Who's the (Multi-)Fairest of Them ALL: Rethinking Interpolation-Based Data Augmentation Through the Lens of Multicalibration

Karina Halevy,[1 2] **Karly Hou,**[2] **Charumathi Badrinath**[2]

[1]Carnegie Mellon University
[2]Harvard University
khalevy@andrew.cmu.edu

Abstract

Data augmentation methods, especially SoTA interpolation-based methods such as Fair Mixup, have been widely shown to increase model fairness. However, this fairness is evaluated on metrics that do not capture model uncertainty and on datasets with only one, relatively large, minority group. As a remedy, multicalibration has been introduced to measure fairness while accommodating uncertainty and accounting for multiple minority groups. However, existing methods of improving multicalibration involve reducing initial training data to create a holdout set for post-processing, which is not ideal when minority training data is already sparse. This paper uses multicalibration to more rigorously examine data augmentation for classification fairness. We stress-test four versions of Fair Mixup on two structured data classification problems with up to 81 marginalized groups, evaluating multicalibration violations and balanced accuracy. We find that on nearly every experiment, Fair Mixup *worsens* baseline performance and fairness, but the simple vanilla Mixup *outperforms* both Fair Mixup and the baseline, especially when calibrating on small groups. *Combining* vanilla Mixup with multicalibration post-processing, which enforces multicalibration through post-processing on a holdout set, further increases fairness.

Code — https://github.com/ENSCMA2/fairest-mixup
Extended version — https://arxiv.org/abs/2412.10575

1 Introduction

Algorithmic fairness has become increasingly important with the ubiquitous application of machine learning (ML). Unfairness can arise from many sources (Huang et al. 2022), including unequal representation of protected groups in data (Guo et al. 2022). For example, people of color can be underrepresented in clinical trials due to access barriers, lack of information, and discrimination (Allison, Patel, and Kaur 2022), leading ML models to have trouble predicting treatment outcomes for non-white patients. One way to mitigate underrepresentation is data augmentation, which creates synthetic individuals from the original data (Chuang and Mroueh 2021; Iosifidis and Ntoutsi 2018; Chawla et al. 2002; Sharma et al. 2020). A particularly promising form of augmentation is Mixup (Zhang et al. 2017) and its fairness-oriented counterpart Fair Mixup (Chuang and Mroueh 2021), which linearly interpolate individuals with features in between majority and minority group attributes. However, existing augmentation literature measures fairness through binary metrics like demographic parity and equalized odds (Chuang and Mroueh 2021), which accumulate loss even when predictors lean toward correct labels. These metrics can be misleading because data often does not include all predictive features, so some notion of uncertainty is appropriate in a good predictor but would be penalized. Furthermore, the methods in Chuang and Mroueh (2021) only assess and optimize fairness for one minority group, but a fair predictor should work well on multiple multi-dimensional intersecting groups.

The metric of multicalibration (MC) (Hebert-Johnson et al. 2018) accounts for this uncertainty and for the presence of multiple groups by comparing predicted probabilities to true probabilities, averaging over groups of interest, and considering subsets of a predictor's support separately. Hebert-Johnson et al. (2018) also introduce an algorithm, with runtime inversely proportional to the size of the smallest group, to post-process a predictor using a holdout set and guarantee a maximum MC violation. Barda et al. (2020) then use this algorithm to learn prediction adjustments from a holdout set and apply those adjustments to test predictions. However, such post-processing subtracts a substantial amount of holdout data from available training data, resulting in even less representation of underrepresented groups in initial training. Moreover, with runtime inversely proportional to group size, enforcing MC for very small groups can be slow. The guarantees of MC enforcement and accuracy tradeoff limits proven in Hebert-Johnson et al. (2018) also only apply to the holdout set, not to unseen test data.

This work examines whether we can combine the desirable properties of MC and data augmentation to supplement the binary outcome insights that demographic parity and equalized odds provide. We ask:

1. RQ1: Under what conditions can Fair Mixup mitigate MC violations of neural network predictors on minority groups while preserving binary classification accuracy?
2. RQ2: When can (Fair) Mixup serve as an alternative to and/or increase the efficiency of MC post-processing?
3. RQ3: What aspects of Fair Mixup contribute to its suc-

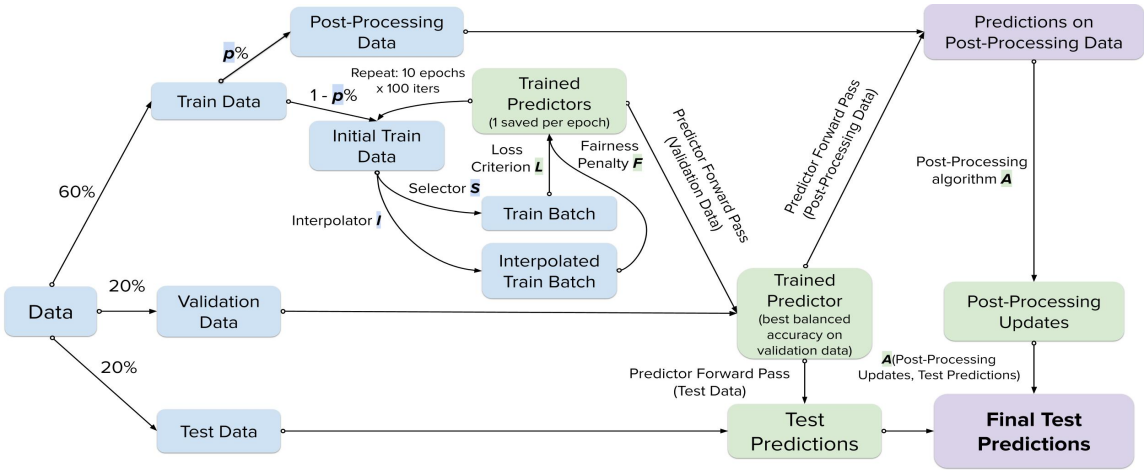

Figure 1: The ML training and evaluation pipelines considered in our work. Each method in our experiments can be characterized by a unique combination of: a percentage p of post-processing data taken from training data, an interpolation-based data augmentation method I, a training batch selection procedure S, a training loss criterion L, a training fairness penalty F, and a post-processing algorithm A. These unique combinations are listed in Table 1.

cess or failure in improving MC-based fairness?

