



Fairness through Aleatoric Uncertainty

Anique Tahir
Arizona State University
Tempe, Arizona, USA
artahir@asu.edu

Lu Cheng
University of Illinois Chicago
Chicago, Illinois, USA
lucheng@uic.edu

Huan Liu
Arizona State University
Tempe, Arizona, USA
huanliu@asu.edu

ABSTRACT

We propose a simple yet effective solution to tackle the often-competing goals of fairness and utility in classification tasks. While fairness ensures that the model's predictions are unbiased and do not discriminate against any particular group or individual, utility focuses on maximizing the model's predictive performance. This work introduces the idea of leveraging aleatoric uncertainty (e.g., data ambiguity) to improve the fairness-utility trade-off. Our central hypothesis is that aleatoric uncertainty is a key factor for algorithmic fairness and samples with low aleatoric uncertainty are modeled more accurately and fairly than those with high aleatoric uncertainty. We then propose a principled model to improve fairness when aleatoric uncertainty is high and improve utility elsewhere. Our approach first intervenes in the data distribution to better decouple aleatoric uncertainty and epistemic uncertainty. It then introduces a fairness-utility bi-objective loss defined based on the estimated aleatoric uncertainty. Our approach is theoretically guaranteed to improve the fairness-utility trade-off. Experimental results on both tabular and image datasets show that the proposed approach outperforms state-of-the-art methods w.r.t. the fairness-utility trade-off and w.r.t. both group and individual fairness metrics. This work presents a fresh perspective on the trade-off between utility and algorithmic fairness and opens a key avenue for the potential of using prediction uncertainty in fair machine learning.

CCS CONCEPTS

• Computing methodologies → Machine learning; • Social and professional topics;

KEYWORDS

fairness, uncertainty quantification, bayesian neural networks

ACM Reference Format:

Anique Tahir, Lu Cheng, and Huan Liu. 2023. Fairness through Aleatoric Uncertainty. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*, October 21–25, 2023, Birmingham, United Kingdom. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3583780.3614875>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0124-5/23/10...\$15.00

<https://doi.org/10.1145/3583780.3614875>

1 INTRODUCTION

Machine learning (ML) algorithms have been widely used in various applications and are becoming increasingly popular in domains such as computer vision, speech recognition, natural language processing, and bioinformatics [33]. Despite their superior performance in terms of prediction accuracy, they have often faced criticism for lacking fairness and discriminating against marginalized groups [53, 23]. Fair ML aims to improve algorithmic fairness. Due to the often competing relation between fairness and utility, a primary challenge in fair ML has been improving the fairness-utility trade-off [23, 37]. Finding a solution that alleviates the trade-off and improves both goals is often deemed impossible yet crucial to ensure that ML algorithms are not only functional but also trustworthy when making predictions [15, 14].

Prior work in fair ML improves training procedures based on certain heuristics (e.g., using an adversary [54]) to achieve a better trade-off [21, 9, 31] (see more works discussed in depth in Section 5). In essence, doing so is analogous to finding a better hypothesis to reduce uncertainty in areas where there is a lack of data or knowledge [34, 18]. This kind of uncertainty is known as *epistemic* or *model uncertainty* [1]. By contrast, this work proposes to explore the connection between fairness and the other kind of predictive uncertainty, known as *aleatoric* [1] or *data uncertainty*, arising from the inherent ambiguity in the data.

Aleatoric uncertainty naturally relates to both algorithmic fairness and utility. When data is ambiguous due to e.g., inherent noise or entangled causal features, we humans tend to make decisions relying on past experience and ambiguous information that might reflect historical inequalities. Similarly, ML models are more likely to make wrong predictions under high aleatoric uncertainty, and even if we train on an infinite amount of data, the model would still be uncertain about the prediction [26]. Therefore, *our central hypothesis is that aleatoric uncertainty is a crucial cause of algorithmic unfairness, and samples with low aleatoric uncertainty are modeled more accurately and fairly than those with high aleatoric uncertainty*. The relation between aleatoric uncertainty and fairness has evaded investigation in the past since aleatoric uncertainty is associated with the impossibility of improvement.

To bridge the gap, this work introduces a simple yet effective approach that leverages aleatoric uncertainty to improve the fairness-utility trade-off with theoretical guarantees. In particular, given the potential confounding effects related to the protected attribute, we first propose effective distributional interventions to prevent noise leakage in uncertainty estimation to enable the disentanglement of aleatoric and epistemic uncertainties. Predictions with low uncertainty tend to be fair while those with high uncertainty tend to be unfair (Section 3.3). Thus, we explicitly model aleatoric uncertainty in the training process: considering heteroscedastic uncertainty (i.e., the uncertainty varies across samples), we prioritize utility

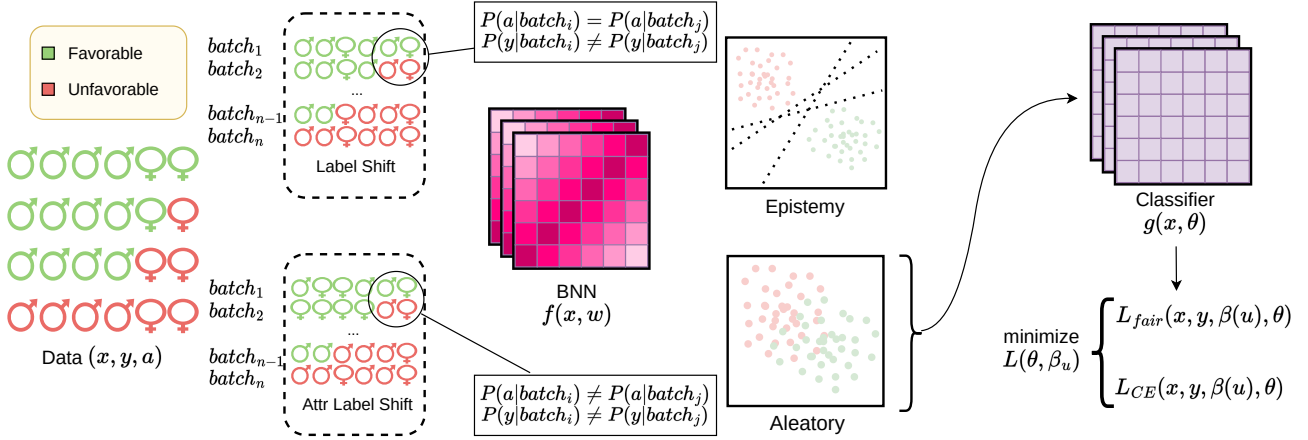


Figure 1: After the distributional intervention, GAIA improves the fairness-utility trade-off by balancing the utility (L_{CE}) and fairness (L_{fair}) loss using aleatoric uncertainty estimated by BNNs.

over fairness when dealing with samples that have low aleatoric uncertainty, and prioritize fairness over utility for samples with high aleatoric uncertainty. The representation of various protected groups is heterogeneous in real-world data. Conventionally, the ground truth labels provided are assumed to be correct. However, ML models learn spurious correlations since subgroups of the population achieve different distributions of favorable or unfavorable outcomes. This results in algorithmic bias. For our approach, we draw a dichotomy between the solution space; (i) where our model is likely to make the correct prediction, resulting in lower algorithmic bias, and (ii) where it is likely to be uncertain, resulting in higher algorithmic bias. By utilizing this knowledge during model training, we can reduce the trade-off between utility and fairness objectives. We evaluate our approach on well-established datasets and compare it to the state-of-the-art baselines that include pre-, in-, and post-processing methods [4, 53]. Experimental results indicate that the proposed approach achieves the best fairness-utility trade-off in terms of both group fairness metrics [19, 53], and shows potential for individual fairness.

