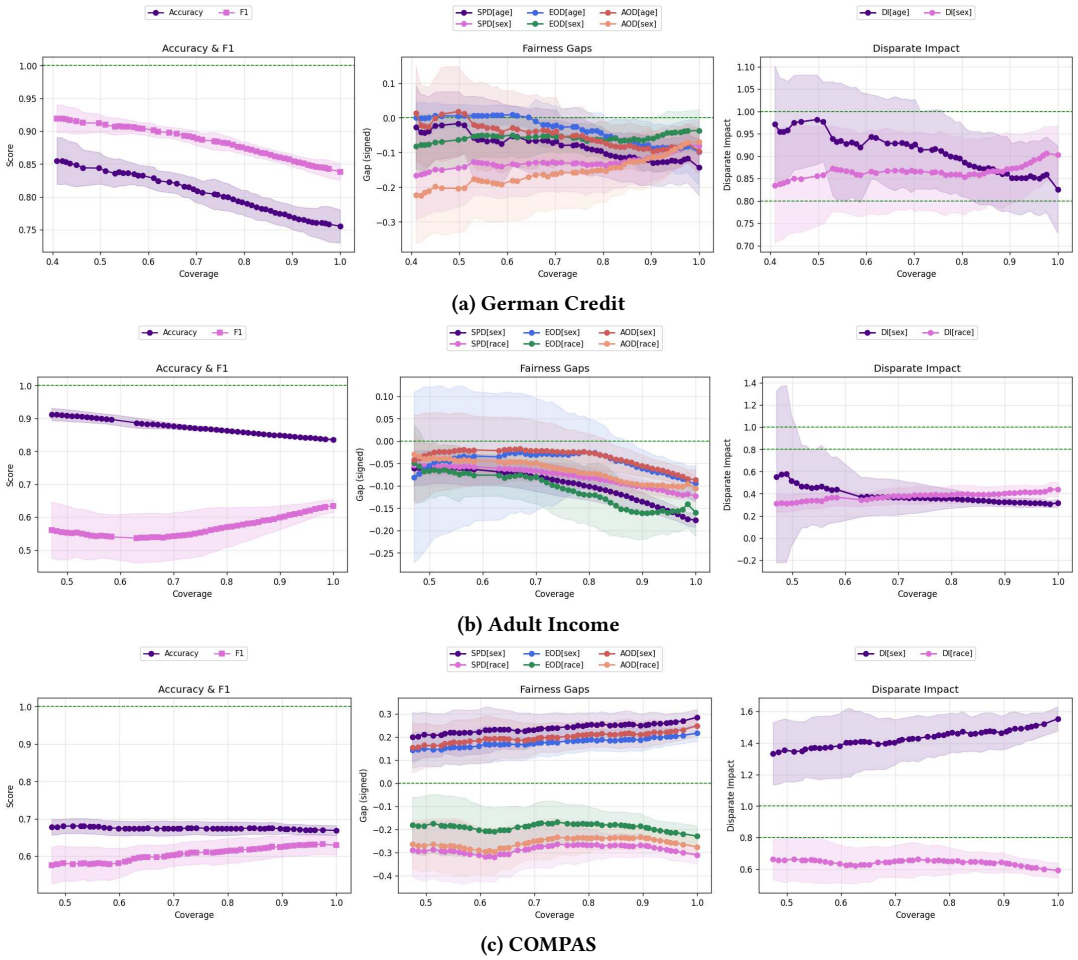
Figure 3 reports the results of the random equalised abstention control. Across all datasets, the fairness metrics remain largely unchanged as coverage decreases, with only minor fluctuations attributable to sampling noise. Neither statistical parity difference nor equal opportunity difference exhibits the monotonic improvement observed under EviFair, and disparate impact stays roughly constant across the entire coverage range. Performance metrics (accuracy and F1) also remain stable, confirming that random abstention neither benefits nor harms model utility in a systematic way.