We contribute the first MC-based investigation of several (Fair) Mixup- and MC-inspired neural network training methods (depicted in Figure 1), stress-testing performance and fairness on intersecting demographic groups and creating a new perspective on whether data augmentation is effective. We find that Fair Mixup can only mitigate MC violations and outperform post-processing under its original design of optimizing one group at a time. However, vanilla Mixup consistently makes predictors fairer and results in an average balanced accuracy/MC violation improvement of up to 14.22% when combined with MC post-processing. We also find that the key performance-enhancing component of Fair Mixup is that it learns from interpolated data points. However, its other components (balancing training data by minority group membership and penalizing pairwise unfairness during training) detract from baseline performance, resulting in average balanced accuracy/MC violation decreases of up to 12.29%.

2 Preliminaries

This section defines calibration (Hebert-Johnson et al. 2018; Chouldechova 2017), multicalibration (Hebert-Johnson et al. 2018), multiaccuracy (Hebert-Johnson et al. 2018), and the data augmentation methods we later expand on.

2.1 Notation

Throughout this paper, $\mathcal{X}$ represents a universe of individuals, x_i represents an individual with index i, $S \subseteq \mathcal{X}$ is a subset of individuals, $\mathcal{C} \subseteq 2^{\mathcal{X}}$ is a set of subsets of individuals, f is a predictor that maps individual x_i to outcome probability f_i, p_i^* is the true outcome probability of x_i, and $y_i \in \{0, 1\}$ is the binarized true outcome for x_i.

2.2 Calibration

For a maximum violation $\alpha \in [0, 1]$, f is **α-calibrated** w.r.t. S if $\exists S' \subseteq S$ with $|S'| \geq (1 - \alpha)|S|$ such that $\forall v \in [0, 1]$,

$$|\mathbb{E}_{x_i \sim (S_v \cap S')}[f_i - p_i^*]| \leq \alpha, \quad (1)$$

where $S_v = \{x_i : f_i = v\}$. In most classification tasks, we only see the binary outcome y_i for x_i. Thus, we use a modification called **observable** calibration (Hebert-Johnson et al. 2018), where y_i replaces p_i^* in Eq. 1.

For example, a tumor malignancy classifier is 0.05-observably calibrated for $v = 0.6$ on Latine patients if of all Latine patients for which it predicts a 60% chance of malignancy, 55% to 65% of these patients have a malignant tumor. The classifier is 0.05-observably calibrated on Latine patients if this holds for all v—of all Latine patients for which it predicts a v chance of malignancy, between $v - 5\%$ and $v + 5\%$ of these patients have a truly malignant tumor.

2.3 Multicalibration

f is **$(\mathcal{C}, \alpha)$-multicalibrated** if it is α-calibrated w.r.t. all $S \in \mathcal{C}$ (Hebert-Johnson et al. 2018). We define MC as in Hebert-Johnson et al. (2018), but we require $S = S'$ (calibration on all of S rather than any $1 - \alpha$ of it). For computational feasibility over datasets with millions of prediction probabilities, we also discretize the predicted probabilities. For integer $d > 0$, the **d-discretized** version of S splits S into $d + 1$ subsets, where

$$S_v = \{x_i : \frac{v}{d} \leq f_i < \frac{v + 1}{d}\} \text{ for } v \in [0, 1, ..., d]. \quad (2)$$

Continuing with the tumor malignancy classifier example, the subset of the 10-discretized S = Latine patients with $v = 6$ would be all Latine patients with a predicted chance of at least 60% but less than 70% malignancy. Suppose all patients in this subset have a prediction of 63%.

0.05-calibration would require that 58 to 68% of these patients have a truly malignant tumor, and that the corresponding conditions hold for all other $v \in [0, 10]$. Given $\mathcal{C} = \{\text{Black patients, Asian patients, Latine patients}\}$, $(\mathcal{C}, 0.05)$-multicalibration requires that this 0.05-calibration must hold for Black, Asian, and Latine patients.

2.4 Multiaccuracy

Multiaccuracy (MA) (Hebert-Johnson et al. 2018) is a looser version of MC. f is $(\mathcal{C}, \alpha)$**-multiaccurate** if $\forall S \in \mathcal{C}$,

$$|\mathbb{E}_{x_i \sim S}[f_i - p_i^*]| \leq \alpha. \tag{3}$$

Rather than requiring the expected prediction error within each S *and* predicted probability to be $\leq \alpha$, MA only requires this error to be $\leq \alpha$ in S overall. Thus, for $\mathcal{C} = \{\text{Black patients, Latine patients}\}$, $(\mathcal{C}, 0.05)$-multiaccuracy means that the average prediction for Black patients is within 5% of the true proportion of Black patients that have a malignant tumor, and likewise for Latine patients. In the rest of this paper, when we say f has an MC or MA violation of α on $\mathcal{C}$, we mean that α is the smallest value for which f is $(\mathcal{C}, \alpha)$-multicalibrated or multiaccurate.

2.5 Mixup

Mixup was proposed to improve the generalizability of neural networks (NN) by training on linear combinations of example pairs, with the intuition that the NN would learn how predictions differ as inputs move continuously between feature sets (Zhang et al. 2017). For training batch size b, mixup draws $(x_1, y_1), ..., (x_b, y_b)$ and $(x'_1, y'_1), ..., (x'_b, y'_b)$ without replacement from the training data. Let $t \sim \text{Beta}(\epsilon, \epsilon)$ where $\epsilon \in (0, \infty)$. Mixup constructs one synthetic point per $i \in [1, ..., b]$:

$$(x''_i, y''_i) = (tx_i + (1-t)x'_i, ty_i + (1-t)y'_i) \tag{4}$$

and trains an NN on $(x''_1, y''_1), \cdots, (x''_b, y''_b)$ instead of the original batch. Zhang et al. (2017) showed that mixup decreased test error on CIFAR-10 and CIFAR-100.

