Learning Fairer Representations with FairVIC

Charmaine Barker [1] **Daniel Bethell** [1] **Dimitar Kazakov** [1]

Abstract

Mitigating bias in automated decision-making systems, particularly in deep learning models, is a critical challenge due to nuanced definitions of fairness, dataset-specific biases, and the inherent trade-off between fairness and accuracy. To address these issues, we introduce FairVIC, an innovative approach that enhances fairness in neural networks by integrating variance, invariance, and covariance terms into the loss function during training. Unlike methods that rely on predefined fairness criteria, FairVIC abstracts fairness concepts to minimise dependency on protected characteristics. We evaluate FairVIC against comparable bias mitigation techniques on benchmark datasets, considering both group and individual fairness, and conduct an ablation study on the accuracy-fairness trade-off. FairVIC demonstrates significant improvements ($\approx 70\%$) in fairness across all tested metrics without compromising accuracy, thus offering a robust, generalisable solution for fair deep learning across diverse tasks and datasets.

1. Introduction

With the ever-increasing utilisation of Artificial Intelligence (AI) in everyday applications, neural networks have emerged as pivotal tools for automated decision making systems in sectors such as healthcare (Esteva et al., 2017), finance (Dixon et al., 2017), and recruitment (Vardarlier and Zafer, 2020). However, bias in the data– stemming from historical inequalities, imbalanced distributions, or flawed feature representations–are often learned by these models, posing significant challenges to fairness. Such bias can lead to adverse decisions affecting real lives. For instance, several studies have shown how bias in facial recognition technologies disproportionately misidentifies individuals of certain ethnic backgrounds (Birhane, 2022; Cavazos et al., 2020), leading to potential discrimination in law enforcement and hiring practices.

Real-world consequences exemplify the urgent need to address these challenges at the core of AI development. Ensuring fairness in deep learning models presents complex challenges, primarily due to the black-box nature of these models, which often complicates understanding and interpreting decisions. Moreover, the dynamic and high-dimensional nature of the data involved, combined with nuances in fairness definitions, further complicates the detection and correction of bias. This complexity necessitates the development of more sophisticated, inherently fair algorithms.

Previous mitigation strategies dealing with algorithmic bias – whether through pre-processing, in-processing, or post-processing – have significant limitations. Pre-processing techniques, which attempt to cleanse biased data, are labour-intensive, dependent on expert intervention (Salimi et al., 2019), and only eliminate considered biases. Current in-processing methods frequently lead to unstable models and often rely upon arbitrary definitions of fairness (Caton and Haas, 2020). Post-processing techniques, which adjust model predictions directly, ignore deeper issues without addressing the underlying biases in the data and model. These approaches lack stability, generalisability, and the ability to ensure fairness across multiple metrics (Berk et al., 2017).

In this paper, we introduce FairVIC (**Fair**ness through **V**ariance, **I**nvariance, and **C**ovariance), a novel approach that embeds fairness directly into neural networks by optimising a custom loss function. This function is designed to minimise the correlation between decisions and protected characteristics while maximising overall prediction performance. FairVIC integrates fairness through the concepts of variance, invariance, and covariance during the training process, making it more principled and intuitive, and universally applicable to diverse datasets. Unlike previous methods that often optimise to a chosen fairness metric, FairVIC offers a robust, generalisable solution that introduces an abstract concept of fairness to significantly reduce bias. Our experimental evaluations demonstrate FairVIC's ability to significantly improve performance in all fairness metrics tested without compromising prediction accuracy. We compare our proposed method against comparable in-

[1]Department of Computer Science, University of York, York, United Kingdom. Correspondence to: Charmaine Barker <charmaine.barker@york.ac.uk>.

processing bias mitigation techniques, such as regularisation and constraint approaches, and highlight the improved, robust performance of our FairVIC model.

Our contributions in this paper are multi-fold:

- A novel, generalisable in-processing bias mitigation technique for neural networks.
- A comprehensive experimental evaluation, using a multitude of comparable methods on a variety of metrics across several datasets, including different modalities such as tabular and text.
- An extended analysis of our proposed method to examine its robustness, including a full ablation study on the lambda weight terms within our loss function.

This paper is structured as follows: Section 2 discusses current approaches to mitigating bias throughout each processing stage. Section 3 describes any preliminary details for this work, including the fairness metrics used in the evaluation. Section 4 outlines our method, including how each term in our loss function is calculated and an algorithm detailing how these terms are applied. Section 5 describes the experiments carried out, Section 6 outlines the results with discussion, and Section 7 concludes this work. Extra information, including the dataset metadata and more extensive experiments, is to be found in the Appendix.

