FairContrast: Enhancing Fairness through Contrastive learning and Customized Augmenting Methods on Tabular Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

As AI systems become more embedded in everyday life, the development of fair and unbiased models becomes more critical. Considering the social impact of AI systems is not merely a technical challenge but a moral imperative. As evidenced in numerous research studies, learning fair and robust representations has proven to be a powerful approach to effectively debiasing algorithms and improving fairness while maintaining essential information for prediction tasks. Representation learning frameworks, particularly those that utilize self-supervised and contrastive learning, have demonstrated superior robustness and generalizability across various domains. Despite the growing interest in applying these approaches to tabular data, the issue of fairness in these learned representations remains underexplored. In this study, we introduce a contrastive learning framework specifically designed to address bias and learn fair representations in tabular datasets. By strategically selecting positive pair samples and employing supervised and self-supervised contrastive learning, we significantly reduce bias compared to existing state-of-the-art contrastive learning models for tabular data. Our results demonstrate the efficacy of our approach in mitigating bias with minimum trade-off in accuracy and leveraging the learned fair representations in various downstream tasks.

1 Introduction

The real-world application of deep learning approaches is expanding rapidly. These approaches are susceptible to stereotypes or societal biases inherited in the data, resulting in models biased against individuals with specific sensitive attributes and treating them unfairly. Well-known examples include a recidivism risk prediction model that predicts reoffending rates for individuals of certain races at twice the rate compared to others [1, 9, 33] , or recruitment models showing bias towards male candidates over equally qualified female candidates[29, 11, 10]. Consequently, there is an increasing amount of research focused on algorithmic fairness, with the primary objective being to guarantee that sensitive attributes have no impact on algorithmic outcomes.

Learning fair and robust representations has shown its potential in effectively debiasing and improving fairness while keeping the essential information for the prediction task [44]. In representation learning frameworks, self-supervised learning and particularly contrastive learning have shown superior robustness and generalizability across various domains, including natural language processing (NLP) and computer vision, even in scenarios with fully labeled or few labeled data [4]. Contrastive learning is more challenging when applied to tabular datasets compared to other data types such as images, text, or speech. This is because these datasets typically contain spatial, semantic, and vocal relationships, which provide structured information not present in tabular data. Although contrastive learning for tabular data is receiving increased attention, the fairness in these learned representations has not been

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

thoroughly explored. Our particular interest lies in investigating whether contrastive learning methods can be utilized to mitigate bias and improve the fairness of representations. We argue that other contrastive learning models for tabular data, such as VIME [45] and SCARF [3], which do not address fairness issues, exhibit bias in their predictions in downstream tasks, leading to discrimination.

In this study, we propose a contrastive learning framework for tabular dataset, to mitigate bias and learn fair representations. For positive pairs, each privileged sample with favorable outcome is paired with one randomly selected unprivileged sample with favorable outcome. Other samples are paired with one randomly selected sample with the same class and the same sensitive attribute. In our supervised and unsupervised framework, supervised contrastive learning loss [27] and InfoNCE loss [17, 32] combined with binary cross-entropy loss are used, respectively. Our proposed method is evaluated on various prevalent datasets in the fairness domain. The results demonstrate a significant reduction in bias compared to existing state-of-the-art contrastive learning frameworks for tabular datasets. Moreover, our framework results in learning fair representations that can be utilized in any downstream tasks.

2 Related Work

Self-Supervised Learning for tabular data. As access to unlabeled data expands, self-supervised learning (SSL) has received attention across various domains, including natural language processing (NLP), computer vision, and speech recognition. SSL approaches are representation learning frameworks that use unlabeled data to learn robust and meaningful representations. SSL also demonstrates its ability in robustness and generalizability even in scenarios with fully labeled or few labeled data [4]. SSL methods highly depend on the correlations within the features of the data. Recently, there has been an increasing focus within the representation learning field on employing SSL for tabular data. Unlike other data types such as images, text, or speech, which possess distinct structures such as spatial, semantic, and vocal relationships respectively, tabular data are more challenging [42]. This is primarily due to the absence of explicit relationships for learning representation, which also varies across different tabular datasets[42].

Several studies [2, 35, 3, 45, 7, 18, 23, 40, 43] have been proposed that deploy SSL methods in tabular datasets. These approaches can be categorized to two groups : 1. employing the pretext tasks and 2. contrastive learning. Deploying the pretext task is the most widely-used category. [45] proposes Value Imputation and Mask Estimation(VIME), a self and semi supervised framework for tabular data. In the self supervised framework, they pre-trained an encoder on unlabeled corrupted/masked data and extracted the features. Then these representations are passed through "mask estimation" and "feature estimation" heads for recovering the binary mask used for corruption, and the value of the feature that has been masked. This pre-trained encoder is utilized in the semi-supervised learning framework as well. Their choice of the corruption strategy is to choose a mask randomly sampled from a Bernoulli distribution. For the filling strategy they utilize CutMix [46], which replaces all masked values of each sample with values from another randomly selected sample. VIME achieved state-of-the-art on clinical and genomics datasets. On the other hand, the main concept of contrastive learning is to pull similar instances (positive pairs) together and push dissimilar instances (negative pairs) away. This is achieved by maximizing agreement (or minimizing the distance) within the embedding space of positive pairs while maximizing the distances with negative pairs. [8, 36, 32, 14]. These methods have been very successful in computer vision. In data types such as images, the positive pairs can be generated through a data augmentation module (i.e., through transformation of the image such as cropping and resizing, rotation, color dissertation, etc.), while negative pairs correspond to other images in the batch [8]. More recently the scope of contrastive learning has been extended to weakly supervised [38], semi-supervised, and supervised setup. These studies have introduced an extra conditional variable such as details about the downstream task [37] or labels from downstream tasks [27, 26], to improve the quality of the representations. In the supervised contrastive learning setup, the positive pairs belong to the same class while negative pairs belong to different classes [27]. In the NLP domain, it has been demonstrated that the model's robustness to noise and data sparsity can be improved when supervised contrastive learning is combined with a cross entropy loss [15, 34]. SCARF proposed by [3] is another state-of-the-art approach that deploys contrastive learning on tabular data. In the SSL setting Scarf method masks 60% of the features for each data point and then uses Random Feature Corruption to replace these masked values. Finally they fine-tune their encoder weights with a classification head on top through some labeled data. The authors of this