2.6 Fair Mixup

Chuang and Mroueh (2021) adapted mixup toward the goal of fairness. **Fair Mixup** (FM) samples $(x_1, y_1), ..., (x_b, y_b)$ from minority group S and $(x'_1, y'_1), ..., (x'_b, y'_b)$ from $S' = \neg S$. Mixup is then performed on these samples as in Section 2.5 to create synthetic points. The loss function applies the standard Binary Cross Entropy (BCE) loss function to the original points, applies the gradient $\mathcal{R}^{\mathcal{M}_S}_{\text{mixup}}$ of a pairwise fairness penalty $\mathcal{M}$ between S and S' to the synthetic points, and adds λ times the fairness penalty to the BCE. Fair Mixup creates better tradeoffs between average precision and the fairness metrics of demographic parity and equalized odds (Chuang and Mroueh 2021).

3 Related Work

3.1 Data Augmentation for Fairness

There are several other data augmentation methods for fairness. In oversampling, minority group samples are duplicated until equal in number to majority group samples (Iosifidis and Ntoutsi 2018). Another method, SMOTE, creates minority group members through linear interpolation among existing minority group members (Chawla et al. 2002). More recently, Sharma et al. (2020) introduce "Ideal World": for each original point, a new sample is created with the same features and label, but the protected attribute is flipped, making both statistical parity difference and average odds difference decrease while preserving accuracy. Outside of structured data, Wadhwa et al. (2022) apply identity pair replacement, identity term blindness, and identity pair swap on text classification. Yucer et al. (2020) introduce data augmentation that improves facial recognition on minority groups.

We focus on structured data classification to minimize the confounding factor of unstructured data featurization. We also choose Fair Mixup as a basis because it minimizes data distribution changes and treats protected attributes as predictive features. Ideal World takes away the predictive information of protected attributes. Oversampling, SMOTE, and Ideal World create additional minority individuals, changing the frequency and composition of minority groups. In contrast, Fair Mixup creates individuals that are neither minority nor majority group members, but rather some interpolated in-between. Thus, while the data distribution may change, the members of boolean circuit-defined groups do not.

3.2 Extensions of MC

Hebert-Johnson et al. (2018) devise algorithms that could enforce MC α's to be below an arbitrary threshold. A related post-processing algorithm, designed for multiaccuracy, is MULTIACCURACY BOOST, which requires a trained auditor on top of a holdout set (Kim, Ghorbani, and Zou 2019). Applying the results of Hebert-Johnson et al. (2018) empirically, Barda et al. (2020) transfer learned post-processing updates to a COVID-19 mortality rate forecasting task. We test a similar application in Section 4.4.

A few works extend (multi-)calibration to more nuanced metrics that handle complex notions of uncertainty. Kumar, Sarawagi, and Jain (2018) add calibration optimization to the training loss function, clamping overconfident predictions while minimizing penalties on true confident predictions. Wald et al. (2021) propose multi-domain calibration to evaluate model generalization to out-of-distribution data, suggesting both isotonic regression post-processing and a training regime that includes calibration from Kumar, Sarawagi, and Jain (2018). Jung et al. (2021) extend MC to higher moments, measuring moment consistency in a way that computes groupwise error inversely proportionally to group size (Jung et al. 2021). Other work extends MC to conformal prediction, which generates prediction sets rather than point estimates (Jung et al. 2023; Foygel Barber et al. 2020). This framework generalizes MC to quantiles of the label's support rather than individual values and is useful for categorical or continuous labels, unlike binary labels, for which MC is already a probabilistic extension. Gopalan et al. (2024) connect MC to multi-group loss minimization.

The most comprehensive investigation of MC post-processing to our knowledge is Hansen et al. (2024), which finds that baseline predictors on tabular data are often decently multicalibrated already, and post-processing does not improve worst-group calibration error for multi-layer per-

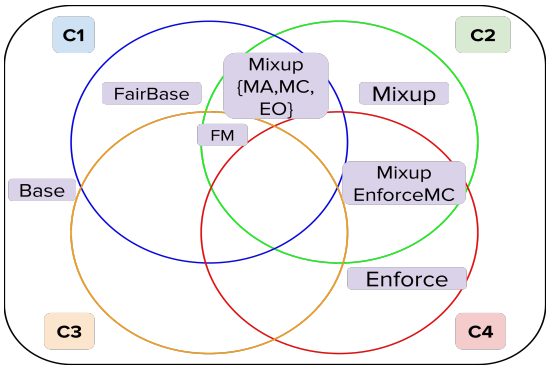

Figure 2: Venn Diagram of each method's core components.

ceptrons, Random Forests, and Logistic Regression but does benefit Support Vector Machines, Decision Trees, and Naive Bayes. When worst-group calibration error improves, there is an overall accuracy tradeoff. They further find that MC enforcement is hyperparameter-sensitive and most effective with huge amounts of data (found in image and language data but not tabular data). They find that calibration algorithms like Platt scaling and isotonic regression sometimes perform nearly on par with MC enforcement while being more efficient. These findings are consistent with previous works suggesting that empirical risk minimization may inevitably yield multicalibrated baseline predictors (Błasiok et al. 2023, 2024). We refer the reader to Hansen et al. (2024) for a more comprehensive MC literature review.