2. Related Work

There exist three broad categories of mitigation strategies for algorithmic bias: pre-processing, in-processing, and post-processing. Each aims to increase fairness differently by acting upon either the training data, the model itself, or the predictions outputted by the model, respectively.

Pre-processing methods aim to *fix* the data before training, recognising that bias is primarily an issue with the data itself (Caton and Haas, 2020). In practice, this can be done a number of different ways, such as representative sampling, or re-sampling the data to reflect the full population (Shekhar et al., 2021; Ustun et al., 2019), reweighing the data such that different groups influence the model in a representative way (Calders and Žliobaitė, 2013; Kamiran and Calders, 2012), or generating synthetic data to balance out the representation of each group (Jang et al., 2021). Another set of approaches utilises causal methods to delineate relationships between sensitive attributes and the target variables within the data (Chiappa and Isaac, 2019; Kusner et al., 2017; Russell et al., 2017). Such techniques as these are labour-intensive and do not generalise well, requiring an expert with knowledge of the data to manually process each case of a new dataset (Salimi et al., 2019). They also cannot provide assurances that all bias has been removed – a model may draw upon non-linear/ complex relationships between features that lead to bias, which are hard for the expert/method to spot.

In-processing methods aim to train models to make fairer predictions, even upon biased data. There are a plethora of ways in which this has been done. For example, Celis et al. (2019) and Agarwal et al. (2018) utilise a chosen fairness metric and perform constraint optimisation during training. This has the effect that a single fairness metric needs to be chosen, introducing human bias (Caton and Haas, 2020), and this metric must perfectly capture the bias within the data to effectively mitigate it. Therefore, fairness cannot be achieved across multiple definitions in this way (Caton and Haas, 2020). Another approach involves incorporating an adversarial component during model training that penalises the model if protected characteristics can be predicted from its outputs (Zhang et al., 2018; Wadsworth et al., 2018; Xu et al., 2019). These methods are often effective but their main shortcoming is seen in their instability. Finally, the most relevant comparisons from previous work to our proposed method are regularisation-based techniques that incorporate fairness constraints or penalties directly into the model's loss function during training. There are a number of ways that this has been done, such as through data augmentation strategies to promote less sensitive decision boundaries (Chuang and Mroueh, 2021) or by incorporating fairness adjustments into the boosting process (Cruz et al., 2023). The performance of these models differs from approach to approach, and those that work by constraining the model by a fairness metric directly suffer from the issue of human bias and misrepresenting the bias within the data/model.

Post-processing techniques involve adjusting model predictions or decision rules after training to ensure fair outcomes. In practice, decision thresholds have been adjusted for different groups to achieve equal outcomes in a particular metric (Hardt et al., 2016). Alternatively, labels near the decision boundary can be altered to favour less biased outcomes (Kamiran et al., 2012; 2018). Calibration (Kim et al., 2018; Noriega-Campero et al., 2019) adjusts the predictions of the model directly so that the proportion of positive instances is equal across each sub-group. This category of methods can oversimplify fairness, and they do not fix the underlying issue within the model. For those techniques that require the specification of a single fairness metric, the same issue applies surrounding this choice as before.

To summarise, there currently lies a number of issues which have not yet been solved in parallel within one technique. These are: stability, generalisability, equal improvements to fairness across metrics (Berk et al., 2017), and built without requirements for user-induced definitions of fairness. In this paper, we solve all these requirements for an effective, generalisable approach to mitigate bias

through FairVIC.

3. Preliminaries

3.1. VICReg

Variance-Invariance-Covariance Regularization (VICReg) (Bardes et al., 2021) has previously been used in self-supervised learning to tackle feature collapse and redundancy. It maximises variance across features to ensure the model produces diverse outputs for different inputs, minimises invariance between augmented representations of the same input to enhance stability, and reduces covariance among features to capture a broader range of information. VICReg is confined to this specific context and objective, and the application of these principles outside of self-supervised methods remains largely unexplored. In contrast, FairVIC extends these principles to supervised learning for bias mitigation. This adaptation addresses the challenges of fairness in decision-making systems, expanding the application of VIC principles beyond their original scope and offering a novel, generalisable solution to fairness in supervised learning models.

3.2. Group Fairness Metrics

In this section, we introduce notation and state the fairness measures that we use to quantify bias.

Equalized Odds Difference requires that both the True Positive Rate (TPR) and False Positive Rate (FPR) are the same across groups defined by the protected attribute, where $TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{FP+TN}$ (Hardt et al., 2016). Therefore, we calculate $\max\left(\left|FPR_u - FPR_p\right|, \left|TPR_u - TPR_p\right|\right)$, where u represents the unprivileged groups and p the privileged group and 0 signifies perfect fairness.