study claim the contrastive learning setting of SCARF is superior compared to pre-trained models such as VIME. They compared their methods with other random feature corruption methods, such as CutMix [46] and MixUp [48] and argue that their proposed random feature corruption method is more effective. As mentioned earlier, CutMix [46] replaces values from one randomly chosen sample for all of the masked values in each sample, whereas SCARF method randomly selects a sample for each masked value. In addition, the corruption method of MixUp [48] is to replace it with a linear combination of the sample and another randomly chosen sample. It is demonstrated that MixUp is more effective when corruption is done on the embedding space rather than input space [35].

Fair Contrastive learning Contrastive learning has been extensively used to address fairness concerns in the field of computer vision [47, 22, 5, 39]. In contrast, while the application of self-supervised contrastive learning to tabular data is gaining attention, its use for fairness-specific objectives remains comparatively underexplored. [30] introduces a conditional contrastive learning (CCL) approach primarily designed for vision tasks, which they also extend to tabular data. Their method selects positive and negative pairs conditioned on the sensitive attribute to minimize sensitive leakage while improving class separability. However, their augmentation strategy, adding isotropic Gaussian noise to standardized tabular features, originates from vision-based contrastive frameworks and is less well-suited for tabular data. Many tabular features are discrete, semantically structured, or non-continuous, making Gaussian perturbations potentially unrealistic or semantically meaningless. In the domains of NLP and computer vision, [34] proposed a contrastive learning-based method for bias mitigation that encourages representations of samples with the same class label to be close, while pushing apart representations that share the same protected attribute. Although developed for unstructured data, the underlying principle of this approach, decoupling class and sensitive attribute representations, is also applicable to tabular data. [6] employs self-supervised setting on tabular data in an encoder- decoder framework and discusses fairness; however, they do not utilize contrastive learning. DualFair [19] presents a self-supervised representation learning framework that jointly addresses both group fairness and counterfactual fairness. It achieves this by generating counterfactual samples using a cyclic variational autoencoder (C-VAE), applying fairness-aware contrastive loss to align embeddings across sensitive groups and counterfactuals, and using self-knowledge distillation to maintain representation quality.

3 Method

Figure 1 illustrates the general framework of our proposed model. A customized technique is used to selectively pair positive samples from the original input based on specific criteria. These positive pairs are integrated into the training process using a contrastive loss, alongside classification tasks, in an end-to-end manner.

The selection of negative pairs in our contrastive learning framework depends on whether the setting is supervised or self-supervised. In the supervised version, negative pairs within a batch are formed from samples of different classes. In the self-supervised version, all other samples in the batch are treated as negatives.

For positive pairs, instances from the same class are further conditioned on the sensitive attribute. That is, we sample pair instances within the same subgroup (defined by outcome * sensitive attribute), such as Female with low income paired with another Female with low income. The only exception is for the privileged group with a favorable outcome. This design helps preserve subgroup-specific characteristics in the learned representations, ensuring the model remains accurate within subgroups while maintaining clear separation between them.

For the privileged group with favorable outcomes, positive pairs are drawn from the unprivileged group with the same favorable outcome, as the privileged group poses challenges to the classifier's ability to ensure fair predictions. For example, we pair a Male high income instance with a Female high income instance. Keeping them in the same class (high income) ensures that the model learns to minimize representational distance between them. This approach encourages the model not to rely on the sensitive attribute in favorable outcome predictions, aligning with the fairness criterion of equality of opportunity by promoting intra-group similarity across sensitive attributes.

The general idea behind contrastive learning is to train a model to bring similar samples closer together in a learned representation space while pushing dissimilar samples apart. Our strategy embeds fairness into the model's learning process without altering the core contrastive learning

mechanism. By encouraging instances from favorable outcomes with different sensitive attributes to be closer in representation space, we naturally achieve fairness goals without relying on additional fairness-specific constraint-based loss functions. Incorporating our custom sampling strategy and optimizing with contrastive loss encourages the embeddings of selected groups to converge, reducing bias and mitigating discrimination while preserving model utility.