This work extends Hansen et al. (2024) in 3 ways. First, expanding on their max of 15 groups (all $\geq 0.5\%$ of their corresponding population), we stress-test our methods on MC w.r.t. up to 81 groups at a time, up to 55 of which are $\leq 0.25\%$ of their corresponding population. We also select these groups in five different ways to investigate effects of group set size on MC. Second, expanding upon their examination of income prediction from `folktables` on Californian residents from 2018, we evaluate our methods on each permutation of the 10 most populous US states and the four most recent American Community Survey data collection years, yielding 40 datasets. We additionally test employment status prediction on these 40 datasets, for a total of 80 tasks. Third, while their work and much of the current MC literature considers data-reductive post-processing, our work takes inspiration from their finding that post-processing works best on huge datasets and instead focuses on data augmentation to maximize initial training data.

4 Methods

We test 13 training methods to determine the features that affect FM's performance and fairness. FM has 3 key features:

1. C1: Training batches are balanced across membership in the minority group for which we wish to ensure fairness.
2. C2: Synthetic data is created by linearly interpolating original points. If C1 is implemented, each synthetic data point is the interpolation of a minority group member and a majority group member. If not, the original points are split in half and paired at random for interpolation.
3. C3 (can only be done if C2 is also implemented): A fairness penalty is added to the loss function for predictions on synthetic points, minimizing a weighted sum of the standard loss and the fairness penalty during training.

Post-processing is distinguished by the following:

4. C4: A post-processing algorithm learns prediction update rules from post-processing data (subtracted from initial training data), and it applies those update rules to the validation and test data during evaluation and deployment.

With these insights (summarized in Fig. 2), this section describes each method mathematically. We motivate each method by explaining how it implements a subset of {C1, C2, C3, C4}, thus isolating the effects of specific components of FM to answer RQ3. C4 also helps answer RQ2 (FM vs. post-processing). Method names are starred if they contain substantial novel elements that we introduce on top of existing work.

4.1 Baselines

BASE trains an NN with mini-batch gradient descent with Binary Cross Entropy loss. We report test-time balanced accuracy for the epoch with the best validation-time balanced accuracy. BASE does not implement C1, C2, C3, or C4.

***FAIRBASE** modifies BASE by balancing training data groupwise (C1). Suppose we have minority groups $\mathcal{C}$ to optimize for fairness and n iterations of gradient descent per training epoch in BASE. Then, FAIRBASE conducts $n \cdot |\mathcal{C}|$ iterations of gradient descent. Each iteration centers around one $S \in \mathcal{C}$: we construct a batch by selecting one sub-batch from S and one sub-batch from its complement $\neg S$. We subsample the larger sub-batch to be equal in size to the smaller sub-batch to ensure balance across membership in S.

4.2 Variants of Mixup

MIXUP is as defined in Section 2.5, implementing C2.

***MIXUP$_{\text{EO}}$** modifies FAIRBASE. Consider minority groups $\mathcal{C}$ and $n \cdot |\mathcal{C}|$ iterations of gradient descent as in FAIRBASE. MIXUP$_{\text{EO}}$ conducts $2n \cdot |\mathcal{C}|$ iterations of gradient descent, each centered around one pair $(S, y) \in (\mathcal{C}, \{0, 1\})$. We construct a batch by selecting one sub-batch of members of S_y (members of S whose true label is y) and one sub-batch of members of S'_y (members of $\neg S$ whose true label is y). Next, we perform mixup by pairing each member of S_y with a member of S'_y within the batch and interpolating each pair. Our loss is a weighted sum of Binary Cross Entropy applied to the original batch and the same loss applied to the interpolated points. MIXUP$_{\text{EO}}$ implements C1, C2, and a control version of C3 (standard loss instead of pairwise fairness, but number of groups under consideration for this loss is adjustable, as elaborated on in Section 4.3). Thus, we can compare it to FAIRBASE to isolate the effect of C2.

***MIXUP$_{\text{MA}}$** creates one balanced batch per $S \in \mathcal{C}$, as in FAIRBASE, yielding $n \cdot |\mathcal{C}|$ gradient descent iterations. We interpolate each batch by pairing members of S with members of $\neg S$ and adding λ times the Binary Cross Entropy on

Method	p	I	S	L	F	A
BASE	0	$I(\cdot) = []$	uniform random	BCE	$F(\cdot) = 0$	$A(\cdot) = []$
FAIRBASE	0	$I(\cdot) = []$	balance by group	BCE	$F(\cdot) = 0$	$A(\cdot) = []$
MIXUP	0	Mixup	uniform random	$L(\cdot) = 0$	BCE	$A(\cdot) = []$
MIXUP_{EO}	0	Mixup	balance by group $\times y_i$	BCE	$\lambda \cdot$ BCE	$A(\cdot) = []$
MIXUP_{MA}	0	Mixup	balance by group	BCE	$\lambda \cdot$ BCE	$A(\cdot) = []$
MIXUP_{MC}	0	Mixup	balance by group $\times f_i$	BCE	$\lambda \cdot$ BCE	$A(\cdot) = []$
FM_{DP}	0	Mixup	balance by group	BCE	$\lambda \cdot \mathcal{R}^{\text{DP}}_{\text{mixup}}$	$A(\cdot) = []$
FM_{EO}	0	Mixup	balance by group $\times y_i$	BCE	$\lambda \cdot \mathcal{R}^{\text{EO}}_{\text{mixup}}$	$A(\cdot) = []$
FM_{MA}	0	Mixup	balance by group	BCE	$\lambda \cdot \mathcal{R}^{\text{MA}}_{\text{mixup}}$	$A(\cdot) = []$
FM_{MC}	0	Mixup	balance by group $\times f_i$	BCE	$\lambda \cdot \mathcal{R}^{\text{MC}}_{\text{mixup}}$	$A(\cdot) = []$
$\text{ENFORCE}_{\text{MA}}$	25	$I(\cdot) = []$	uniform random	BCE	$F(\cdot) = 0$	Hebert-Johnson et al. (2018) Alg. 3.1
$\text{ENFORCE}_{\text{MC}}$	25	$I(\cdot) = []$	uniform random	BCE	$F(\cdot) = 0$	Hebert-Johnson et al. (2018) Alg. 3.2
$\text{MIXUP}_{\text{ENFORCE}_{\text{MC}}}$	25	Mixup	uniform random	$L(\cdot) = 0$	BCE	Hebert-Johnson et al. (2018) Alg. 3.2

Table 1: Post-processing data split percentages p, data augmentors I, training batch selectors S, loss criteria L (applied to original data), fairness penalties F (applied to synthetic data), and post-processing algorithms A that uniquely characterize each method described in Section 4 and diagrammed in Fig. 1. BCE stands for Binary Cross Entropy loss.

the interpolated points. MIXUP_{MA} also implements C1, C2, and a control version of C3, though C1 is slightly different than in MIXUP_{EO}, allowing us to compare variations of C1.