Average Absolute Odds Difference averages the absolute differences in the false positive rates and true positive rates between groups, defined as $\frac{1}{2}(|FPR_u - FPR_p| + |TPR_u - TPR_p|)$, where u represents the unprivileged groups and p the privileged group, with 0 signifying perfect fairness.

Statistical Parity Difference evaluates the difference in the probability of a positive prediction between groups, aiming for 0 to signify perfect fairness. Formally, $DP = |P(\hat{Y} = 1|u) - P(\hat{Y} = 1|p)|$, where u represents the unprivileged groups, p the privileged group, and $\hat{Y} = 1$ a positive prediction (Dwork et al., 2012).

Disparate Impact compares the proportion of positive outcomes for the unprivileged group to that of the privileged group, with a ratio of 1 indicating no disparate impact, and therefore perfect fairness. Denoted as $DI = \frac{P(\hat{Y}=1|u)}{P(\hat{Y}=1|p)}$, where u represents the unprivileged groups, p the privileged group, and $\hat{Y} = 1$ a positive prediction (Feldman et al., 2015).

3.3. Individual Fairness

While FairVIC aims to increase group fairness, the invariance term promotes direct improvements in individual fairness. This can be observed in our evaluations through counterfactual fairness (Kusner et al., 2017). Counterfactual fairness ensures that decisions made by an algorithm are fair even when considering hypothetical (counterfactual) scenarios. For each individual, the sensitive attribute is switched to assess the model's ability to perform equally in both the original and counterfactual scenarios.

Formally, if u denotes the unprivileged group, p the privileged group and $\hat{Y}$ is the decision outcome, then the model is considered counterfactually fair if $\hat{Y}_u = \hat{Y}_p$ for different groups u and p of the sensitive attribute while all non-sensitive features remain the same.

4. Approach

We propose FairVIC (**Fair**ness through **V**ariance, **I**nvariance, and **C**ovariance), a novel loss function that enables a model's ability to learn fairness in a robust manner. FairVIC is comprised of three terms: variance, invariance, and covariance. Optimising for these three terms encourages the model to be stable and consistent across protected characteristics, thereby reducing bias during training. By adopting this broad, generalised approach to defining bias, FairVIC significantly improves performance across a range of fairness metrics. This makes it an effective strategy for reducing bias across various applications, ensuring more equitable outcomes in diverse settings.

4.1. FairVIC Training

To understand how FairVIC operates, it is crucial to define variance, invariance, and covariance within the context of fairness:

Variance: This term promotes diversity in the latent representations by penalizing low variance in the bottleneck embeddings of the neural network. It ensures the embeddings capture sufficient information, not relying upon a trivial relationship such as the protected characteristic in order to find a solution.

$$L_{\text{var}} = \frac{1}{N}\sum_{i=1}^{N} \max(0, \gamma - \sigma(z)) \tag{1}$$

where $\sigma(z)$ represents the standard deviation of the embeddings, γ is a margin parameter that controls the desired variability, and N is the number of samples.

Algorithm 1 FairVIC Loss Function

Input: Model M, Epochs E, Batch size B, Data D, Protected attribute P, Weights $(\lambda_{\text{acc}}, \lambda_{\text{var}}, \lambda_{\text{inv}}, \lambda_{\text{cov}})$
Output: Trained Model M
Initialise M
for $e \in \{1, \ldots, E\}$ **do**
  Shuffle data D
  for each batch $\{(X, Y)\} \in D$ with size B **do**
    $\hat{Y} \leftarrow M(X)$
    $Z \leftarrow \text{BottleneckLayer}(X)$
    Calculate FairVIC Loss:
      $L_{\text{acc}} \leftarrow \text{AccuracyLoss}(Y, \hat{Y})$
      $L_{\text{var}} \leftarrow \text{VarianceLoss}(Z)$
      $L_{\text{inv}} \leftarrow \text{InvarianceLoss}(\hat{Y}, M(\text{Flip}(X, P)))$
      $L_{\text{cov}} \leftarrow \text{CovarianceLoss}(\hat{Y}, P)$
      $L_{\text{total}} \leftarrow \lambda_{\text{acc}} L_{\text{acc}} + \lambda_{\text{var}} L_{\text{var}} + \lambda_{\text{inv}} L_{\text{inv}} +$
        $\lambda_{\text{cov}} L_{\text{cov}}$
    Compute gradients $\nabla L_{\text{total}} \leftarrow \frac{\partial L_{\text{total}}}{\partial M}$
    Update model parameters $M \leftarrow M - \alpha \nabla L_{\text{total}}$
  end for
end for

Invariance: This term ensures the model's predictions remain consistent when the protected attribute is flipped, promoting individual/counterfactual fairness.

$$L_{\text{inv}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - \hat{y}_i^*)^2 \tag{2}$$

where $\hat{y}_i$ is the prediction for the original input, $\hat{y}_i^*$ is the prediction for the input with the protected attribute P flipped to its complement, and N is the number of samples.