To implement this approach, our architecture includes an encoder $z = Enc(x)$, which maps the input to a representation. These representations are then used to calculate both the contrastive loss and the classification loss. Within this framework, we investigate a range of contrastive loss functions:

- **Self-supervised contrastive loss:** Self-supervised contrastive learning does not require explicit labels for training. Instead, it leverages positive and negative pairs of the data sample. In this context, given a mini-batch with a set of N randomly selected samples, let $i \in \{1...N\}$ be the index of an arbitrary sample, called the *anchor* and let j be the index of random augmentations (a.k.a., "views"), also called the *positive*, the corresponding mini-batch consists of $2N$ pairs where the other $2(N-1)$ indices $\{1...N\} \setminus \{i\}$ are called the *negatives*. Here $z_i = Enc(x_i)$, $\tilde{z}_j = Enc(\tilde{x}_j)$ denotes the embeddings generated from the encoder and the self-supervised contrastive loss is calculated as [8, 36, 21]:

$$L^{self} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k\not\equiv i}\exp(\text{sim}(z_i, z_k)/\tau)} \tag{1}$$

  where **sim** function is the Cosine Similarity (Eq. 2).

$$\text{sim}(z_i, z_j) = \frac{z_i^T \tilde{z}_j}{\|z_i\|_2.\|\tilde{z}_j\|_2} \tag{2}$$

  $\tau \in R^+$ is a scalar temperature parameter controlling softness.

- **Supervised contrastive loss:** In the realm of supervised learning, the contrastive loss outlined in Eq. 1 encounters limitations when multiple samples are known to belong to the same class [27]. Eq. 3 presents the most direct approaches for extending Eq. 1 to include supervision [27]:

$$L^{sup} = -\frac{1}{N}\sum_{k=1}^{N}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(z_i \cdot z_p/\tau)}{\sum_{q\in Q(i)}\exp(z_i \cdot z_k/\tau)} \tag{3}$$

  where the symbol $\cdot$ denotes the inner (dot) product and $Q(i) \equiv \{1...N\} \setminus \{i\}$, $P(i) \equiv \{p \in Q(i) : y_p = y_i\}$ is the set of indices of all positives in the batch distinct from i, and $|P(i)|$ is its cardinality.

As shown in Figure 1, our final objective function is formulated as a weighted combination of a binary cross-entropy loss and contrastive loss.

$$L_{total} = \alpha L_{BCE} + L_{SCL} \tag{4}$$

3.1 Theoretical Analysis

Let $(X, Y, S) \sim p_{\text{data}}$, where $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$ are features, $Y \in \{0, 1\}$ is the target, and $S \in \{0, 1\}$ is a binary sensitive attribute. An encoder $f_\theta : \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^{d_z}$ maps a sample to a representation $Z = f_\theta(X)$. Similarity between two representations is measured by $g_\tau(z, z') = \exp\big(\langle z, z'\rangle/\tau\big)$ with temperature $\tau > 0$.

Positive-pair sampler. Given an anchor sample (x, y, s) with $y = 1$ (favourable label), we draw the positive according to the mixture

$$p_{\text{pos}} = \pi\, p_{\text{cross}} + (1-\pi)\, p_{\text{within}}, \tag{5}$$

where

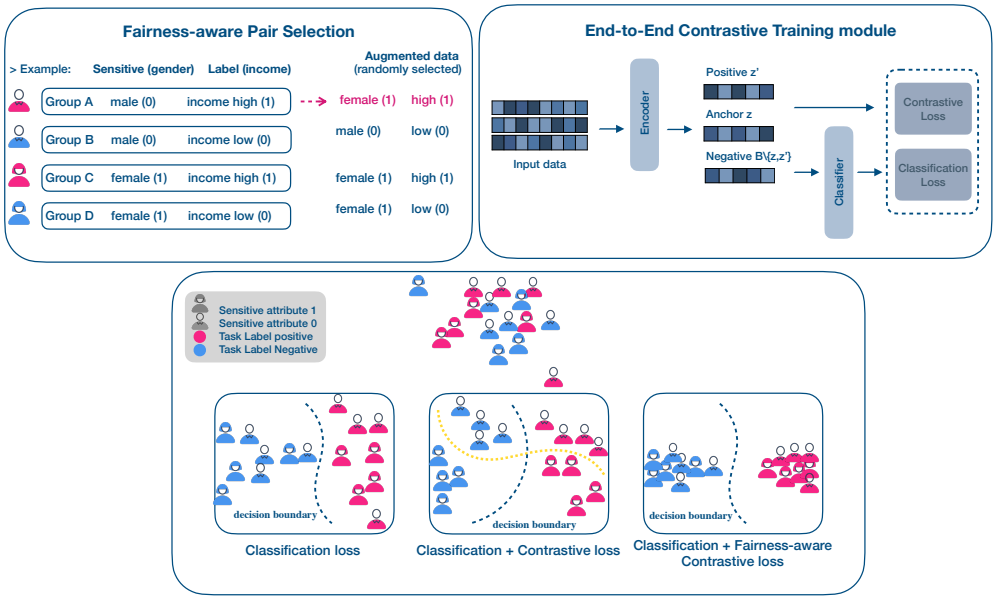


Figure 1: The schematic diagram illustrates the proposed fairness-aware contrastive learning framework. Our approach involves selectively sampling positive pairs based on specific criteria and integrating them into the training process with a contrastive loss in an end-to-end manner. Although combining supervised contrastive learning with cross-entropy loss improves model robustness, contrastive loss without explicit bias mitigation can unintentionally separate instances across sensitive attributes in the representation space. Our proposed fairness-aware contrastive loss, together with cross-entropy, reduces this separation by bringing positive-class instances from different sensitive groups closer, thereby improving fairness without requiring additional fairness-specific constraint loss functions.

- $p_{\text{cross}}\big((x, y, s), (x^+, y^+, s^+)\big)$ places probability mass only on pairs satisfying $y^+ = y = 1$ and $s^+ \neq s$;
- p_{within} places mass on pairs with $y^+ = y = 1$ *and* $s^+ = s$.