*$\textbf{MIXUP}_{\textbf{MC}}$ creates $d+1$ batches $\forall S$ by creating d-discretized intervals of f_i's, yielding $(d+1)n|\mathcal{C}|$ gradient descent iterations. For each S_v^d (members of S with predicted probability in $[\frac{v}{d}, \frac{v+1}{d})$), we create a batch with half its points from members of S_v and the other half from members of $(\neg S)_v^d$. We interpolate and calculate loss as in MIXUP_{EO}. MIXUP_{MC} also uses C1, C2, and a control version of C3, providing another way to compare specifics of C1.

4.3 Variants of Fair Mixup

FM implements C1, C2, and C3. Though Chuang and Mroueh (2021) introduce two versions of FM (with $\mathcal{M}$ as demographic parity difference and equalized odds difference), their framework generalizes to any pairwise metric. We first show how to modify FM to accommodate multiple groups simultaneously and then define the 2 versions of FM from Chuang and Mroueh (2021), followed by 2 versions with new metrics that try to include a notion of MC in the penalty. These methods test the effects of C3.

Modifying (Fair) Mixup for Multiple Groups Consider metric $\mathcal{M}$ and group gradients $\mathcal{R}^{\mathcal{M}_{S_1}}_{\text{mixup}}, ..., \mathcal{R}^{\mathcal{M}_{S_{|\mathcal{C}|}}}_{\text{mixup}}$. For $k \in \{1, ..., |\mathcal{C}|\}$, the penalty is the mean of the k highest group gradients (preliminary experiments show that sums produce higher MC αs). We make k adjustable to prevent overfitting. The rest of the computation proceeds as in Chuang and Mroueh (2021). Given $\mathcal{M}_S$, we first transform it into an integral so it can be computed for interpolated data (Chuang and Mroueh 2021). We differentiate the integral to get $\mathcal{R}^{\mathcal{M}_S}_{\text{mixup}}$.

FM_{DP} is the first version of FM in Chuang and Mroueh (2021). $\mathcal{M}_S$ is demographic parity, the difference between the average f_i on members of S vs. its complement S':

$$\mathcal{M}_S = \Delta\text{DP}_S(f) = |\mathbb{E}_{x_i \sim S}[f_i] - \mathbb{E}_{x_i \sim S'}[f_i]|. \quad (5)$$

FM_{EO} is the second version of FM, with the equalized odds difference (Hardt, Price, and Srebro 2016) that modifies DP by considering true outcomes separately:

$$\mathcal{M}_S = \Delta\text{EO}_S(f) = \sum_{y \in \{0,1\}} |\mathbb{E}_{x_i \sim S_y}[f_i] - \mathbb{E}_{x_i \sim S'_y}[f_i]|, \quad (6)$$

where $S_y = \{x_i \in S : y_i = y\}$, and $S' = \neg S$.

*$\textbf{FM}_{\textbf{MA}}$ is our first extension of FM, with a version of MA modified to be pairwise. We measure the mean difference in prediction errors $e_i = f_i - p_i^*$ between S and S':

$$\mathcal{M}_S = \Delta\text{MA}_S(f) = |\mathbb{E}_{x_i \sim S}[e_i] - \mathbb{E}_{x_i \sim S'}[e_i]|, \quad (7)$$

*$\textbf{FM}_{\textbf{MC}}$ is our second extension of FM, with a pairwise modification of MC, which modifies MA by considering one interval $S_v^d = \{x_i \in S : f_i \in [\frac{v}{d}, \frac{v+1}{d})\}$ at a time:

$$\mathcal{M}_S = \Delta\text{MC}_S(f) = \sum_{v=0}^{d} |\mathbb{E}_{x_i \sim S_v^d}[e_i] - \mathbb{E}_{x_i \sim S_v'^d}[e_i]|. \quad (8)$$

4.4 Post-Processing

We test whether MC and MA enforcement (implementing C4) improve test performance as in Barda et al. (2020).

$\text{ENFORCE}_{\text{MA}}$ post-processes predictions to minimize MA violations. We feed (1) predictions on a holdout post-processing set and (2) a set of minority groups $\mathcal{C}$ as inputs to Algorithm 3.1 in Hebert-Johnson et al. (2018). However, we augment Algorithm 3.1 with a list of rules mapping each $S \in \mathcal{C}$ to a float a_S to be added to predictions on members of S. In other words, the algorithm learns how much to adjust predictions for each group. At validation and test time, we add a_S to initial predictor outputs for members of S.

$\text{ENFORCE}_{\text{MC}}$ post-processes predictions to minimize MC violations. It proceeds as in $\text{ENFORCE}_{\text{MA}}$, but we add an integer d as a third input to Algorithm 3.2 in Hebert-Johnson et al. (2018). We augment Algorithm 3.2 with a list of rules mapping each group S_v (members of $S \in \mathcal{C}$ where $f_i \in [\frac{v}{d}, \frac{v+1}{d})$), to a float $a_{S,v}$ to be added to predictions on members of S whose initial predictions are in $[\frac{v}{d}, \frac{v+1}{d})$.