In summary, we introduce several important contributions to the field of fairness in ML:

- (i) we provide the first empirical results regarding the relationship among fairness, utility, and aleatoric uncertainty in classification tasks;
- (ii) we propose a simple yet effective approach that leverages aleatoric uncertainty to improve the fairness-utility trade-off with a theoretical guarantee; and
- (iii) we provide empirical evidence of its efficacy on real-world datasets. Experimental results also highlight the importance of distributional intervention for uncertainty estimation that would otherwise lead to algorithmic unfairness.

2 METHODOLOGY

Problem Setting. We consider the standard fair binary classification setting where the samples $X \in \mathcal{X} \subset \mathbb{R}^n$, labels $Y \in \mathcal{Y} = \{0, 1\}$, and protected attribute $A \in \{0, 1\}$ are provided as the input. Our

objective is to train a classifier $g: \mathbb{R}^n \rightarrow [0, 1]$ such that its predictions $\hat{Y} \in [0, 1]$ are accurate i.e., $P(\hat{Y}|X) = P(Y|X)$, and fair across different demographic groups. The proposed approach, **Guided Algorithm for Integrating Aleatory (GAIA)**, draws from the inherent relation between fairness and aleatoric uncertainty due to data ambiguity which leads a model to rely on biased priors. With high aleatoric uncertainty, it becomes infeasible to improve the utility; however, we can still improve fairness since it does not necessarily rely on utility. We empirically and theoretically prove that GAIA improves the fairness-utility trade-off. GAIA consists of three major steps highlighted in the following subsections.

2.1 Distributional Intervention

Traditional ML algorithms use Empirical Risk Minimization (ERM) [17] and rely on the independent and identically distributed (*i.i.d.*) assumption. Prior work shows that distribution shift exacerbates both fairness and predictive performance [48, 43]. In addition, due to the skewed distributions for different protected groups, standard uncertainty estimation methods such as BNNs cannot be directly applied to estimating the model uncertainty given its sensitivity to data imbalance [42]. To assuage this issue, we intervene in the data distribution and identify two instances of data bias that can be controlled: the label distribution is skewed resulting in the model relying on (i) the prior distribution of the label (*Label Shift*), and (ii) the spurious correlation between the protected attribute and the label (*Attribute Label Shift*). An example for (i) is when the majority of the data has a specific label (e.g., non-fraud transactions in fraud detection). Here, a trained model may rely on shortcut learning [22] to predict the majority label. Similarly, for (ii), a trained model may rely on the protected attribute for prediction if it displays a significant correlation with the label. If the protected attribute is correlated with the label, the non-protected covariates affected by the protected attribute also resonate with the correlation. Thus, if we intervene in the correlation between the protected attribute and the label, it also results in the intervention of the factors resonating with the protected attributes in the non-protected covariates.

Distribution intervention can mitigate unfairness and lead to better uncertainty disentanglement. Note that this step can be replaced with other heuristics for achieving better utility and fairness as highlighted in Section. 5. This is because using a good heuristic reduces epistemic uncertainty, leading to better estimation of aleatoric uncertainty.

2.1.1 Label Shift. Label shift aims to change the distribution of the labels in every mini-batch during training. This will result in a model that does not favor the majority label in the original data distribution. Formally, let $(X, Y) \in \mathcal{D}$ be instances in the dataset $\mathcal{D}$, where X denotes the feature matrix and $Y \in \{0, 1\}$ denotes the binary label vector. We define the sets of indices $M_1 = \{i \in \mathcal{D} \mid Y_i = 1\}$ and $M_0 = \{i \in \mathcal{D} \mid Y_i = 0\}$, corresponding to samples with favorable (e.g., low credit risk) and unfavorable outcomes, respectively. $|M_1| = n_1$ and $|M_0| = n_0$.

A random percentage of favored samples, p , is determined by sampling randomly from the uniform distribution $\mathcal{U}(0, 1)$. We then define the scaled sets of indices $M'_1 = \{i \in M_1 \mid p\}$ and $M'_0 = \{i \in M_0 \mid 1 - p\}$. These sets are used to calculate the probability of selecting each sample, $P_i^1 = \frac{|i \in M'_1|}{n_1}$, $P_i^0 = \frac{|i \in M'_0|}{n_0}$. A batch of size m is selected from the dataset by randomly picking samples without replacement according to the probabilities P_i . We denote the set of indices of the selected samples by $I = \{i_1, i_2, \dots, i_m\} \subset \mathcal{D}$. This results in a counterfactual batch of training samples with the intervention of label distribution (**LabelShift, LS**).

2.1.2 Attribute Label Shift. Intervening only on the label distribution may be insufficient to reduce the spurious correlations in the data. We further intervene on the protected attribute to resolve its confounding effect. However, this is often infeasible with observational data. Therefore, we introduce an estimation of intervention by changing the correlation of protected attribute and label distributions across different mini-batches during training. The underlying assumption is that there is a sufficient number of non-causal factors in the covariates such that the interventional changes are large enough for the model to distinguish the non-causal factors from the causal ones. Thus, Attribute Label Shift aims to intervene on both the protected attribute a and the label y . Let $M_{p^1} = \{i \in \mathcal{D} \mid a_i = 1\}$ and $M_{p^0} = \{i \in \mathcal{D} \mid a_i = 0\}$ be the sets of indices for samples belonging to the protected group and non-protected group, respectively. $n_{p^1} = |M_{p^1}|$ and $n_{p^0} = |M_{p^0}|$.

A random percentage p_1 of samples from the protected group is determined by sampling randomly from the uniform distribution $\mathcal{U}(0, 1)$. We then define the scaled sets of indices $M'_{p^1} = \{i \in M_{p^1} \mid p_1\}$ and $M'_{p^0} = \{i \in M_{p^0} \mid 1 - p_1\}$. These sets are used to calculate the probability of selecting each sample $P_{p,i}$ from the protected or non-protected group, $P_{p^1,i} = \frac{|i \in M'_{p^1}|}{n_{p^1}}$, $P_{p^0,i} = \frac{|i \in M'_{p^0}|}{n_{p^0}}$. Similarly, the probability of selecting each sample $P_{f,i}$ from the favored or unfavored class is defined as $P_{1,i} = \frac{|i \in M'_1|}{n_1}$, $P_{0,i} = \frac{|i \in M'_0|}{n_0}$. The final probability of selecting each sample is the product, $P_i = P_{p,i} * P_{f,i}$. This gives us a batch of training samples with interventions on the correlation of protected attribute and label (**AttrLabelShift, ALS**).