Consequently, $\pi = \Pr[S^+ \neq S \mid Y = 1]$ is determined entirely by the data.

Contrastive loss. With one positive and K negatives, the InfoNCE loss is

$$\mathcal{L}_{\text{NCE}}(\theta) = -\mathbb{E}_{(x,x^+)\sim p_{\text{pos}}}\left[\log \frac{g_\tau(z, z^+)}{g_\tau(z, z^+) + \sum_{k=1}^{K} g_\tau(z, z_k^-)}\right]. \tag{6}$$

Let $C \in \{\text{cross}, \text{within}\}$ be the indicator

$$C = \begin{cases} \text{cross} & \text{if } S^+ \neq S, \\ \text{within} & \text{if } S^+ = S \end{cases} \tag{7}$$
$$\Rightarrow \quad \Pr[C = \text{cross}] = \pi, \quad \Pr[C = \text{within}] = 1 - \pi.$$

Lemma 3.1 (InfoNCE lower bound [32])**.** *For any encoder f_θ and any positive-pair distribution,*

$$-\mathcal{L}_{\text{NCE}}(\theta) + \log K = \underbrace{I(Z; Z^+)}_{\textit{MI of pairs}} - \underbrace{D_{\text{KL}}\big(p_{Z,Z^+} \,\|\, p_Z p_{Z^+}\big)}_{\geq 0} \tag{8}$$
$$\leq I(Z; Z^+).$$

Thus minimising $\mathcal{L}_{\text{NCE}}$ maximises the mutual information $I(Z; Z^+)$.

Assumptions (for this proposition only).

1. (*Pair-wise Markov property*)

$$Z \perp\!\!\!\perp Z^+ \mid \begin{cases} Y & \text{if } C = \text{cross}, \\ (Y, S) & \text{if } C = \text{within}. \end{cases}$$

2. C is a *deterministic* function of (S, S^+); hence $C \perp\!\!\!\perp Z \mid (Y, S)$ and $C \perp\!\!\!\perp Z^+ \mid (Y, S)$.

Proposition 3.2 (Mutual-information decomposition)**.** *Under Assumptions 1–2,*

$$I(Z; Z^+) \;=\; I(Z; Y) \;+\; (1 - \pi)\, I(Z; S \mid Y). \tag{9}$$

Proof (chain rule only). Start from the law of total expectation for mutual information:

$$I(Z; Z^+) = \underbrace{I(Z; Z^+ \mid C)}_{\text{case analysis}} + \underbrace{I(Z; C)}_{=0 \text{ by (Assumption 2)}} - \underbrace{I(Z; C \mid Z^+)}_{=0 \text{ by (Assumption 2)}}. \tag{10}$$

Because $I(Z; C) = 0$ and $I(Z; C \mid Z^+) = 0$, only the first term remains. Expand it with the definition of conditional mutual information:

$$\begin{aligned} I(Z; Z^+ \mid C) = \pi \cdot I(Z; Z^+ \mid C = \text{cross}) \\ + (1 - \pi) \cdot I(Z; Z^+ \mid C = \text{within}). \end{aligned} \tag{11}$$

the cross-S branch:
When $C = \text{cross}$, Assumption 1 gives the Markov chain $Z \perp\!\!\!\perp Z^+ \mid Y$. By the chain rule,

$$I(Z; Z^+ \mid C = \text{cross}) = I(Z; Y \mid C = \text{cross}) \quad (Z \perp\!\!\!\perp Z^+ \mid Y). \tag{a}$$

the within-S branch:
When $C = \text{within}$, the Markov chain is $Z \perp\!\!\!\perp Z^+ \mid (Y, S)$. A second application of the chain rule yields

$$\begin{aligned} I(Z; Z^+ \mid C = \text{within}) \\ = I(Z; Y, S \mid C = \text{within}) \\ = I(Z; Y \mid C = \text{within}) + I(Z; S \mid Y, C = \text{within}). \end{aligned} \tag{b}$$

dropping the C-condition inside the MI terms. Because C is a function of (S, S^+) and is independent of Z once (Y, S) is fixed (Assumption 2), conditioning on C adds no information beyond (Y, S):

$$\begin{aligned} I(Z; Y \mid C) = I(Z; Y), \\ I(Z; S \mid Y,\, C = \text{within}) = I(Z; S \mid Y). \end{aligned} \tag{c}$$

Plugging (a)–(c) into the weighted sum gives

$$\begin{aligned} I(Z; Z^+) = \pi\, I(Z; Y) + (1 - \pi)\big[\, I(Z; Y) + I(Z; S \mid Y)\big] \\ = I(Z; Y) \;+\; (1 - \pi)\, I(Z; S \mid Y), \end{aligned}$$

which is exactly (9). □

Theorem 3.3 (InfoNCE $\Longleftrightarrow$ information bottleneck)**.** *Let $\lambda := 1 - \pi$. Under the assumptions of Proposition 3.2,*

$$\mathit{argmin}_\theta\, \mathcal{L}_{\mathrm{NCE}}(\theta) \;=\; \mathit{argmax}_\theta \Big\{ I(Z; Y) - \lambda\, I(Z; S \mid Y) \Big\}. \tag{12}$$

Proof. Combine Lemma 3.1 and the exact identity (9):

$$\begin{aligned} \mathcal{L}_{\mathrm{NCE}}(\theta) = -I(Z; Z^+) + \log K \\ = -I(Z; Y) - (1 - \pi) I(Z; S \mid Y) + \log K. \end{aligned} \tag{13}$$

Since $\log K$ is constant in θ, minimising $x\mathcal{L}_{\mathrm{NCE}}$ is equivalent to maximising the right-hand side of (12). □

Table 1: Hyperparameter configuration used for training on the Adult, Health, and German datasets. All models were optimized using the Adam optimizer with a fixed learning rate and temperature for the contrastive loss.