MIXUP$_{\text{ENFORCE}_{\text{MC}}}$ performs MIXUP on a reduced pre-training set followed by ENFORCE$_{\text{MC}}$ on a holdout post-processing set. This tests C2 ∪ C4, as we ultimately find that MIXUP performs best overall among methods that do not use C4. We implement MIXUP$_{\text{ENFORCE}_{\text{MC}}}$ to see if data augmentation improves the performance of ENFORCE$_{\text{MC}}$ (RQ2).

5 Experiments

This section describes our data and experimental settings.

5.1 Datasets

We test 2 tasks from `folktables` (Ding et al. 2021), a superset of Adult Income data (Becker and Kohavi 1996) from the American Community Survey. We choose these tasks based on experiments in Jung et al. (2023). We also seek problems with reasonable baseline balanced accuracy ($\geq 80\%$) to focus on making useful classifiers fairer. We have $p_i^* \in \{0, 1\}$, but $f_i \in [0, 1]$. Table 2 summarizes the data. Full data statistics are at http://tiny.cc/mfm-stats.

EMPLOYMENT The task is to predict whether an individual is employed. Our preprint specifies exact input features.

INCOME The task is to predict whether an individual's annual income is higher than the median income for that year in their state according to Data Commons (Google 2024).

We run 40 geographically and temporally varied datasets each for EMPLOYMENT and INCOME: the 10 most populous US states × 4 most recent years.

5.2 Experimental Settings

To measure effects of $|\mathcal{C}|$ and $|S|$, we run 5 settings:

ALL $\mathcal{C} = \cup\{$all n computationally possible racial groups, disabled people, disabled members of each racial group$\}$. A racial group is computationally possible if for all random seeds, at least one disabled member of that group is in each of the train, validation, and test splits. $|\mathcal{C}| = \mathbf{2n + 1}$.

BIG $\mathcal{C} = \cup\{b$ racial groups each comprising $> 0.25\%$ of the total dataset, disabled people, disabled members of each of the b racial groups.$\}$ $|\mathcal{C}| = \mathbf{2b + 1}, \mathbf{b << n}$.

SMALL $\mathcal{C} = \cup\{s$ racial groups each comprising $\leq 0.25\%$ of the total dataset, disabled people, disabled members of each of the s racial groups$\}$. $|\mathcal{C}| = \mathbf{2s + 1}, \mathbf{s << n}$.

DIS This setting is closest to what FM has already been tested on: $\mathcal{C} = \{$disabled individuals$\}$, so $|\mathcal{C}| = \mathbf{1}$.

DLFR $\mathcal{C}$ = disabled people, members of the least frequent (computationally possible) racial group (LFR), and disabled members of the LFR, hence $|\mathcal{C}| = \mathbf{3}$.

6 Results

To capture both fairness and overall performance, we compute the mean across all 40 (state, year) pairs of the following quantities for each experiment: (1) % increase in balanced accuracy over BASE for the corresponding state, year, and task and (2) % decrease over BASE in worst (highest) individual group MC violation α. Table 3 reports these mean percentages, showing that for all (task, setting) pairs except for DIS (both tasks), (EMPLOYMENT, DLFR), and (INCOME, SMALL), MIXUP$_{\text{ENFORCE}_{\text{MC}}}$ shows the biggest average balanced accuracy and MC α improvement. The other best methods are MIXUP$_{\text{MA}}$ for (EMPLOYMENT, DIS), MIXUP for (INCOME, DIS) and (INCOME, SMALL), ENFORCE$_{\text{MC}}$ for (EMPLOYMENT, DLFR) (though MIXUP$_{\text{ENFORCE}_{\text{MC}}}$ is close), but FM$_{\text{DP}}$ has the best α for (EMPLOYMENT, DIS). If we consider only methods that perform post-processing or augmentation/data balancing (i.e. all methods except MIXUP$_{\text{ENFORCE}_{\text{MC}}}$), the best method is ENFORCE$_{\text{MC}}$, except (EMPLOYMENT, DIS) (MIXUP$_{\text{MA}}$ was best), (INCOME, SMALL), and (INCOME, DIS) (MIXUP was best). One note is that for 28 of 40 datasets, we had $s = 0$ and thus $\mathcal{C}$ = just disabled people, so (INCOME, SMALL) results may be more characteristic of single-group calibration. We also note that all methods except for BASE had negative mean increases in balanced accuracy (up to -1.25%), so positive values in Table 3 indicate fairness improvements.

For FM, all variants but (EMPLOYMENT, DIS) worsened fairness. For (EMPLOYMENT, DIS), FM improved fairness and preserved balanced accuracy, confirming per Chuang and Mroueh (2021) that FM works on one larger group.

Comparing MIXUP$_{\text{ENFORCE}_{\text{MC}}}$ and ENFORCE$_{\text{MC}}$, we observe that while MIXUP$_{\text{ENFORCE}_{\text{MC}}}$ outperforms ENFORCE$_{\text{MC}}$ in many cases, it sometimes makes the ENFORCE$_{\text{MC}}$ component of MIXUP$_{\text{ENFORCE}_{\text{MC}}}$ less efficient. On the EMPLOYMENT dataset, the number of iterations to convergence of the ENFORCE$_{\text{MC}}$ post-processing algorithm increased by a percentage in the range $(0.34\%, 5.8\%)$, with the greatest percentage increase for SMALL (+5.8%) and the greatest decrease for BIG (-1.58%). For INCOME, all methods took fewer iterations, in the range $(-3.68\%, -0.08\%)$.

Finally, we analyze correlations between results and data statistics. We largely find either no correlation or low correlations, with some exceptions. One exception is that the mean MC α across groups $> 0.25\%$ of the population on ALL has a moderate correlation with total dataset size (lower violations for bigger datasets) for all non-BASE methods that do not involve ENFORCE$_{\text{MC}}$. This mean α on ALL also moderately correlates with the number of groups bigger than 0.25% of the population for FM$_{\text{MA}}$, and it also moderately correlates with number of groups smaller than 0.25%, total number of minority groups, number of disabled individuals, and number of non-white individuals for several methods. Looking at efficiency, the number of iterations to convergence of ENFORCE$_{\text{MC}}$ and MIXUP$_{\text{ENFORCE}_{\text{MC}}}$ both strongly correlate with dataset size, but there is no correlation with the % change in number of iterations between methods.