Covariance: This component seeks to reduce the model's reliance on protected attributes when making predictions, ensuring that decisions are made independently of these attributes. By doing so, it promotes group fairness. The loss function is designed to minimize this covariance, as defined by the following equation:

$$L_{\text{cov}} = \frac{\sqrt{\sum_{i=1}^{N} \left((\hat{y}_i - \mathbb{E}[\hat{y}])^\top \cdot P_i \right)^2}}{N} \tag{3}$$

where $\hat{y}$ is the model's prediction, P is the protected attribute, and N is the number of samples.

During the training of a deep learning model, the model iterates over epochs E. Data is shuffled into batches, upon which the model predicts to produce a set of predictions $\hat{Y}$. Typically, the true labels Y and predictions $\hat{Y}$ are then passed into a suitable accuracy loss function (e.g., binary cross-entropy, hinge loss, Huber loss, etc.) and the resulting loss attempts to be minimised by an optimiser.

In the case of FairVIC, in addition to computing a suitable accuracy loss L_{acc}, we also calculate our three novel terms L_{var}, L_{inv}, and L_{cov} using Equations 1, 2, and 3 respectively. Each of these individual loss terms is then multiplied by its respective weighting factor λ and summed to form the total loss L_{total}. Subsequently, gradients are computed, and the optimiser adjusts the model parameters with respect to this combined loss. Further details are provided in Algorithm 1.

The multipliers λ enable users to balance the trade-off between fairness and predictive performance, which is typical in bias mitigation techniques. Assigning a higher weight to λ_{acc} directs the model to prioritise accuracy while increasing the weights of $(\lambda_{\text{var}}, \lambda_{\text{inv}}, \lambda_{\text{cov}})$ shifts the focus towards enhancing fairness in the model's predictions. In our implementation, the lambda coefficients $(\lambda_{\text{acc}}, \lambda_{\text{var}}, \lambda_{\text{inv}}, \lambda_{\text{cov}})$ are constrained such that their sum equals one. In other words, $\lambda_{\text{acc}} = 1 - \lambda_{\text{var}} - \lambda_{\text{inv}} - \lambda_{\text{cov}}$. This normalisation ensures the optimisation will not produce multiple solutions in the form $\{k.\lambda_{\text{acc}}, k.\lambda_{\text{var}}, k.\lambda_{\text{inv}}, k.\lambda_{\text{cov}}\}, k \in \mathcal{R}$.

5. Experiments

In our experimental evaluation, we assess the performance of FairVIC[1] against a set of comparable in-processing bias mitigation methods on a series of datasets known for their bias. Here, we describe the datasets used and the methods we compare against.

5.1. Datasets

We evaluate FairVIC on five datasets that are used as benchmarks in bias mitigation evaluation due to their known biases towards certain subgroups of people within their sample population. These datasets allow for highlighting the generalisable capabilities of FairVIC across different demographic disparities.

Tabular datasets. The main body of evaluation is done using three tabular datasets: Adult Income (Becker and Kohavi, 1996), COMPAS (Angwin et al., 2022), and German Credit (Hofmann, 1994), all of which are binary classification tasks. Adult Income aims to predict whether an individual's income is over $\$50$K or not. It is known for its gender and racial biases in economic disparity. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset is frequently used for evaluating debiasing techniques. It has a classification goal of predicting recidivism risks and is infamous for its racial biases. Finally, the German Credit dataset was used to assess creditworthiness by classifying individuals as either good or bad credit risks, with known biases related to age and

[1]The code for our FairVIC implementation is available at: `https://anonymous.4open.science/r/FairVIC-BEE7`

gender (Kamiran and Calders, 2009).

Language datasets. To show the ability of FairVIC to work for different data modalities, we also utilise CivilComments-WILDS (Koh et al., 2021) and BiasBios (De-Arteaga et al., 2019) – natural language datasets. We collected a stratified sample of 50,000 text instances from each dataset, ensuring equal representation of both binary classification outcomes. CivilComments-WILDS is comprised of a collection of comments on the Civil Comments platform, with a binary classification goal to label each comment as toxic or non-toxic. We take ethnicity as the protected characteristic where comments are marked as white or non-white. BiasBios is a collection of professional biographies, with gender as the protected characteristic. We take a set of privileged and unprivileged professions, typically associated with a certain gender, as our binary classification goal.

Detailed metadata for each dataset, including our selections for protected groups and classification goals, can be found in Appendix A.1.