Hyperparameter	**Adult**	**Health**	**German**
Encoder hidden layers	[64, 64, 64]	[128, 64, 64]	[32, 32, 32]
Classifier hidden layers	[16]	[16]	[16]
Epochs	100	100	100
Contrastive loss temperature (τ)	1	1	1
Learning rate	1×10^{-3}	1×10^{-3}	1×10^{-3}
Optimizer	Adam	Adam	Adam

Corollary 3.3.1. *The hyper-parameter λ that trades off predictive utility $I(Z;Y)$ against conditional leakage $I(Z;S \mid Y)$ is* entirely data-driven*: large values of π (many cross-group pairs) automatically reduce the penalty on $I(Z;S \mid Y)$ and vice-versa.*

Implications. Equation (12) shows that our *pair-selection policy alone* turns standard contrastive learning into an **information bottleneck** that *(i)* preserves label-relevant bits and *(ii)* suppresses sensitive bits *conditioned* on the label. Unlike adversarial critics or explicit MI estimators, no additional modules or tunable coefficients are needed; fairness–accuracy trade-offs emerge directly from the data distribution. Empirically, we observe that increasing π tightens demographic-parity and equal-opportunity gaps while maintaining task performance, corroborating the theoretical guarantee.

4 Experiments

We provide the full experimental setup details, including model architecture and training hyperparameters, in Table 1. These configurations were selected through empirical tuning based on validation performance. The model training was performed using an NVIDIA GeForce RTX 3090 GPU.

4.1 Datasets.

We validate our model on three benchmark datasets in the fairness domain.

- **Adult.** [12] This dataset contains demographic data of 48,842 individuals, and the main task is to predict whether the income of the individual is greater than 50k or not. The sensitive attribute is gender.
- **German Credit** [12] This dataset consists of 1000 individuals and their banking information. The primary task is to predict an individual's credit in repaying their loan. The sensitive attribute is age. Consistent with the setup in [13], we binarize the age feature into "younger" and "older" groups, treating the "older" group as the privileged class. The threshold of 25 years is chosen based on findings from [25], which identified this split as having the highest potential for discriminatory impact.
- **Heritage Health**[1] This dataset contains information of about 50k patients and their corresponding medical conditions. The task is to predict the Charlson Comorbidity Index, a 10-year mortality risk index. The sensitive attribute is age, which has been categorized into nine values. Prior analyses have shown that the dataset exhibits bias against older individuals.

4.2 Evaluation Metrics.

Various fairness notions have been defined and utilized in the fairness domain. Group fairness, including metrics such as Demographic Parity (DP), Equalized Odds (EO), and Equal Opportunity (EOPP) [20, 41], is a commonly used concept in the literature. In this study, we adopt **Demographic Parity (DP)**, also known as statistical parity, as our main fairness metric. Demographic parity ensures that the probability of receiving a favorable outcome is being equitably distributed across groups(privileged and unprivileged). Specifically, this metric requires that the likelihood of all

[1] https://foreverdata.org/1015/index.html

positive predictions (both true positives and false positives) be similar across these groups. Thus, discrimination or disparities can be quantified by measuring the difference between the conditional probabilities of positive predictions for the privileged and unprivileged groups:

$$P(\hat{Y} = 1 \mid X, S = 1) = P(\hat{Y} = 1 \mid X, S = 0) \tag{14}$$

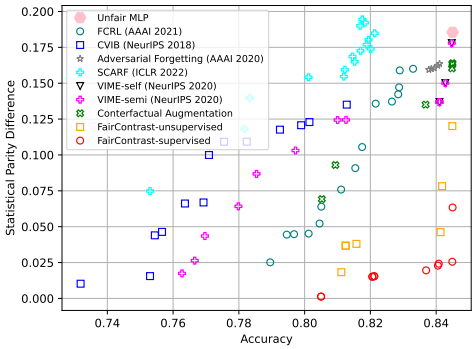
4.3 Baselines.

We evaluate our proposed model, FairContrast, in both unsupervised and self-supervised settings and compare our method with various baselines:

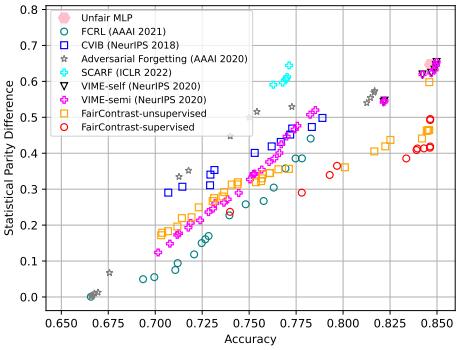
- **Base MLP** We train an MLP classifier without incorporating any fairness measures as our biased base model.
- **FCRL** [16] A fair representation learning framework that introduces a robust method to control parity using mutual information based on contrastive information estimators. By constraining mutual information between representations and sensitive attributes, [16] controls the parity of any downstream classifier.
- **CVIB** [31] A fair representation learning framework that proposes a conditional variational autoencoder to derive representations invariant to sensitive attribute. Their approach is based on a single information-theoretic optimization without adversarial training.
- **Adversarial forgetting** [24] A novel representation learning framework for invariance induction through the "forgetting" mechanism as an information bottleneck to learn invariant representations.
- **Counterfactual Data Augmentation** To demonstrate the importance of data augmentation in contrastive learning, we also conducted a comparison between our method and counterfactual data augmentation. This approach is based on counterfactual fairness [28], where a decision is considered fair if it is the same in both the actual world and the counterfactual world. We generate counterfactual data points by converting the sensitive attribute to counterfactual values, while leaving all other attributes unchanged. We then integrate this data augmentation into our supervised contrastive learning framework. This data aumentation is only applicable on binary sensitive attributes.
- **SCARF** [3] As discussed earlier, SCARF is a state-of-the-art contrastive learning approach on tabular data. They mask 60% of the features and use Random Feature Corruption to replace these masked values. They deploy these data augmentations as positive pairs in their contrastive learning framework. They use InfoNCE as their loss function.
- **VIME** [45] As discussed earlier, VIME is a state-of-the-art self-supervised framework for tabular data. They pre-train an encoder on unlabeled masked data to extract representations. They used Bernoulli distribution to randomly generate the mask and the CutMix method [46] to fill the masked value. We evaluated this model in both semi- and self-supervised settings.