2.2 Decoupling Aleatoric and Epistemic Uncertainty

GAIA uses BNN via backpropagation (Bayes by Backprop) [7] to conveniently decouple aleatoric from epistemic uncertainty while also maintaining its ability to be integrated into existing neural architectures. Bayes by Backprop is computationally efficient and theoretically sound. Given C classes, aleatoric uncertainty is formulated as the expected entropy for the prediction [1, 26],

$$H_{\text{alea}}(\mathbf{x}) = \int_{\theta} \sum_i^C -p(y_i|\mathbf{x}, \theta) \log p(y_i|\mathbf{x}, \theta) d\theta, \quad (1)$$

where $p(y_i|\mathbf{x}, \theta)$ is the predictive probability of the i -th class from the model parameterized by θ . Epistemic uncertainty is represented by the model's predictive variance [1],

$$\sigma_{\text{epi}}^2(\mathbf{x}) = \text{Var}_j[p(y|\mathbf{x}, \theta_j)], \quad (2)$$

where j denotes the j -th sample of the BNN weights. BNN involves finding the maximum a posteriori (MAP) weights:

$$\theta^{MAP} = \arg \max_{\theta} \log P(\theta|\mathcal{D}) = \arg \max_{\theta} P(\mathcal{D}|\theta) + \log P(\theta). \quad (3)$$

The final prediction of BNNs is the expected value of the predicted label $\hat{y}$ for an unseen sample $\hat{\mathbf{x}}$ over the posterior distribution of the weights, $P(\theta|\mathcal{D})$ i.e., $P(\hat{y}|\hat{\mathbf{x}}) = \mathbb{E}_{P(\theta|\mathcal{D})}[P(\hat{y}|\hat{\mathbf{x}}, \theta)]$. We can then utilize each candidate prediction, $P(\hat{y}|\hat{\mathbf{x}}, \theta_j)$, where $\theta_j \sim P(\theta|\mathcal{D})$ to efficiently evaluate both aleatoric and epistemic uncertainties using Eq. 1 and 2, respectively.

For tractable estimation, the common practice in variational inference estimates the posterior using a surrogate, $q(\theta|w)$, by minimizing the Evidence Lower Bound (ELBO) loss [52]. Further, we assume heteroscedastic uncertainty, i.e., uncertainty varies across different samples [10], given its practicability. Hence, the uncertainty metrics between predictions are on a per-sample basis. By explicitly modeling aleatoric and epistemic uncertainty, GAIA traces whether the uncertainty stems from ambiguity or lack of data.

2.3 Improving Fairness-Utility Trade-off

The goal of GAIA is leveraging aleatoric uncertainty to bridge the gap between fairness and accuracy based on the hypothesis that samples with low aleatoric uncertainty are modeled more accurately and fairly than those with high uncertainty. Thus, to achieve a better trade-off, we design a model to improve fairness when aleatoric uncertainty is high and improve utility elsewhere. We first describe the function $\beta(u) : \mathbb{R}^m \rightarrow \mathbb{R}^m$ that assigns weights to samples based on the estimated aleatoric uncertainty u :

$$\beta(u) = \left(\frac{u - u_{\min}}{u_{\max} - u_{\min}} \right)^k, \quad (4)$$

where the hyper-parameter k helps to weigh one objective in favor of the other, and $u_{\min}$ and $u_{\max}$ are two hyperparameters to normalize the weights. The overall objective function of GAIA (Eq. 7) is a bi-objective loss corresponding to both utility and fairness. It maximizes utility for the samples with low aleatoric uncertainty; for samples with high aleatoric uncertainty, there is little improvement to be made in terms of utility due to the inherent ambiguity of the data. Thus, the aim of GAIA is to steer the objective toward improving the fairness of samples with high aleatoric uncertainty.

Given a batch of training data $\mathcal{S} \subseteq \mathcal{D}$ and a classifier parameterized by θ , the utility loss is a weighted cross-entropy loss:

$$\mathcal{L}_{CE}(\mathcal{S}, \beta) = -\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \beta_i (y_i \log(p(y_i|x_i, u_i)) + (1 - y_i) \log(1 - p(y_i|x_i, u_i))), \quad (5)$$

where y_i is the label for sample x_i and $\beta_i = \beta(u_i)$. The conditioning on prediction $p(y_i|x_i, u_i)$, allows the model to make an informed choice based on the uncertainty. We define fairness as the difference in the mean cross-entropy between instances of different protected attributes. We show in Section 3 that our proposed metric acts as a feasible surrogate to cover common group fairness metrics. Let $\mathcal{S}_0$ and $\mathcal{S}_1$ be the sets of samples whose protected attribute is 0 and 1, respectively. We define fairness as follows:

$$\mathcal{L}_{fair}(\mathcal{S}, 1 - \beta) = |\mathcal{L}_{CE}(\mathcal{S}_0, 1 - \beta) - \mathcal{L}_{CE}(\mathcal{S}_1, 1 - \beta)| \quad \mathcal{S}_0 \cup \mathcal{S}_1 = \mathcal{S}. \quad (6)$$

The objective function of GAIA, $\mathcal{L}$, is the sum of Eq. 5 and Eq. 6:

$$\mathcal{L}(\mathcal{S}, \beta) = \mathcal{L}_{CE}(\mathcal{S}, \beta) + \mathcal{L}_{fair}(\mathcal{S}, 1 - \beta). \quad (7)$$

3 THEORETICAL GUARANTEE TO IMPROVE THE TRADE-OFF

In this section, we theoretically prove GAIA can guarantee to improve the fairness-accuracy trade-off through the following three key hypotheses: (i) as aleatoric uncertainty increases, accuracy will decrease; (ii) we can improve fairness in regions of high aleatoric uncertainty; and (iii) binary cross-entropy (BCE) difference across separate protected groups (Eq. 6) is proportional to common group fairness metrics such as equal opportunity difference (EOD) and average odds difference (AOD). The proof consists of two propositions. First, we show divergence on the optimal utility under aleatoric uncertainty. Second, we show the convergence for fairness under BCE difference between protected and non-protected groups. As per convention from the problem setting and for the sake of simplicity, we consider the binary classification case. We use AOD for illustration and similar formulation extends to other group fairness metrics such as EOD.

3.1 Relation between Aleatoric Uncertainty and Accuracy

THEOREM 3.1. *As the aleatoric uncertainty increases, the model's accuracy approaches random chance:*

$$\lim_{\mathbb{E}[H[q(y|x)]] \rightarrow \inf} \text{accuracy} = \frac{1}{C},$$

where C is the number of classes.

Proof of Theorem 3.1. We first define the predictive entropy for the model. Let $p(y|x)$ be the predicted probability distribution of the target class y given the input instance x . In a binary classification problem where $y \in \{0, 1\}$, the expected predictive entropy is the average predictive entropy over all instances in the dataset. This represents the aleatoric uncertainty (Eq. 1).