These findings validate that the fairness gains achieved by EviFair are not a result of discarding data, but instead arise from its ability to identify and selectively abstain on bias-sensitive samples. By comparing against this randomised baseline, we demonstrate that evidential sensitivity provides meaningful signal about fairness risk, enabling abstention decisions that actively reduce group disparities while maintaining high predictive performance.

6.4 EviFair-Ranked Post-Processing

To test EviFair's ability to prioritise the most problematic decisions, we run a post-processing mitigation strategy outcome flip study study: for each dataset, we fit Equalized Odds post-processing [8] on the trained model's test predictions to obtain per-group, per-direction flip counts/budgets, and then compare (i) the method's native flip selection with (ii) an EviFair variant that preserves the same per-cell counts but chooses which instances to flip by ranking from the bias score.

Across all datasets, ranking flips by our bias score while keeping Equalized Odds' per-group/per-direction budgets fixed, yields better accuracy-fairness trade-offs and more stable trajectories than Equalized Odds' native instance selection. On Adult Income, our targeted flips reach the EOD/AOD targets at smaller flip fractions, while maintaining higher accuracy, showing that EviFair can help post-processing methods hit fairness goals with less intervention.



(a) German Credit

(b) Adult Income

(c) COMPAS

Figure 3: Coverage–fairness–performance trade-offs for the multi-attribute experiments across datasets. Left: predictive performance (Accuracy, F1). Middle: group fairness gaps (SPD, EOD, AOD). Right: Disparate Impact (DI). Shaded regions denote the standard deviation over ten random seeds.

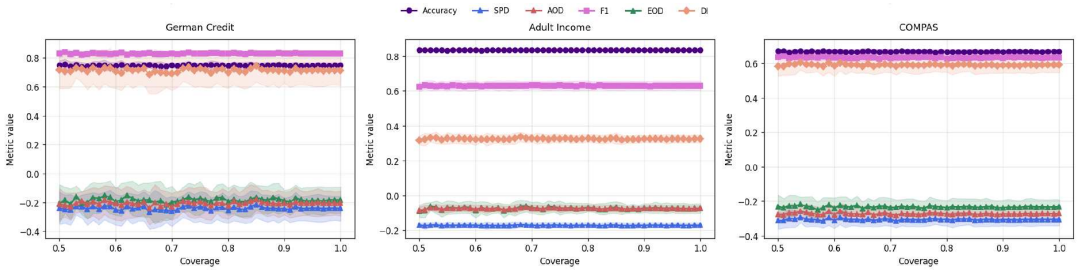
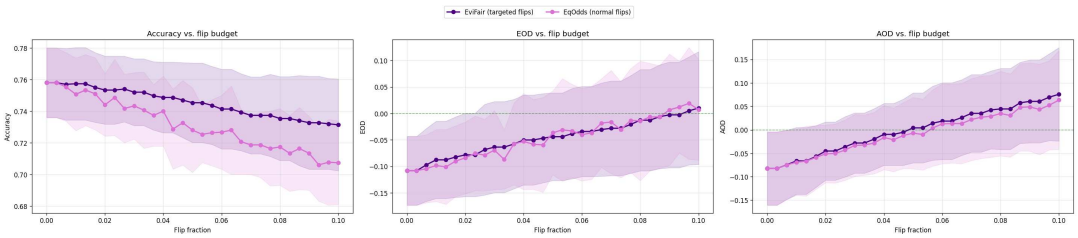
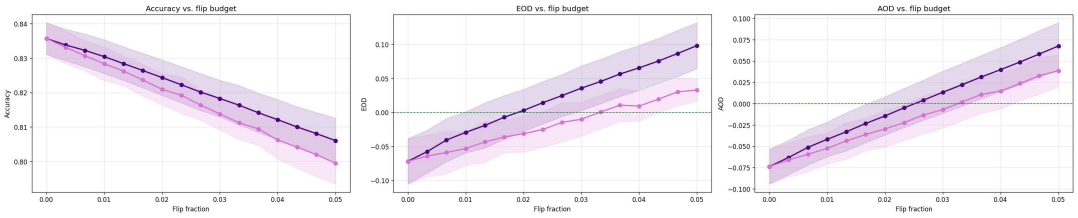


Figure 4: Coverage–fairness–performance trade-offs for the random equalised abstention experiment. Left: German Credit. Middle: Adult Income. Right: COMPAS. Shaded regions denote the standard deviation over ten random seeds.