5.2. Comparable Techniques

To highlight the performance of FairVIC, we evaluate against five comparable in-processing bias mitigation methods. These are:

Adversarial Debiasing. This method leverages an adversarial network that aims to predict protected characteristics based on the predictions of the main model. The primary model seeks to maximise its own prediction accuracy while minimising the adversary's prediction accuracy (Zhang et al., 2018).

Exponentiated Gradient Reduction. This technique reduces fair classification to a sequence of cost-sensitive classification problems, returning a randomised classifier with the lowest empirical error subject to a chosen fairness constraint (Agarwal et al., 2018).

Meta Fair Classifier. This classifier allows a fairness metric as an input and optimises the model with respect to regular performance and the chosen fairness metric (Celis et al., 2019).

Fair MixUp. This technique generates synthetic samples by linearly interpolating between pairs of training data points by protected attribute to smooth decision boundaries. The loss function is then further constrained by a fairness metric (Chuang and Mroueh, 2021).

FairGBM. This method uses a gradient-boosting decision tree model that integrates fairness constraints directly into the boosting process by adjusting the loss function to account for fairness metrics (Cruz et al., 2023).

Alongside these comparisons, a baseline neural network model using only binary cross-entropy loss was implemented, which exhibits the biases present in the datasets used. Details on the neural network architecture/ hyperparameters used for both the baseline model and the FairVIC model can be found in Appendix A.2.

6. Evaluation

6.1. Core Results Analysis

To assess the prediction and fairness performance of FairVIC[2] and state-of-the-art approaches, we test all methods across each tabular dataset to enable a fair comparison. Table 1 shows these results. We have also provided Figure 1, which visualises the absolute difference from the ideal value of each metric, highlighting how far each method deviates from perfect accuracy and fairness on each tabular dataset.

Across all three datasets, the baseline performs poorly in fairness but obtains higher performance scores, which is expected. For example, in the Adult Income dataset, the baseline model shows a relatively high accuracy (0.8444), while exhibiting poor fairness with regard to Disparate Impact (0.2853). The baseline highlights the need for a bias mitigation approach that works across all metrics simultaneously, as the lower bias in terms of Equalised Odds (0.1330) and Absolute Odds (0.1172) alone could misleadingly suggest that the model is fair, when in reality, the bias may only become evident when captured through a different perspective. This means that approaches relying on a single fairness constraint, such as Exponentiated Gradient Reduction, often fail to address significant bias present in the data.

Overall, FairVIC outperforms all other comparable methods by demonstrating consistent improvements in both fairness and accuracy retention. As seen in Figure 1a, our FairVIC model achieves the lowest cumulative absolute error from perfect accuracy and fairness in the Adult Income dataset, effectively balancing the fairness-accuracy trade-off. The trend is also consistent across the COMPAS and German Credit datasets as seen in Figures 1b and 1c. This again exemplifies the ability of FairVIC's approach to generalise across datasets, making it a more versatile solution.

Other comparable methods are generally not as effective as FairVIC, each exhibiting different shortcomings. For instance, MetaFair often struggles to improve even upon the baseline in cumulative absolute difference from the ideal value, and many techniques struggle to balance the improvements across all fairness metrics, often prioritising Equalised and Absolute Odds over Disparate Impact, particularly in the Adult Income dataset. Similarly, Fair-

[2]FairVIC weights $L_{\text{acc}} = 0.2$, $L_{\text{var, inv}} = 0.1$, and $L_{\text{cov}} = 0.6$ for the Adult Income dataset, and $L_{\text{acc, var, inv}} = 0.1$, and $L_{\text{cov}} = 0.7$ for the COMPAS and German Credit datasets, see Appendix B.3 for discussion on these selections.

Table 1. FairVIC accuracy and fairness results, compared with the baseline model, and five other comparable methods for bias mitigation in-processing for each of the three tabular datasets.