7 Discussion

Our results reveal the importance of stress-testing fairness optimization on multiple groups of varying sizes and on metrics that capture uncertainty. To answer RQ1, **the only condition under which FM improves MC is the one it was designed for**: fairness for one minority group (DIS) on a truly binary problem (EMPLOYMENT). This holds irrespective of the train-time fairness penalty. This leads to an an-

Dataset	Size	# Features (Binary, Categorical, Continuous)	# Non-White	# Disabled	Max # Minority Groups	Max # Groups $\leq$ 0.25% of Population	Mean Size of Smallest Group
EMPLOYMENT	6,993,839	5, 9, 2	2,160,161	1,036,251	81	55	28.5
INCOME	3,543,292	2, 3, 6	1,014,632	250,074	51	28	31.85

Table 2: Summary statistics of the EMPLOYMENT and INCOME datasets. "Size" is the number of individuals summed over all 40 subsets. Maxes and means are taken over these subsets. "Smallest Group" disabled members of the LFR.

Method	EMPLOYMENT					INCOME				
	ALL	BIG	SMALL	DIS	DLFR	ALL	BIG	SMALL	DIS	DLFR
BASE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FAIRBASE	-2.69	-2.36	-4.28	2.08	-3.76	-3.04	-4.26	-11.08	-12.29	-6.34
MIXUP	2.89	3.11	2.56	-23.20	2.03	1.22	1.30	**3.12**	**3.20**	1.63
MIXUP_{EO}	-2.54	-2.21	-4.66	4.45	-4.07	-3.39	-4.14	-9.73	-12.09	-6.15
MIXUP_{MA}	-3.31	-3.08	-4.70	**4.50**	-5.35	-3.10	-3.51	-11.51	-10.80	-7.27
MIXUP_{MC}	-2.91	-2.55	-5.30	3.09	-5.93	-3.08	-3.02	-10.34	-10.78	-6.58
FM_{DP}	-3.41	-2.22	-5.23	3.84	-5.29	-3.34	-3.39	-10.75	-10.11	-7.67
FM_{EO}	-2.26	-3.36	-4.05	1.73	-4.52	-3.55	-3.58	-10.00	-11.12	-5.55
FM_{MA}	-2.95	-3.99	-6.35	1.40	-5.79	-3.74	-3.41	-10.33	-11.92	-7.36
FM_{MC}	-2.72	-2.36	-4.94	0.61	-4.12	-4.32	-3.23	-10.41	-11.30	-5.52
$\text{ENFORCE}_{\text{MA}}$	-0.46	0.02	0.46	0.43	-0.32	-0.02	0.70	-1.27	-2.76	-1.66
$\text{ENFORCE}_{\text{MC}}$	8.31	12.03	7.89	-15.28	**8.74**	9.87	9.97	-6.61	-17.82	8.28
$\text{MIXUP}_{\text{ENFORCE}_{\text{MC}}}$	**10.36**	**12.97**	**9.23**	-29.27	8.72	**11.64**	**14.22**	-2.15	-11.37	**13.06**

Table 3: Summary (mean of 10 trials) of methods across 40 (state, year) pairs $\times$ EMPLOYMENT and INCOME. Each number is the mean of the % increase in balanced accuracy and % decrease in worst-group MC α.

swer to RQ2: under a **single-group truly-binary condition**, FM (especially FM_{DP}) outperforms post-processing in ensuring fairness for disabled people. Based on raw αs, this could be because disabled people are a relatively big, non-monolithic group for which the BASE NN is already more calibrated than the more fine-grained racial groups. Thus, to further improve the BASE α, it may help more to examine more disabled individuals and their full feature sets during training (as in FM) rather than apply a fixed adjustment to disabled individuals unconditionally (as in post-processing). However, this fixed post-processing adjustment may work well for smaller racial groups because their smaller sizes make race more informative than disability.

Regular MIXUP presents a robust alternative to $\text{ENFORCE}_{\text{MA}}$ in nearly all settings and to $\text{ENFORCE}_{\text{MC}}$ when considering one group in a continuous-to-binary prediction problem as in (INCOME, DIS). More powerfully, **combining MIXUP and $\text{ENFORCE}_{\text{MC}}$ enhances performance** of $\text{ENFORCE}_{\text{MC}}$ alone in most settings, especially when more than one group is considered. However, it is unclear whether this enhancement entails higher efficiency for the $\text{ENFORCE}_{\text{MC}}$.

For RQ3, **MIXUP is the overall best post-processing-free method**. Comparing MIXUP with MIXUP_{EO}, MIXUP_{MA}, MIXUP_{MC}, and FAIRBASE, we observe that **using interpolated data contributes more to fairness improvements** than groupwise balancing of training batches. Looking at FAIRBASE, MIXUP_{EO}, MIXUP_{MA}, and MIXUP_{MC}, we further suggest that **data balancing may adversely affect performance and fairness**, since the key factor that sets MIXUP apart from worse-performing methods of FM, MIXUP_{EO}, MIXUP_{MA}, and MIXUP_{MA} is C1. This may be because having limited minority instances means we learn less about majority instances as well (and since groups intersect, some instances that are minorities in one way but majorities in another are seen less). Finally, comparing MIXUP_{EO}, MIXUP_{MA}, and MIXUP_{MC} to FM variants, we see that C3 effects (train-time fairness penalty) are inconclusive, as outcomes fluctuate by method and setting. Thinking more generally about why MIXUP outperforms FM so often, we hypothesize that in addition to the adverse effect of data balancing in FM, MIXUP has a more manageable amount of learning (normal BCE loss, with more data to learn from, net positive), while the pairwise fairness component of FM loss may be differently valued across demographic groups, thus possibly leading to less stable/effective learning (added complexity to the loss might also be a negative that worsens with the number of groups).