Next, we will show that the lower bound on the accuracy approaches random chance as the expected predictive entropy increases. In binary classification, random chance corresponds to an

accuracy of $1/2$, suggesting that the model is not better than random guessing. We first derive a lower bound on the accuracy using Fano's inequality [45]. Fano's inequality relates the conditional probability of error in predicting the target class y given the input instance x with the mutual information between y and x :

LEMMA 3.2 (FANO'S INEQUALITY).

$$H(\epsilon) + \epsilon \log(C - 1) \geq H(Y|X),$$

where $H(\epsilon)$ is the binary entropy function of ϵ , the probability of error in predicting the target class, and $H(Y|X)$ is the conditional entropy of the true conditional probability distribution. In a binary classification problem, $C = 2$ and we can simplify Fano's inequality as follows:

$$H(\epsilon) + \epsilon \log(1) \geq H(Y|X). \quad (8)$$

Since $\log(1) = 0$, the inequality becomes:

$$H(\epsilon) \geq H(Y|X). \quad (9)$$

The probability of error ϵ is related to the accuracy by the following relationship:

$$\epsilon = 1 - \text{Accuracy}. \quad (10)$$

We can then reformulate Fano's inequality in terms of accuracy:

$$H(1 - \text{Accuracy}) \geq H(Y|X). \quad (11)$$

Since the binary entropy function $H(p)$ is a monotonically increasing function for $0 \leq p \leq 1/2$ and a monotonically decreasing function for $1/2 \leq p \leq 1$, the maximum entropy is achieved when $p = 1/2$. Thus, the entropy of the error probability is maximized when the accuracy is at random chance:

$$H(1 - 1/2) = H(1/2) = 1. \quad (12)$$

Therefore, as the expected predictive entropy $\mathbb{E}[H[q(y|x)]]$ increases, the lower bound on the accuracy given by Fano's inequality approaches the maximum entropy state, which corresponds to random chance.

3.2 Relation between BCE Loss Difference and Fairness

THEOREM 3.3. *The expected difference in BCE losses between the protected and non-protected groups defined in Eq. 6 is proportional to the Average Odds Difference (AOD).*

$$\mathbb{E}[\Delta L(y)] = \frac{1}{N_1} \sum_{i \in A_1} \Delta L(y_i) - \frac{1}{N_2} \sum_{j \in A_2} \Delta L(y_j) \propto \text{AOD}.$$

Proof of Theorem 3.3. Let us denote the protected attribute instances as A_1 and A_2 . Let p_i be the predicted probability of the positive class $y = 1$ for instances in the group with protected attribute A_i , where $i \in \{1, 2\}$.

PROPOSITION 3.1. *The Binary Cross-Entropy (BCE) loss for instances with protected attribute A_i is given by*

$$L_i(y, p_i) = -y \log(p_i) - (1 - y) \log(1 - p_i).$$

This proposition follows directly from the definition of BCE for binary classification problems. For group fairness metrics, we are concerned with True Positive Rate (TPR_i) difference and False Positive Rate (FPR_i) difference between different groups.

LEMMA 3.4 (AVERAGE ODDS DIFFERENCE). *The Average Odds Difference (AOD) between group A_1 and group A_2 is given by*

$$AOD = \frac{|TPR_1 - TPR_2| + |FPR_1 - FPR_2|}{2}.$$

Now, let us analyze the difference between the BCE losses for the protected ($L_1(\cdot)$) and non-protected ($L_2(\cdot)$) groups:

LEMMA 3.5. *The difference in BCE losses between the two protected attribute groups A_1 and A_2 can be expressed as*

$$\begin{aligned} \Delta L(y) &= L_1(y, p_1) - L_2(y, p_2) \\ &= -y \log\left(\frac{p_1}{p_2}\right) - (1-y) \log\left(\frac{1-p_1}{1-p_2}\right). \end{aligned}$$

Let N_1 and N_2 be the total number of instances in the protected A_1 and non-protected groups A_2 , respectively. To prove Theorem 3.3, we compute the expected differences in BCE losses for the true positive and false positive cases separately.

3.2.1 Equal Opportunity Difference and BCE Difference. First, consider the true positive cases where $y = 1$. In this case, $\Delta L(y = 1) = -\log\left(\frac{p_1}{p_2}\right)$ (from Lemma 3.5). The expected difference in BCE losses for true positives in both groups can be expressed as:

$$\begin{aligned} \mathbb{E}[\Delta L(y = 1)] &= \frac{1}{N_1} \sum_{i \in A_1, y_i=1} -\log\left(\frac{p_1}{p_2}\right) - \frac{1}{N_2} \sum_{j \in A_2, y_j=1} -\log\left(\frac{p_1}{p_2}\right) \\ &\propto |TPR_1 - TPR_2| = EOD. \end{aligned} \quad (13)$$

3.2.2 Average Odds Difference and BCE Difference. Next, consider the false positive cases where $y = 0$. In this case, $\Delta L(y) = -\log\left(\frac{1-p_1}{1-p_2}\right)$. The expected difference in BCE losses for false positives in both groups can be expressed as:

$$\begin{aligned} \mathbb{E}[\Delta L(y = 0)] &= \frac{1}{N_1} \sum_{i \in A_1, y_i=0} -\log\left(\frac{1-p_1}{1-p_2}\right) \\ &\quad - \frac{1}{N_2} \sum_{j \in A_2, y_j=0} -\log\left(\frac{1-p_1}{1-p_2}\right) \\ &\propto |FPR_1 - FPR_2|. \end{aligned} \quad (14)$$

Finally, by combining the expected differences in BCE losses for true positive (Eq. 13) and false positive (Eq. 14) cases with Lemma 3.4, we get:

$$\begin{aligned} \mathbb{E}[\Delta L(y)] &= \mathbb{E}[\Delta L(y = 1)] + \mathbb{E}[\Delta L(y = 0)] \\ &\propto |TPR_1 - TPR_2| + |FPR_1 - FPR_2| = AOD \times 2. \end{aligned} \quad (15)$$

Thus, Eq. 15 shows that the expected difference in BCE losses between the two protected attribute groups is proportional to AOD. This implies that minimizing the difference in BCE losses can lead to fairer outcomes with respect to AOD. EOD is a subset of AOD as demonstrated by Eq. 13.

3.2.3 A Closer Look. Here, we elaborate on why Eq. 13 and Eq. 15 hold. From Lemma 3.5, for the true positive cases where $y = 1$, we have $\Delta L(y = 1) = -\log\left(\frac{p_1}{p_2}\right)$. We first analyze the relationship between the expected difference in BCE losses and the TPR for the two protected attribute groups.