Dataset	Model	Accuracy	F1 Score	Equalized Odds	Absolute Odds	Statistical Parity	Disparate Impact
Adult Income	Baseline (Biased)	0.8444 ± 0.0065	0.6685 ± 0.0118	0.1330 ± 0.0317	0.1172 ± 0.0289	-0.2173 ± 0.0291	0.2853 ± 0.0329
	Adversarial Debiasing	0.8065 ± 0.0048	0.4773 ± 0.0708	0.2127 ± 0.0828	0.1172 ± 0.0443	-0.0405 ± 0.0679	0.7874 ± 0.2185
	Exponentiated Gradient Reduction	0.8027 ± 0.0026	0.4056 ± 0.0052	0.0238 ± 0.0115	0.0167 ± 0.0061	-0.0601 ± 0.0026	0.4602 ± 0.0237
	Meta Fair Classifier	0.5171 ± 0.0602	0.4744 ± 0.0219	0.4826 ± 0.0894	0.2935 ± 0.0497	-0.2098 ± 0.0542	0.7140 ± 0.0812
	Fair MixUp	0.7785 ± 0.0069	0.3815 ± 0.0521	0.1137 ± 0.0928	0.0786 ± 0.0649	-0.0859 ± 0.0591	0.4830 ± 0.2174
	FairGBM	0.8731 ± 0.0026	0.7122 ± 0.0079	0.0658 ± 0.0131	0.0583 ± 0.0092	-0.1707 ± 0.0044	0.3363 ± 0.0151
	FairVIC	$\mathbf{0.8284 \pm 0.0088}$	$\mathbf{0.5314 \pm 0.0509}$	$\mathbf{0.2993 \pm 0.0683}$	$\mathbf{0.1637 \pm 0.0371}$	$\mathbf{-0.0088 \pm 0.0249}$	$\mathbf{0.9803 \pm 0.2220}$
COMPAS	Baseline (Biased)	0.6622 ± 0.0150	0.6118 ± 0.0252	0.3281 ± 0.0574	0.2635 ± 0.0452	-0.2941 ± 0.0459	0.6223 ± 0.0504
	Adversarial Debiasing	0.6581 ± 0.0185	0.6253 ± 0.0124	0.1707 ± 0.0694	0.1363 ± 0.0504	-0.0902 ± 0.1367	0.8982 ± 0.2614
	Exponentiated Gradient Reduction	0.5574 ± 0.0169	0.2981 ± 0.0407	0.0630 ± 0.0333	0.0432 ± 0.0231	-0.0393 ± 0.0257	0.9545 ± 0.0293
	Meta Fair Classifier	0.3471 ± 0.0147	0.4312 ± 0.0380	0.2951 ± 0.1038	0.2257 ± 0.1095	0.2526 ± 0.1070	2.5876 ± 0.6627
	Fair MixUp	0.6122 ± 0.0191	0.5356 ± 0.0437	0.1180 ± 0.0774	0.0871 ± 0.0597	-0.0496 ± 0.0998	0.9427 ± 0.1470
	FairGBM	0.6440 ± 0.0151	0.6254 ± 0.0153	0.2015 ± 0.1128	0.1466 ± 0.0961	0.0881 ± 0.1225	1.2828 ± 0.4058
	FairVIC	$\mathbf{0.6501 \pm 0.0173}$	$\mathbf{0.5934 \pm 0.0357}$	$\mathbf{0.0976 \pm 0.0375}$	$\mathbf{0.0719 \pm 0.0305}$	$\mathbf{-0.0602 \pm 0.0678}$	$\mathbf{0.9139 \pm 0.1135}$
German Credit	Baseline (Biased)	0.7255 ± 0.0284	0.8077 ± 0.0275	0.2234 ± 0.0974	0.1641 ± 0.0936	-0.2218 ± 0.0901	0.7140 ± 0.1203
	Adversarial Debiasing	0.5815 ± 0.1513	0.6302 ± 0.2581	0.1020 ± 0.0418	0.0737 ± 0.0404	-0.0657 ± 0.0335	0.8084 ± 0.2130
	Exponentiated Gradient Reduction	0.7465 ± 0.0300	0.8321 ± 0.0208	0.1232 ± 0.0631	0.0796 ± 0.0348	-0.1084 ± 0.0746	0.8692 ± 0.0896
	Meta Fair Classifier	0.7575 ± 0.0260	0.8291 ± 0.0229	0.2215 ± 0.1112	0.1444 ± 0.0810	-0.1052 ± 0.1315	0.8601 ± 0.1755
	Fair MixUp	0.6925 ± 0.0225	0.7837 ± 0.0208	0.0661 ± 0.0389	0.0465 ± 0.0252	-0.0461 ± 0.0446	0.9347 ± 0.0629
	FairGBM	0.7460 ± 0.0348	0.8255 ± 0.0283	0.1922 ± 0.0906	0.1345 ± 0.0756	-0.1539 ± 0.0773	0.8081 ± 0.0915
	FairVIC	$\mathbf{0.7250 \pm 0.0239}$	$\mathbf{0.8108 \pm 0.0237}$	$\mathbf{0.1443 \pm 0.0796}$	$\mathbf{0.1017 \pm 0.0464}$	$\mathbf{0.0016 \pm 0.0604}$	$\mathbf{1.0037 \pm 0.0764}$

MixUp, though initially promising and achieving second place after FairVIC in the COMPAS and German Credit datasets, fails to maintain its performance on the Adult Income dataset, where its results only just beat the baseline. In many cases, such as FairMixUp on the COMPAS and German Credit datasets, comparable techniques improve fairness but at the cost of accuracy, failing to achieve a balanced tradeoff.