4.4 Experimental Results.

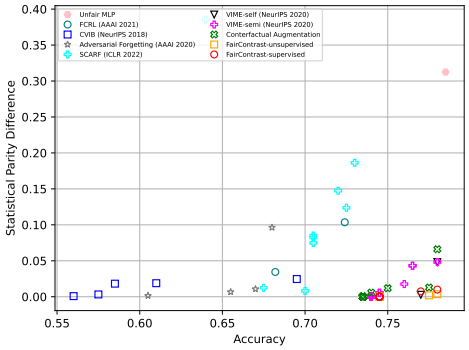
The trade-off between accuracy and fairness across three datasets is shown in Figure 2. The optimal region of the graph is in the lower right corner, indicating higher accuracy and fair outcome (lower demographic parity). Similarly to [16], the results reported for various benchmarks are average accuracy and maximum demographic parity over five runs with random seeds. As evident from the figures, in the Adult dataset, our FairContrast-supervised model stands out, achieving the highest accuracy within the demographic parity range of $0 \sim 0.075$, demonstrating its ability to provide fair predictions while maintaining strong performance. Within the demographic parity range of $0.075 \sim 0.125$, our FairContrast-unsupervised model outperforms others, further showcasing the robustness of our framework even without supervision. In contrast, models such as VIME and SCARF, which are not explicitly designed to address fairness, exhibit a higher bias in their results, reflected by higher DP values, similar to the unfair MLP. For the German dataset, both the FairContrast-supervised and FairContrast-unsupervised models continue to demonstrate superior performance, particularly in the demographic parity range of $0 \sim 0.05$. Comparatively, models such as Adversarial Forgetting and



(a) UCI Adult



(b) Heritage Health



(c) German Credit

Figure 2: Accuracy-fairness trade-off and comparison to various benchmark models across three benchmark datasets: (a) UCI Adult dataset, (b) Heritage Health dataset, and (c) German Credit. The optimal region on the graph is the lower right corner, representing high accuracy and low demographic parity. Our model demonstrates a superior fairness-accuracy trade-off.

CVIB achieve lower DP values, but at the cost of significant accuracy loss. VIME model shows less bias in this dataset, as indicated by their position on the graph. In the Health dataset, our FairContrast-supervised model achieves the highest accuracy within the demographic parity range of $0.3 \sim 0.5$, confirming its effectiveness in providing fair and accurate predictions even in the more challenging dataset. Our FairContrast-unsupervised model shows comparable performance within this range, further underscoring the versatility of our approach. When focusing on the demographic parity range of $0.2 \sim 0.3$, both our supervised model and FCRL exhibit comparable accuracy, indicating that FCRL is also competitive in this particular fairness range. However, models like SCARF and VIME again demonstrate higher bias, as reflected by their positions further up in the DP range. Across all three datasets, our FairContrast models, both supervised and unsupervised, consistently occupy the optimal region of the trade-off graphs, balancing high accuracy with low demographic parity difference. This confirms the effectiveness of our approach in achieving fairness without compromising performance. In contrast, state-of-the-art models like VIME and SCARF, which do not explicitly target fairness, exhibit bias levels similar to the Unfair MLP, as evidenced by their higher DP values across the datasets. This analysis highlights the robustness and effectiveness of our FairContrast framework to ensure that models not only perform well, but also adhere to fairness constraints, making it a valuable contribution to the field of fair representation learning.

For quantitative comparison, we also report the best accuracy corresponding to the worst-case scenario of demographic parity results for all models on the three datasets, summarized in Table 2. Our proposed FairContrast-supervised model consistently demonstrates superior performance across all three datasets—Adult, German, and Health—achieving the best or nearly the best results in both accuracy and fairness (lowest DP). Specifically, in the Adult dataset, FairContrast-supervised achieved

Table 2: Accuracy and demographic parity (DP) on three benchmark datasets. Lower DP indicates higher fairness. Gray highlights the baseline model (Unfair MLP), Green highlights our method (FairContrast), **bold** marks the best performance among all methods, and *italic* denotes the second-best.