Denote the total number of true positive instances for each group as N_1^{TP} and N_2^{TP} , and let TPR_1 and TPR_2 be the true positive rates for the groups A_1 and A_2 , respectively. The expected difference in BCE losses for the true positive instances is represented as

$$\begin{aligned} \mathbb{E}[\Delta L(y = 1)] &= \frac{1}{N_1^{TP}} \sum_{i \in A_1, y_i=1} -\log\left(\frac{p_1}{p_2}\right) \\ &\quad - \frac{1}{N_2^{TP}} \sum_{j \in A_2, y_j=1} -\log\left(\frac{p_1}{p_2}\right). \end{aligned} \quad (16)$$

We reformulate Eq. 16 using TPR values as follows:

$$\begin{aligned} \mathbb{E}[\Delta L(y = 1)] &= \frac{1}{TPR_1 N_1} \sum_{i \in A_1, y_i=1} -\log\left(\frac{p_1}{p_2}\right) \\ &\quad - \frac{1}{TPR_2 N_2} \sum_{j \in A_2, y_j=1} -\log\left(\frac{p_1}{p_2}\right). \end{aligned} \quad (17)$$

Eq. 17 indicates that as the difference between TPR_1 and TPR_2 increases, $\mathbb{E}[\Delta L(y = 1)]$ also increases. This means that if there is a notable difference in the TPR between the two groups, it will result in a substantial dissimilarity in the BCE losses as well. Therefore, we can conclude that the expected difference in BCE losses for the true positive cases, $\mathbb{E}[\Delta L(y = 1)]$, is indeed proportional to the difference in TPR between the two protected attribute groups. Similarly, we can establish the proportionality of the expected difference in BCE losses for false positive cases, $\mathbb{E}[\Delta L(y = 0)]$, to the difference in FPR between the groups. Combining the results for true positive and false positive cases, we demonstrate that the expected difference in BCE losses between the two protected attribute groups is proportional to the AOD, as stated in Theorem 3.3. In other words, the expected difference in BCE losses for true positive cases captures the difference in TPR and FPR between the two protected attribute groups, which is an essential component of common group fairness metrics such as EOD and AOD.

3.3 On the Fairness-Utility Trade-off

Under Theorem 3.3, we show that by minimizing the BCE loss difference in regions of high aleatoric uncertainty, we indirectly improve group fairness, as reducing the loss entails minimizing the disparities across different groups. In these regions, the model's predictions are more susceptible to biases and disparities since it relies on learned priors, leading to unfair predictions. By prioritizing fairness in these regions, we aim to mitigate the adverse effects of aleatoric uncertainty on marginalized groups. As per Theorem 3.1, it is not feasible to improve accuracy in such regions.

For regions of high confidence (i.e., low uncertainty), accuracy converges to 1 (due to the law of large numbers). Thus, when the uncertainty is low, fairness improves. We can achieve fairness by optimizing utility. Based on Lemma 3.4, we have

$$\begin{aligned} \lim_{accuracy \rightarrow 1} AOD &= \frac{|TPR_1 - TPR_2| + |FPR_1 - FPR_2|}{2} \\ &= \frac{|1 - 1| + |0 - 0|}{2} = 0. \end{aligned} \quad (18)$$

According to Theorems 3.1 and 3.3, GAIA targets utility and fairness in the respective regions where the other metric is non-conflicting. This results in the improvement of both utility and fairness while minimizing the trade-off.

4 EXPERIMENTS

In this section, we show empirical evidence of the effectiveness of GAIA. We aim to answer the following research questions:

- **RQ1:** How does GAIA fare against the state-of-the-art baselines in terms of the fairness-utility trade-off?
- **RQ2:** How does empirical evidence support our hypothesis regarding aleatoric uncertainty, fairness, and utility?
- **RQ3:** While designed for group fairness, what role does GAIA play in improving individual fairness?

4.1 Experimental Setup

Experiments are conducted for both tabular and image datasets. For tabular data, we compare GAIA with seven baselines including common pre-processing, in-processing, and post-processing approaches. We use two benchmark tabular datasets and four fairness metrics including both group and individual fairness metrics. In particular, for **RQ. 1-2**, we use EOD and AOD as the group fairness metrics. We use Generalized Entropy Error (GE) [47] and Consistency Score (CS) [53] to measure individual fairness for **RQ. 3**. For utility measure, we use balanced accuracy, which is conventionally used in fairness literature since it captures balanced protected groups. For the image classification task, we use one benchmark dataset and two additional state-of-the-art approaches as baselines to validate the generalizability of GAIA.

Datasets. The benchmark tabular datasets and image dataset for fair machine learning are detailed below:

- **Adult [49]:** This dataset consists of multiple features ranging from work class, age, education, and sex. Each instance has a binary label based on whether an individual's income exceeds \$50,000/yr. This dataset consists of 48,842 samples.
- **German [25]:** This dataset consists of features related to the financial status of individuals. The label represents whether the attributes represent good or bad credit risk. This dataset consists of 1,000 samples.
- **CelebA [35]:** This dataset contains aligned faces of celebrities with annotations of various attributes, such as gender, age, expression, hair type, and attractiveness. This dataset contains 202,599 face images from 10,177 celebrities.

Gender is considered as the protected attribute in each dataset. Features in tabular datasets are binarized, preprocessed, and scaled following Bellamy et al. [4]. Preprocessing for CelebA follows the conventions established by Chuang and Mroueh [16].

Baselines. For tabular data, we compare GAIA against seven well-established baseline approaches. These approaches can be divided into pre-processing, in-processing, and post-processing methods.

- **Reweighting [27]:** Reweighting is a *pre-processing* approach that adjusts the weight assigned to examples in each (group, label) pairing to promote fairness prior to classification.

- **Learning Fair Representations (LFR) [53]:** A *preprocessing* technique aimed at discovering a latent representation that effectively encodes the data while concealing information pertaining to protected attributes.
- **Optimized Preprocessing [9]:** Optimized preprocessing is a *pre-processing* approach that employs a probabilistic transformation to modify both features and labels in the data while considering fairness with respect to groups, minimizing individual distortion, and preserving data integrity.
- **Adversarial Debiasing [54]:** Adversarial debiasing is an *in-processing* technique that trains a classifier to achieve high prediction accuracy while simultaneously reducing the adversary's capacity to infer protected attributes from the predictions. This results in a fair classifier, as the predictions are rendered devoid of any group discrimination information that could be leveraged by the adversary.
- **MetaFair [11]:** An *in-processing* meta-algorithm for fair classification that handles a broad range of fairness constraints, including non-convex linear fractional constraints such as predictive parity.
- **Calibrated Equalized-Odds [41]:** A *post-processing* technique which uses the calibrated predicted scores to adjust the labels towards better equalized-odds.
- **Reject Option Classification (ROC) [28]:** A *post-processing* technique that balances favorable outcomes between privileged and unprivileged groups by altering the decision boundary in regions of the highest uncertainty.

To further examine the effectiveness of the incorporated aleatoric uncertainty, we compare GAIA against its two sub-module variants: **BNN LS** is the uncertainty estimation component where a BNN is trained using Label Shift (Section 2.1.1), and **BNN ALS** where it is trained using Attribute Label Shift (Section 2.1.2).

The baseline methods for tabular data are not designed for image modality. Thus, for fair comparisons, we consider the following two state-of-the-art approaches for fair image classification:

- **FairBatch: [44]** FairBatch seeks to improve the batch selection process through bi-level optimization such that the downstream model achieves improved fairness.
- **FairMixup [16]:** FairMixup uses data augmentation to improve the fairness-utility tradeoff by making the underlying model more generalizable through regularization on interpolates.

Implementation Details. For the sake of simplicity in our experiments, we employ a logistic regression model, which is essentially a multi-layer perceptron (MLP) without any hidden layers. The uncertainties utilized for training the classification model are generated using a BNN that consists of three hidden layers. The activation functions employed for the BNN and MLP are LeakyReLU [36] and ReLU [2], respectively. When necessary, we utilize the Adam optimizer [29]. Both the BNN and MLP are designed using the JAX framework [8] and Oryx [20] for sampling from distributions. For image classification, ResNet-18 [24] is used as the backbone for both the BNN and the final classifier. We provide the source code for our implementation¹.