Overall, FairVIC's ability to consistently balance the trade-off between fairness and accuracy, adapt to various datasets, and handle all fairness metrics comprehensively makes it the most effective method. Its consistent performance across different datasets, as evidenced by the lowest cumulative absolute error in performance and fairness, solidifies its superiority over other comparable methods.

6.2. Individual Fairness Analysis

To emphasise further FairVIC's ability to perform well across all fairness metrics, we also evaluate upon individual fairness by outputting the results of the counterfactual model, as described in Section 3.3. The full results, alongside the absolute difference in averages for each metric across the regular and counterfactual models, are seen in Table 5 in Appendix B.2.

The FairVIC model shows considerable promise in enhancing individual fairness across different datasets when compared with the baseline models. The counterfactual results from the FairVIC model with invariance term weighted heavily (FairVIC Invariance) exhibits lower absolute differences in metrics across all datasets. For example, in the German Credit dataset, the mean absolute difference across all six metrics between the regular and the counterfactual baseline model is ≈ 0.0277, while for FairVIC Invariance's regular and counterfactual models it is lower at ≈ 0.0108. This suggests a more stable and fair performance under counterfactual conditions. This capability highlights FairVIC's strength in not only addressing group fairness but also ensuring that individual decisions remain consistent and fair when hypothetical scenarios are considered. In the FairVIC model with recommended lambdas, we prioritise group fairness so invariance is weighted less. Even with this lower invariance weighting, FairVIC still achieved improved individual fairness.

6.3. Lambda Ablation Study Analysis

The FairVIC loss terms are combined with binary cross entropy for training the neural network to enable optimisation of both accuracy and fairness, minimising the tradeoff. The effect of FairVIC on the overall loss function can be increased and decreased by changing the weight λ for each FairVIC term. To evaluate this effect, we train a number of neural networks with the architecture described in Appendix A.2, with a different λ_{acc} weighting each time. In this experiment, we evaluate the effect of weighting the FairVIC loss terms equally, so that $\lambda_{\text{var}} = \lambda_{\text{inv}} = \lambda_{\text{cov}} = \frac{(1-\lambda_{\text{acc}})}{3}$, where $0 < \lambda_{\text{acc}} < 1$. The performance and fairness measures for each model are listed in Table 6 in Appendix C, and visualisations for the absolute difference in performance and fairness from ideal values for each run are

(a) Adult Income dataset.

(b) COMPAS dataset.

(c) German Credit dataset.

Figure 1. Absolute differences from the ideal value (e.g., perfect accuracy and fairness) in performance (left) and fairness (right) metrics of comparable techniques, sorted in ascending order on all three tabular datasets.

visualised in Figure 6 in Appendix C.

In Figure 6 (Appendix C), the trade-off between accuracy and fairness is evident. As λ_{acc} increases, predictive performance improves, but the fairness metrics deviate further from the ideal value. In contrast, when λ_{acc} is lower, fairness improves, but this time with only a negligible drop in accuracy. This suggests that lower λ_{acc} values provide a better overall performance balance. This trend is much more prevalent for the larger Adult dataset, where more complex relationships could lead to a larger accuracy-fairness trade-off. In the COMPAS and German Credit datasets, this trade-off, while still following the same pattern, is much smaller.

To evaluate upon the effect of each individual VIC term

Table 2. FairVIC and baseline comparison results of both performance and fairness for the CivilComments-WILDS and BiasBios datasets, including FairVIC's counterfactual (CF) model results and the absolute differences (ADs) between each model.