Dataset	Model	Accuracy ↑	DP ↓
Adult	Unfair MLP	84.5	0.1855
	FCRL	83.29	0.16
	CVIB	81.28	0.1350
	Adversarial forgetting	84.1	0.1635
	Counterfactual	84.49	0.1639
	VIME-semi	84.27	0.15
	VIME-self	84.47	0.1779
	SCARF	82.13	0.1848
	FairContrast (Ours)-unsupervised	**84.4**	*0.1201*
	FairContrast (Ours)-supervised	**84.4**	**0.0255**
German	Unfair MLP	78.5	0.3125
	FCRL	72.4	0.1035
	CVIB	69.5	*0.0244*
	Adversarial forgetting	68	0.0963
	Counterfactual	78	0.066
	VIME-semi	76.5	0.0431
	VIME-self	78	0.0482
	SCARF	73	0.1862
	FairContrast (Ours)-unsupervised	**78**	0.0297
	FairContrast (Ours)-supervised	**78**	**0.0099**
Health	Unfair MLP	84.64	0.6468
	FCRL	78.27	*0.4407*
	CVIB	78.9	0.4982
	Adversarial forgetting	81.68	0.5733
	VIME-semi	82.2	0.5463
	VIME-self	84.22	0.6192
	SCARF	77.12	0.6444
	FairContrast (Ours)-unsupervised	*84.19*	0.4410
	FairContrast (Ours)-supervised	**84.3**	**0.4135**

an accuracy of 84.4 % with a DP of 0.0255, indicating a substantial reduction in bias compared to other models. Similarly, in the German dataset, our model maintained strong accuracy at 78 % while achieving the lowest DP of 0.0099, further confirming its ability to mitigate bias effectively. Our proposed unsupervised model also performs well, with relatively low DP scores compared to other models, though its accuracy is slightly lower than that of the FairContrast-supervised model. Although other models like FCRL and CVIB offer competitive alternatives, particularly in fairness, they often fall short in achieving the same level of accuracy or in minimizing bias as effectively as FairContrast. State-of-the-art models, such as VIME and SCARF, which are not specifically focused on enhancing fairness, achieve accuracy comparable to our supervised model. However, the bias in their representations is similar to that found in the unfair MLP model. Overall, our FairContrast framework represents a significant advancement in contrastive learning for tabular data, offering a robust solution that does not compromise on fairness while maintaining strong predictive performance.Our results suggest that contrastive learning, when properly supervised and designed with fairness in mind, can lead to models that perform well both in terms of accuracy and fairness.

4.5 Ablation on Classification Loss Weight

To further analyze the impact of the loss weight α on the fairness–accuracy trade-off, we conduct an ablation study using the Adult dataset. Specifically, we evaluate the Area Over the Fairness–Accuracy Pareto Curve (AOC) at varying values of α in both supervised and unsupervised settings. The AOC summarizes the feasible region in the parity–accuracy space and offers a quantitative proxy for a method's capacity to provide accurate predictions under fairness constraints. A higher AOC indicates that a method can achieve better utility while satisfying a wider range of demographic parity thresholds.

Following the interpretation presented in Gupta et al. [16], the parity–accuracy curve reflects the achievable frontier between accuracy and fairness, where methods that shift the curve closer to the bottom right are more desirable. Thus, the area under this frontier—the AOC—represents the

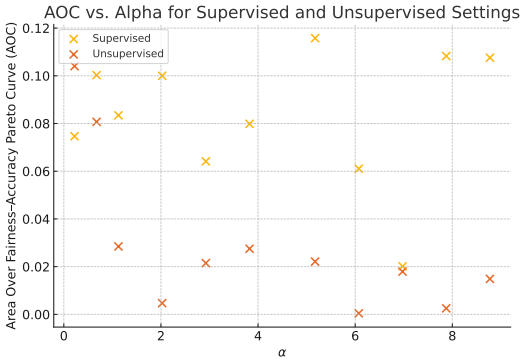


Figure 3: Effect of varying α on the Area Over the Fairness–Accuracy Pareto Curve (AOC) for supervised and unsupervised settings. Each point represents the AOC score at a specific α value. The trade-off stabilizes for $\alpha > 1$, indicating consistent fairness–accuracy performance in both learning modes.

volume of favorable outcomes. In our results (Fig. 3), performance stabilizes across both learning settings when $\alpha > 1$, suggesting that moderate weighting of the classification loss produces robust representations with respect to both utility and fairness.

5 Conclusion

Contrastive learning has shown its effectiveness in improving model robustness and generalizability across various domains, including Natural Language Processing (NLP), computer vision, and speech recognition. Recently, there has been an increasing interest in applying self-supervised contrastive learning to tabular data. Although handling data types such as images, text, or speech is less challenging due to their feature correlations, semantic relationships, and structured information, tabular datasets pose unique challenges due to the lack of explicit relationships within their features, which can vary across different datasets.

In this paper, we argue that current state-of-the-art models for tabular data, such as VIME and SCARF, do not address fairness issues. The fairness of the learned representations has not been thoroughly examined, and these models exhibit biases in their predictions, leading to discrimination in the downstream tasks. To address this, we propose supervised and self-supervised contrastive learning frameworks for tabular data to mitigate bias and improve fairness. Our approach involves selective pairing of samples based on specific criteria and incorporating these pairs into the training process with a contrastive loss. This method encourages the embeddings of paired instances to be closer together, reducing discrimination based on sensitive attributes.

We evaluated our proposed method using three benchmark datasets in the fairness domain. The results show a significant reduction in bias compared to existing state-of-the-art frameworks for tabular data. Furthermore, these fair representations can be applied to any downstream tasks.

Although our framework achieves promising results, several limitations should be noted.