8 Conclusion

We conduct the first investigation of how data augmentation via interpolation affects MC-based fairness on multiple minority groups of multiple sizes for binary tabular data classification. We find that while Fair Mixup is not so fair on multiple groups, regular mixup mitigates MC violations across many groups, both by itself and with MC post-processing. Our investigation opens several avenues of future work, with our pipeline being extensible to data augmentation on probabilistic fairness in other modalities and ML problems.

Ethical Statement

Augmentation introduces synthetic data and alters demographic representation to present the illusion that certain groups are well-represented. We urge creators and users

of augmented datasets to be transparent about augmentation methods used. We lead by example as we release our datasets with full methodological descriptions. Furthermore, we caution that our implemented augmentation methods can substantially alter outcomes in real-world decision-making settings, and examining multicalibration is meant to be a supplement to, not a replacement for, frameworks addressing binary, individual-level fairness.

Acknowledgements

We thank Professor Cynthia Dwork and Pranay Tankala for teaching the course that inspired this work and for mentoring us through the initial stages of the project. We also thank Professor Maarten Sap and Alfredo Gomez for helpful feedback during the paper writing and rebuttal process.

References

Allison, K.; Patel, D.; and Kaur, R. 2022. Assessing multiple factors affecting minority participation in clinical trials: Development of the clinical trials participation barriers survey. *Cureus*, 14(4): e24424.

Barda, N.; Riesel, D.; Akriv, A.; Levy, J.; Finkel, U.; Yona, G.; Greenfeld, D.; Sheiba, S.; Somer, J.; Bachmat, E.; Rothblum, G. N.; Shalit, U.; Netzer, D.; Balicer, R.; and Dagan, N. 2020. Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nature Communications*, 11(1): 4439.

Becker, B.; and Kohavi, R. 1996. Census Income Data Set.

Błasiok, J.; Gopalan, P.; Hu, L.; Kalai, A. T.; and Nakkiran, P. 2024. Loss Minimization Yields Multicalibration for Large Neural Networks. In Guruswami, V., ed., *15th Innovations in Theoretical Computer Science Conference, ITCS 2024, January 30 to February 2, 2024, Berkeley, CA, USA*, volume 287 of *LIPIcs*, 17:1–17:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Błasiok, J.; Gopalan, P.; Hu, L.; and Nakkiran, P. 2023. When Does Optimizing a Proper Loss Yield Calibration? In *Thirty-seventh Conference on Neural Information Processing Systems*.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2): 153–163. PMID: 28632438.

Chuang, C.-Y.; and Mroueh, Y. 2021. Fair Mixup: Fairness via Interpolation. In *International Conference on Learning Representations*.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Foygel Barber, R.; Candès, E. J.; Ramdas, A.; and Tibshirani, R. J. 2020. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2): 455–482.

Google. 2024. Data Commons.

Gopalan, P.; Okoroafor, P.; Raghavendra, P.; Shetty, A.; and Singhal, M. A. 2024. Omnipredictors for Regression and the Approximate Rank of Convex Functions. In *COLT*.

Guo, L. N.; Lee, M. S.; Kassamali, B.; Mita, C.; and Nambudiri, V. E. 2022. Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-A scoping review. *J. Am. Acad. Dermatol.*, 87(1): 157–159.

Hansen, D.; Devic, S.; Nakkiran, P.; and Sharan, V. 2024. When is Multicalibration Post-Processing Necessary? arXiv:2406.06487.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 3323–3331. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819.

Hebert-Johnson, U.; Kim, M.; Reingold, O.; and Rothblum, G. 2018. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1939–1948. PMLR.

Huang, J.; Galal, G.; Etemadi, M.; and Vaidyanathan, M. 2022. Evaluation and mitigation of racial bias in clinical machine learning models: Scoping review. *JMIR Med. Inform.*, 10(5): e36388.

Iosifidis, V.; and Ntoutsi, E. 2018. Dealing with Bias via Data Augmentation in Supervised Learning Scenarios. *Jo Bates Paul D. Clough Robert Jäschke*, 24.

Jung, C.; Lee, C.; Pai, M.; Roth, A.; and Vohra, R. 2021. Moment Multicalibration for Uncertainty Estimation. In Belkin, M.; and Kpotufe, S., eds., *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, 2634–2678. PMLR.

Jung, C.; Noarov, G.; Ramalingam, R.; and Roth, A. 2023. Batch Multivalid Conformal Prediction. In *International Conference on Learning Representations*.

Kim, M. P.; Ghorbani, A.; and Zou, J. 2019. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 247–254. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.

Kumar, A.; Sarawagi, S.; and Jain, U. 2018. Trainable Calibration Measures for Neural Networks from Kernel Mean Embeddings. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2805–2814. PMLR.

Sharma, S.; Zhang, Y.; Ríos Aliaga, J. M.; Bouneffouf, D.; Muthusamy, V.; and Varshney, K. R. 2020. *Data Augmentation for Discrimination Prevention and Bias Disambiguation*, 358–364. New York, NY, USA: Association for Computing Machinery. ISBN 9781450371100.

Wadhwa, M.; Bhambhani, M.; Jindal, A.; Sawant, U.; and Madhavan, R. 2022. Fairness for Text Classification Tasks with Identity Information Data Augmentation Methods.

Wald, Y.; Feder, A.; Greenfeld, D.; and Shalit, U. 2021. On Calibration and Out-of-Domain Generalization. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Yucer, S.; Akcay, S.; Moubayed, N. A.; and Breckon, T. 2020. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Computer Vision and Pattern Recognition Workshops*. IEEE. To be presented at the Workshop on Fair, Data Efficient and Trusted Computer Vision.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond Empirical Risk Minimization. *CoRR*, abs/1710.09412.