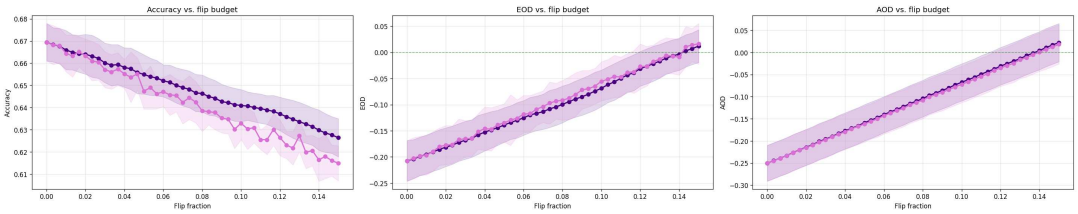
On German Credit and COMPAS, the flip budgets needed to reach the fairness thresholds are similar, but there is no downside to EviFair's targeting and for matched budgets it preserves substantially more accuracy (especially on German). In short, for post-processing, the selection of *which* instances are flipped matters, so utilising tools such as EviFair to help make these decisions can mean less intervention and higher accuracy retention.



(a) German Credit



(b) Adult Income



(c) COMPAS

Figure 5: Flip-budget trade-offs across datasets. Left: Accuracy vs. flip fraction. Middle: Equalized Opportunity Difference. Right: Average Odds Difference. Curves compare EviFair targeted flips to Equalised Odds native flips. Shaded regions show the standard deviation over ten random seeds and dashed lines mark fairness targets.

7 Conclusion

This paper presented EviFair, a fairness-aware abstention framework that treats bias sensitivity as a safety problem. By combining evidential uncertainty modelling with feature-path analysis, EviFair identifies predictions whose confidence depends disproportionately on protected attributes and flags them for further scrutiny. Through the use of an equalised coverage policy, it ensures balanced abstention behaviour across groups, converting fairness assurance into an explicit coverage–fairness trade-off. Experiments across benchmark datasets demonstrated that even modest abstention yields substantial reductions in group disparities with minimal loss in predictive performance.

EviFair offers a practical mechanism for responsible model deployment, enabling systems to manage fairness risk dynamically at inference time. Future work will explore integration with real-world decision workflows. These directions aim to advance fairness monitoring toward actionable safety guarantees in machine learning.

References

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

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[4] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20. (1996).

[5] Rachel KE Bellamy et al. 2019. Ai fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63, 4/5, 4–1.

[6] Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2014. Certifying and removing disparate impact. *arXiv preprint arXiv:1412.3756*.

[7] Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

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

[9] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77. (1994).

[10] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*. IEEE, 1–6.

[11] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33, 1, 1–33.

[12] Faisal Kamiran, Sameen Mansha, Asim Karim, and Xiangliang Zhang. 2018. Exploiting reject option in classification for social discrimination control. *Information Sciences*, 425, 18–33.

[13] Joshua K Lee, Yuheng Bu, Deepta Rajan, Prasanna Sattigeri, Rameswar Panda, Subhro Das, and Gregory W Wornell. 2021. Fair selective classification via sufficiency. In *International conference on machine learning*. PMLR, 6076–6086.

[14] Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31.

[15] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54, 6, 1–35.

[16] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: choices, assumptions, and definitions. *Annual review of statistics and its application*, 8, 1, 141–163.

[17] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems*, 30.

[18] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

[19] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.

[20] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.