First, the work mainly addresses group fairness through metrics such as Demographic Parity. While these are useful for capturing disparities between subgroups, they do not fully account for individual fairness, which ensures that similar individuals are treated similarly. Future research could explore approaches that jointly address both group and individual level fairness.

Second, our method currently emphasizes demographic parity. In real-world scenarios, multiple fairness definitions may be relevant, and these can sometimes conflict with one another. Extending the framework to accommodate several fairness criteria simultaneously would increase its practical flexibility.

Third, although our approach was designed with tabular data in mind, the underlying methodology could also be extended to other data types, including images, text, or multimodal systems. Exploring these extensions remains an open avenue for future work.

References

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

[2] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.

[3] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. *arXiv preprint arXiv:2106.15147*, 2021.

[4] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

[5] Carlo Alberto Barbano, Benoit Dufumier, Enzo Tartaglione, Marco Grangetto, and Pietro Gori. Unbiased supervised contrastive learning. *arXiv preprint arXiv:2211.05568*, 2022.

[6] Souradip Chakraborty, Ekansh Verma, Saswata Sahoo, and Jyotishka Datta. Fairmixrep: Self-supervised robust representation learning for heterogeneous data with fairness constraints. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 458–463. IEEE, 2020.

[7] Suiyao Chen, Jing Wu, Naira Hovakimyan, and Handong Yao. Recontab: Regularized contrastive representation learning for tabular data. *arXiv preprint arXiv:2310.18541*, 2023.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

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

[10] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2022.

[11] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014.

[12] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[13] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[15] Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*, 2020.

[16] Umang Gupta, Aaron M Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7610–7619, 2021.

[17] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[18] Ehsan Hajiramezanali, Nathaniel Lee Diamant, Gabriele Scalia, and Max W Shen. Stab: Self-supervised learning for tabular data. In *NeurIPS 2022 First Table Representation Workshop*, 2022.

[19] Sungwon Han, Seungeon Lee, Fangzhao Wu, Sundong Kim, Chuhan Wu, Xiting Wang, Xing Xie, and Meeyoung Cha. Dualfair: Fair representation learning at both group and individual levels via contrastive self-supervision. In *Proceedings of the ACM Web Conference 2023*, pages 3766–3774, 2023.

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

[21] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[22] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34:26449–26461, 2021.

[23] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

[24] Ayush Jaiswal, Daniel Moyer, Greg Ver Steeg, Wael AbdAlmageed, and Premkumar Natarajan. Invariant representations through adversarial forgetting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4272–4279, 2020.

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

[26] Minguk Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. *Advances in Neural Information Processing Systems*, 33:21357–21369, 2020.

[27] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[28] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[29] Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science*, 65(7):2966–2981, 2019.

[30] Martin Q Ma, Yao-Hung Hubert Tsai, Paul Pu Liang, Han Zhao, Kun Zhang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Conditional contrastive learning for improving fairness in self-supervised learning. *arXiv preprint arXiv:2106.02866*, 2021.

[31] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. *Advances in neural information processing systems*, 31, 2018.

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[33] Dana Pessach and Erez Shmueli. Algorithmic fairness. In *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, pages 867–886. Springer, 2023.

[34] Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Contrastive learning for fair representations. *arXiv preprint arXiv:2109.10645*, 2021.

[35] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

[36] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[37] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

[38] Yao-Hung Hubert Tsai, Tianqin Li, Weixin Liu, Peiyuan Liao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Integrating auxiliary information in self-supervised learning. *arXiv preprint arXiv:2106.02869*, 2021.

[39] Yao-Hung Hubert Tsai, Tianqin Li, Martin Q Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Conditional contrastive learning with kernel. *arXiv preprint arXiv:2202.05458*, 2022.

[40] Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34: 18853–18865, 2021.

[41] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[42] Wei-Yao Wang, Wei-Wei Du, Derek Xu, Wei Wang, and Wen-Chih Peng. A survey on self-supervised learning for non-sequential tabular data. *arXiv preprint arXiv:2402.01204*, 2024.

[43] Zifeng Wang and Jimeng Sun. Transtab: Learning transferable tabular transformers across tables. *Advances in Neural Information Processing Systems*, 35:2902–2915, 2022.

[44] Mehdi Yazdani-Jahromi, Ali Khodabandeh Yalabadi, AmirArsalan Rajabi, Aida Tayebi, Ivan Garibay, and Ozlem Garibay. Fair bilevel neural network (fairbinn): On balancing fairness and accuracy via stackelberg equilibrium. *Advances in Neural Information Processing Systems*, 37:105780–105818, 2024.

[45] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela Van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33:11033–11043, 2020.

[46] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[47] Fengda Zhang, Kun Kuang, Long Chen, Yuxuan Liu, Chao Wu, and Jun Xiao. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *The Eleventh International Conference on Learning Representations*, 2022.

[48] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly outline the paper's scope and key contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations and prospective research directions are presented in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assumptions and proofs are mentioned in methodology section 3.1.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The required details for reproducing this model have been provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All the datasets used in this study are publicly available. the link to the code can be provided at any time. For the review submission, we excluded the link to ensure anonymous evaluation.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The hyperparameters details are provided in Table 1 in the Experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The significance of our work is clearly demonstrated through the provided tables and figures.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As noted in the Experiments section, the computations for this study were carried out on a system equipped with an NVIDIA GeForce RTX 3090 GPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This study does not present any potential violations.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The assets employed in this study are entirely original, publicly available (e.g., datasets), or properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: he GitHub repository containing the code and datasets will be shared upon acceptance. To preserve anonymity during review, the link is withheld for now.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.