¹<https://github.com/aniqueahir/GAIA>

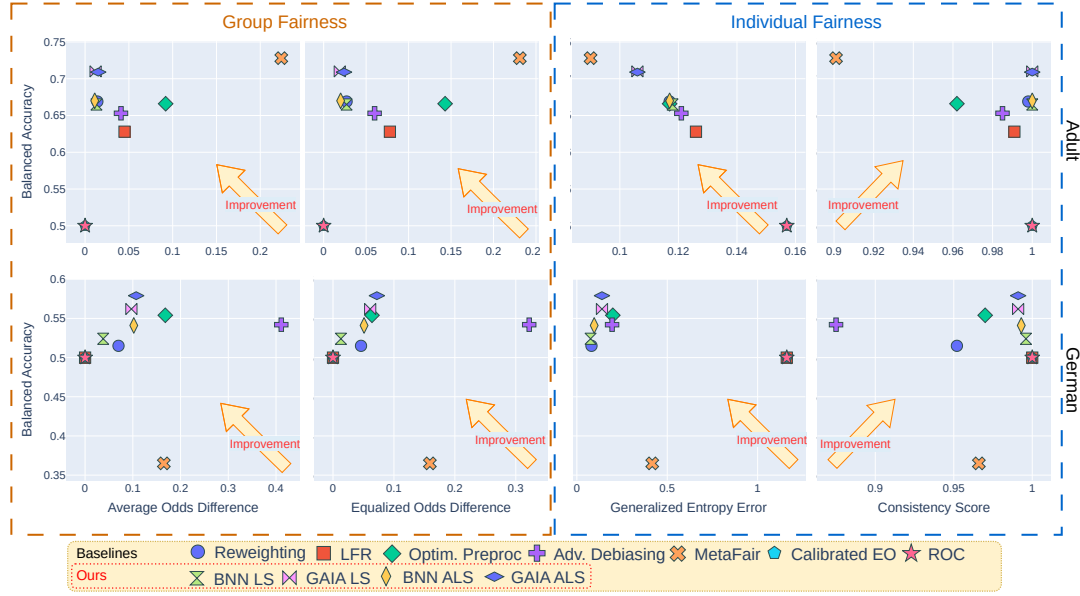


Figure 2: Comparison of Group (left) and Individual (right) Fairness for the Adult and German Datasets. Various approaches fall on different places on the Pareto front representing the fairness-utility trade-off.

To select the best model from training, we use a simple approach: During the training phase, between each mini-batch, we calculate the smoothed training prediction accuracy by using a running average. We select the model parameters corresponding to the best-smoothed accuracy during training for inference. For the baselines, we use standard implementations provided by the AI Fairness 360 Toolkit [4] using the recommended hyper-parameters where needed. For image baselines, we follow the open-source code provided by the authors, respectively [44, 16].

4.2 Experimental Results

Tabular Data. We present the experimental results for **RQ1** regarding the trade-off between models’ utility and fairness. We visualize the comparison of Pareto fronts regarding group fairness in Fig. 2 (left). Our model displays pareto dominance in most of the cases overall. We observe that the in-processing approaches (Adversarial Debiasing, MetaFair) prefer fairness over utility. In contrast, pre- (Reweight, LFR, Optimized Preprocessing) and post-processing (Calibrated EO, ROC) approaches have a more balanced trade-off. We also observe a difference in the trade-off across the Adult and German datasets due to variations in their sample sizes. The Adult dataset (~48k samples) is significantly larger compared to the German dataset (1,000 samples). This may cause each method to perform distinctly from the perspective of the fairness-utility trade-off.

For the Adult dataset, we see a smaller disparity between the performance for versions of our approach using Label Shift (LS) and Attribute Label Shift (ALS). We hypothesize that this is due to the larger size of the Adult dataset compared to the German dataset. The larger dataset size allows the model to make better generalizations and reduce the uncertainty overall. Thus, the shift used in the BNN training is less relevant. By comparison, we see a more diverse performance for the German dataset. The ALS counterparts of both

	FairBatch	FairMixup	GAIA
Bal Acc $\uparrow$	0.562 (0.138)	0.549 (0.035)	0.602 (0.065)
AOD $\downarrow$	0.047 (0.105)	0.041 (0.032)	0.108 (0.068)
EOD $\downarrow$	0.035 (0.077)	0.044 (0.040)	0.021 (0.018)
GE $\downarrow$	0.086 (0.070)	0.260 (0.144)	0.079 (0.022)

Table 1: GAIA shows an overall improvement over baselines w.r.t. balanced accuracy, group (AOD and EOD), and individual fairness (GE) metrics on CelebA image dataset.

BNN and GAIA outperform LS in terms of utility. However, we see slightly better fairness from the LS counterparts. We believe this is due to the LS versions falling closer towards random chance which increases fairness since instances of the protected attribute are treated equally random. For GAIA LS and GAIA ALS, the disparity between fairness is less pronounced since both versions perform comparatively better than random chance.

Fig. 2 also illustrates the value of uncertainty-guided training in GAIA which considers a weighted sum of utility and fairness objectives. Even though BNN with distribution shift (BNN LS and BNN ALS) by itself shows competitive performance compared to the baselines, GAIA consistently outperforms the BNN in terms of utility while matching it in terms of fairness. This improvement is more pronounced in the Adult dataset, where there are more samples for GAIA to leverage the disparity between ambiguous and non-ambiguous subsets of data. Our results highlight the viability of GAIA in improving the fairness-utility trade-off (**RQ1**).

Image Data. To analyze the generalizability of our approach, we also evaluate its performance in the image domain using the Celebrity Faces dataset (CelebA) [35]. We do not report the Consistency Score

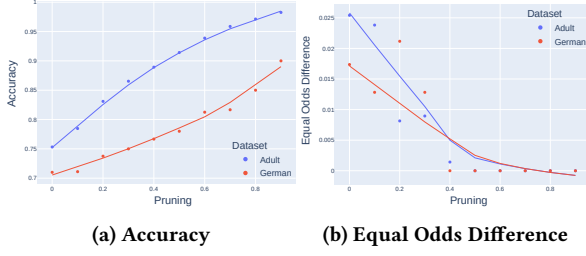


Figure 3: Pruning the most uncertain samples leads to an improvement in both utility and fairness for the Adult and German datasets. We make similar findings across datasets and various versions of our approach.

for fair image classification since the consistency distance in image data at a pixel level is affected by spurious features, such as the background. We highlight our results in Table 1.

For multi-objective optimization, an outcome is considered Pareto dominant if both utility and fairness are improved [5]. GAIA is Pareto dominant over FairMixup and FairBatch for all compared fairness metrics except for AOD. FairBatch is Pareto dominant in the same metrics over FairMixup. While FairMixup is not Pareto dominant for AOD since it has lower accuracy, it shows superior AOD performance. We hypothesize this is due to its predictions being closer to random chance since random predictions are considered fair under AOD.

FairBatch uses meta-optimization of the batch selection process to make the underlying model training to be fair. GAIA uses a similar idea for batch selection using Label Shift (LS), and Attribute Label Shift (ALS). However, while our approach explicitly intervenes in the label distribution and the attribute-label correlation, FairBatch uses an outer loss that attempts to train the model in batch selection. In addition, GAIA is capable of premonition regarding uncertainty, allowing it to make informed predictions that lead to a better trade-off. In contrast, Fair-Mixup uses data augmentation. The counterfactuals generated by data augmentation through interpolation may not reflect reality. However, when the batch selection process is changed in FairBatch and GAIA, each sample comes from the training data. Thus, while the data distribution changes, each sample reflects a real sample. This explains the superior performance of both FairBatch and GAIA over FairMixup.

4.3 Relation among Aleatoric Uncertainty, Utility, and Fairness

To test our central hypothesis that samples with high aleatoric uncertainty contribute more to algorithmic unfairness and prediction errors, we conduct additional experiments for tabular data to examine how GAIA performs in terms of utility and fairness when removing samples with high aleatoric uncertainty (RQ2). Fig. 3 shows results for pruning samples with high aleatoric uncertainty. For both Adult and German datasets, we observe improved accuracy and EOD as we filter out the most uncertain predictions. Group fairness metrics, such as EOD and AOD, consider the difference between the TPR and FPR. When the predictions completely match the ground truth, these metrics approach 1 and 0, respectively, for

all instances of the protected group. The result is an improvement in both accuracy and fairness. Thus, if we consider the samples with the most confident predictions, the likelihood of improving both utility and fairness increases. This serves as sound empirical evidence in favor of our main hypothesis which targets the dichotomy between samples based on aleatoric uncertainty for shifting focus between fairness and utility.

4.4 Individual Fairness

Our fairness notion is inspired by group fairness metrics since it optimizes over the cross-entropy difference for separate instances of the protected attribute. This raises concern over its applicability for individual fairness (RQ3). However, empirical evidence from both the tabular data (Fig. 2, right) and image data (Table 1) shows that GAIA also performs well on the trade-off when individual fairness metrics are of particular interest. To understand these results, we again consider the dichotomy between regions of high and low aleatoric uncertainty and the two individual fairness metrics we used (GE and CS).

The Generalized Entropy Error (GE) is a metric that quantifies the entropy index within each group. When we have low aleatoric uncertainty within a single group, the predictor tends to closely match the ground truth for each sample. This is because higher confidence increases the likelihood of a prediction aligning with the actual label. On the other hand, when the aleatoric uncertainty is high, GAIA aims to optimize for equal cross-entropy between groups, which contributes to improved fairness. However, it is important to note that in scenarios where aleatoric uncertainty is high, the labels themselves are inherently noisy. Consequently, the predictive output for each sample tends to be closer to a random assignment. Thus, at an individual level, samples are treated equally. The Consistency Score (CS) is a metric that evaluates how a classifier treats its k nearest neighbors. In essence, it quantifies the impact of high aleatoric uncertainty, which signifies increased variability among the labels of neighboring samples. As this noise is considered theoretically irreducible, our hypothesis is that leveraging aleatoric uncertainty can effectively identify areas where consistency can be enhanced. This approach offers insights into the improved empirical performance observed in relation to this metric.

4.5 Summary

Since prior works focused on epistemic uncertainty, we study the connection between aleatoric uncertainty and fairness. We show how our approach compares against both group and individual fairness. The results complement the findings by Binns [6], who suggest that group and individual fairness may not always be conflicting objectives. Our experiments also suggest that ALS introduces an improvement over LS. In addition, we observe that GAIA outperforms BNN consistently in terms of utility, while the BNN has a minuscule advantage in terms of fairness. BNN has a coherent representation due to the regularization effect of the variational inference on the encoding space, where the encoder must output a probabilistic distribution over the latent variables that approximates the true posterior. This encourages similar samples to have similar encodings, leading to a more organized and smoother latent space representation. Therefore, it is not surprising that BNNs

demonstrate high performance on individual fairness metrics, as they evaluate the consistency in the treatment of similar covariates. Both GAIA and BNN outperform baseline approaches consistently in terms of the fairness-utility trade-off. Results over both image and tabular datasets show the generalizability of GAIA. Different architectures can be plugged in and sampling from a distribution over the model weights can be used to measure uncertainty.

5 RELATED WORK

Current work on fairness ML relies on identifying and mitigating spurious correlations or reducing epistemic uncertainty. We highlight the novelty of our approach in comparison.

5.1 Bias Mitigation and Fairness

There are three main types of methods for reducing bias in machine learning, which depend on where in the model training process they are applied: (i) pre-processing, (ii) in-processing, and (iii) post-processing. In addition, there are various metrics for evaluating fairness that can be grouped into group fairness or individual fairness metrics. Preprocessing methods [27, 53, 21, 9, 12] aim to reduce bias by modifying the data, labels, or sample importance in the dataset. For example, the Disparate Impact Remover [21] technique attempts to adjust the label distribution to ensure that protected attributes have the same median outcome. The Learning Fair Representations (LFR) [53] approach creates a latent representation of the data to obscure protected attributes. In-processing methods [3, 54, 13] rely on the model architecture to achieve fairness. Adversarial Debiasing [54] involves an adversary that tries to predict the protected attribute. The goal is to make the best predictions in a way that prevents the adversary from distinguishing the protected attribute. Post-processing methods [28, 23, 41] adjust the predictions of a trained model after inference to make them unbiased. There are various approaches with different debiasing objectives. Some methods target specific fairness metrics, such as Calibrated Equal Odds Difference [41], which aims to minimize Equalized Odds.

5.2 Uncertainty based Learning

Deep Learning has achieved unprecedented success in making accurate predictions in various domains; therefore, it is increasingly important to evaluate the reliability and uncertainty of AI systems before deployment. The principles of uncertainty play an important role in AI settings such as concrete learning algorithms [38] and active learning [40]. There are two main types of uncertainty, i.e., aleatoric (or data) uncertainty and epistemic (or model) uncertainty [26]. Common techniques used in uncertainty quantification include Bayesian [39, 51] and Ensemble [55, 32] methods. The highlights come in the form of popular variational inference approaches such as Variational Auto-Encoders (VAE) [30]. The specialty of VAE comes from the estimation of a distribution in the latent space rather than a specific latent representation. Similarly, Bayesian Neural Networks (BNNs) use a distribution over the weights, rather than specific weights to estimate the uncertainty for predictions. One common variation of BNNs is Bayes by Backprop [7] which leverages the standard backpropagation used in traditional NNs. Despite the popularity of uncertainty quantification, approaches using uncertainty to improve fairness are scarce. ROC [28] is one

such instance. Liu et al. [34] use a multi-task model for predicting the under-represented class label in addition to the classification label to create a robust representation space. Singh et al. [46] propose an approach for fair ranking where the probability of being assigned a higher rank is in proportion to the estimated merit.

Our approach complements past work by incorporating aleatoric uncertainty in particular. While prior works suggest good heuristics and processing techniques to overcome the challenge of lack of data, our approach suggests that when the model is likely to make the correct prediction, it is also likely to be fair. Conversely, when the model is unlikely to make the correct prediction due to data ambiguity, we optimize it to ensure fairness. Past approaches can easily be incorporated into our proposed framework by substituting them with the utility objective.

6 CONCLUSION AND FUTURE WORK

This study introduces a novel concept balancing fairness and utility via aleatoric uncertainty. By optimizing objectives based on uncertainty levels, our approach improves fairness and utility trade-off. Aleatoric uncertainty informs model decisions for better trade-off. To mitigate the confounding effects associated with protected attributes, we propose a distributional intervention approach when estimating uncertainty using BNN. We then optimize for fairness in the solution space with high aleatoric uncertainty, and utility elsewhere. The proposed GAIA approach yields an improved fairness-utility trade-off regarding both group and individual fairness. A thorough evaluation of our approach is conducted using multiple datasets across various domains, various metrics, and comparisons to established baseline methods. The theoretical analyses and empirical evidence provide insights into the advantages, limitations, and areas for further improvement in our concept.

Our work significantly contributes to the field of ML by offering a new solution to the balance between fairness and utility. The study highlights the potential link between fairness and predictive uncertainty, and future research will delve into the robustness, scalability, and potential applications of this concept in other domains.

While our approach demonstrates promising results, we acknowledge a few limitations. GAIA relies on the differences in uncertainty between training samples. If the majority of samples consistently exhibit low uncertainty, it suggests both high utility and fairness, even with simple approaches that do not specifically focus on fairness, such as Empirical Risk Minimization [50]. However, if most samples consistently exhibit high uncertainty, our training objective leans toward maintaining fairness rather than utility.

Altering the uncertainty quantification backbone architecture, such as using an auto-encoder, could provide additional insights, and our design allows for such modifications. We separate the downstream model from the uncertainty model, enabling easy integration of GAIA with existing architectures for downstream tasks.

ACKNOWLEDGMENTS

This work received support from the National Science Foundation (NSF) under grant number 2036127, as well as from the Cisco Research Gift Grant (Lu Cheng). The opinions, interpretations, conclusions, and recommendations presented herein solely reflect those of the authors.

REFERENCES

- [1] Moloud Abdar et al. 2021. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Information Fusion*, 76, 243–297.
- [2] Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.
- [3] 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.
- [4] Rachel K. E. Bellamy et al. 2018. AI Fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. (Oct. 2018). <https://arxiv.org/abs/1810.01943>.
- [5] B Douglas Bernheim, Bezalel Peleg, and Michael D Whinston. 1987. Coalition-proof nash equilibria i. concepts. *Journal of economic theory*, 42, 1, 1–12.
- [6] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 514–524.
- [7] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International conference on machine learning*. PMLR, 1613–1622.
- [8] [SW] James Bradbury et al., JAX: composable transformations of Python+NumPy programs version 0.3.13, 2018. URL: <http://github.com/google/jax>.
- [9] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30.
- [10] Kaidi Cao, Yining Chen, Junwei Lu, Nikos Arechiga, Adrien Gaidon, and Tengyu Ma. 2020. Heteroskedastic and imbalanced deep learning with adaptive regularization. *arXiv preprint arXiv:2006.15766*.
- [11] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with fairness constraints: a meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, 319–328.
- [12] Lu Cheng, Nayoung Kim, and Huan Liu. 2022. Debiasing word embeddings with nonlinear geometry. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1286–1298.
- [13] Lu Cheng, Ahmadrza Mosallanezhad, Yasin N Silva, Deborah L Hall, and Huan Liu. 2021. Mitigating bias in session-based cyberbullying detection: a non-compromising approach. In *ACL-IJCNLP*. Vol. 1.
- [14] Lu Cheng, Kush R Varshney, and Huan Liu. 2021. Socially responsible ai algorithms: issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, 71, 1137–1181.
- [15] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big data*, 5, 2, 153–163.
- [16] Ching-Yao Chuang and Youssef Mroueh. [n. d.] Fair mixup: fairness via interpolation. In *International Conference on Learning Representations*.
- [17] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31.
- [18] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. 2020. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*. PMLR, 2803–2813.
- [19] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.
- [20] Sharad Vikram et al. 2022. Oryx. (2022). <https://github.com/jax-ml/oryx>.
- [21] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.
- [22] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2, 11, 665–673.
- [23] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- [25] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. (1994).
- [26] Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110, 457–506.
- [27] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33, 1, 1–33.
- [28] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining*. IEEE, 924–929.
- [29] Diederik P Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [30] Diederik P Kingma, Max Welling, et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12, 4, 307–392.
- [31] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33, 728–740.
- [32] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.
- [33] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open*.
- [34] Jeremiah Zhe Liu, Krishnamurthy Dj Divjiotham, Jihyeon Lee, Quan Yuan, Balaji Lakshminarayanan, and Deepak Ramachandran. [n. d.] Pushing the accuracy-group robustness frontier with introspective self-play. In *The Eleventh International Conference on Learning Representations*.
- [35] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*. (Dec. 2015).
- [36] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml* number 1. Vol. 30. Atlanta, Georgia, USA, 3.
- [37] 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.
- [38] Tom M Mitchell. 1980. *The need for biases in learning generalizations*. CiteSeer.
- [39] Radford M Neal. 2012. *Bayesian learning for neural networks*. Vol. 118. Springer Science & Business Media.
- [40] Vu-Linh Nguyen, Sébastien Destercke, and Eyke Hüllermeier. 2019. Epistemic uncertainty sampling. In *Discovery Science: 22nd International Conference, DS 2019, Split, Croatia, October 28–30, 2019, Proceedings 22*. Springer, 72–86.
- [41] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems*, 30.
- [42] Aleksandr Podkopaev and Aaditya Ramdas. 2021. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*. PMLR, 844–853.
- [43] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. 2021. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 11. Vol. 35, 9419–9427.
- [44] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. [n. d.] Fairbatch: batch selection for model fairness. In *International Conference on Learning Representations*.
- [45] Jonathan Scarlett and Volkan Cevher. 2019. An introductory guide to fano’s inequality with applications in statistical estimation. *arXiv preprint arXiv:1901.00555*.
- [46] Ashudeep Singh, David Kempe, and Thorsten Joachims. 2021. Fairness in ranking under uncertainty. *Advances in Neural Information Processing Systems*, 34, 11896–11908.
- [47] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2239–2248.
- [48] Anique Tahir, Lu Cheng, Ruocheng Guo, and Huan Liu. 2022. Distributional shift adaptation using domain-specific features. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 5593–5597.
- [49] UCI. 1996. Adult. UCI Machine Learning Repository. (1996).
- [50] Vladimir Vapnik. 1991. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4.
- [51] Kuan-Chieh Wang, Paul Vicol, James Lucas, Li Gu, Roger Grosse, and Richard Zemel. 2018. Adversarial distillation of bayesian neural network posteriors. In *International conference on machine learning*. PMLR, 5190–5199.
- [52] Xitong Yang. 2017. Understanding the variational lower bound. *variational lower bound, ELBO, hard attention*, 22, 1–4.
- [53] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.
- [54] 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.
- [55] Jize Zhang, Bhavya Kaikhura, and T Yong-Jin Han. 2020. Mix-n-match: ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*. PMLR, 11117–11128.