Dataset	Model	Accuracy	F1 Score	Equalized Odds	Absolute Odds	Statistical Parity	Disparate Impact
CivilComments-WILDS	Baseline	0.7624 ± 0.0055	0.7566 ± 0.0091	0.3095 ± 0.0297	0.1832 ± 0.0236	0.2639 ± 0.0212	1.9390 ± 0.1135
	Baseline CF	0.7608 ± 0.0031	0.7608 ± 0.0069	0.3104 ± 0.0296	0.1848 ± 0.0222	-0.2648 ± 0.0199	0.4940 ± 0.0400
	Baseline AD	0.0016 ± 0.0041	0.0041 ± 0.0081	0.0009 ± 0.0110	0.0016 ± 0.0091	0.5287 ± 0.0348	1.4449 ± 0.1360
	FairVIC	0.7243 ± 0.0755	0.6613 ± 0.1954	0.1457 ± 0.0661	0.1030 ± 0.0429	0.0562 ± 0.0517	1.1344 ± 0.1452
	FairVIC CF	0.6323 ± 0.1057	0.5722 ± 0.2128	0.1316 ± 0.0846	0.0953 ± 0.0819	0.0233 ± 0.1006	1.0687 ± 0.2324
	FairVIC AD	0.0921 ± 0.0907	0.0892 ± 0.2381	0.0141 ± 0.0711	0.0077 ± 0.0751	0.0329 ± 0.0648	0.0657 ± 0.1844
BiasBios	Baseline	0.8818 ± 0.0034	0.8811 ± 0.0044	0.0558 ± 0.0103	0.0456 ± 0.0083	-0.2489 ± 0.0098	0.6038 ± 0.0159
	Baesline CF	0.8794 ± 0.0041	0.8797 ± 0.0032	0.0563 ± 0.0153	0.0461 ± 0.0120	0.2481 ± 0.0119	1.6401 ± 0.0292
	Baseline AD	0.0041 ± 0.0029	0.0037 ± 0.0024	0.0123 ± 0.0093	0.0066 ± 0.0042	0.4970 ± 0.0206	1.0363 ± 0.0421
	FairVIC	0.8653 ± 0.0070	0.8646 ± 0.0059	0.0992 ± 0.0258	0.0830 ± 0.0147	-0.1217 ± 0.0158	0.7817 ± 0.0352
	FairVIC CF	0.8587 ± 0.0071	0.8580 ± 0.0074	0.1472 ± 0.0304	0.1193 ± 0.0186	0.0844 ± 0.0194	1.1890 ± 0.0568
	FairVIC AD	0.0066 ± 0.0066	0.0066 ± 0.0062	0.0480 ± 0.0392	0.0363 ± 0.0193	0.2061 ± 0.0276	0.4074 ± 0.0735

within the loss function, we can suppress the lambda terms from two out of three of variance, invariance, and covariance to leave only one remaining. We keep $\lambda_{\text{acc}} = 0.1$ since the previous lambda experiment showed this to be most effective and revealing in terms of the effect on fairness, while the chosen FairVIC loss term is assigned a weighting of 0.9. Similarly, we can also suppress a single term at a time, assigning two out of the three VIC terms a weighting of 0.45. The performance and fairness results for each experiment with different weightings are listed in Table 7.

It can be concluded that each term has a different effect. The variance term is shown to have the lowest standard deviation across all metrics and all tabular data in Table 7, offering stability to FairVIC. The covariance term makes the greatest contribution to group fairness, as seen in Table 7. The invariance term aims to give similar outputs to similar inputs, regardless of the protected attribute; therefore, it should have more of an effect towards individual fairness. Table 5 corroborates this hypothesis, as the FairVIC Invariance model (FairVIC with the invariance loss term weighted to 0.9, and accuracy loss of 0.1) consistently has a lower absolute difference than the baseline between the regular and counterfactual models across all metrics and tabular datasets, signalling greater individual fairness. Therefore, we conclude that the combination of all three terms would aim to improve both group and individual fairness, and increase stability.

6.4. Language Dataset Results

To show FairVIC's versatility across data modalities, our approach was applied to the CivilComments-WILDS and BiasBios datasets. The results are shown in Table 2, where FairVIC uses the lambdas $L_{\text{acc, var, inv}} = 0.1$, and $L_{\text{cov}} = 0.7$ for the CivilComments-WILDS dataset and $L_{\text{acc, var, inv}} = 0.05$, and $L_{\text{cov}} = 0.85$ for the BiasBios dataset.

From Table 2, the same trend can be seen as in the tabular dataset results, where FairVIC gives fairer results across all tested fairness metrics, the most notable being the improvement to disparate impact from 1.9390 to 1.1344 in the CivilComments-WILDS and 0.6038 to 0.7817 in the BiasBios datasets. In terms of individual fairness, the CivilComments-WILDS' baseline model has a mean absolute difference across every metric between the regular and counterfactual model of ≈ 0.3303 and ≈ 0.0503 for FairVIC. For the BiasBios dataset, the baseline model has a mean absolute difference across every metric between the regular and counterfactual model of ≈ 0.2563 and ≈ 0.1185 for our FairVIC model. Therefore, FairVIC is not confined to one modality, due to its ability to effectively reduce both individual and group fairness, while not seeing a drop of more than 3.81% and 1.87% in accuracy for the CivilComments-WILDS and BiasBios datasets respectively. The use of a different model architecture also proves FairVIC's adaptability to be utilised within different neural networks.

7. Conclusion and Future Work

In this paper, we introduced FairVIC, an in-processing bias mitigation technique that introduces three new terms into the loss function of a neural network- variance, invariance, and covariance. Across our experimental evaluation, FairVIC significantly improves scores for all fairness metrics, with minimal drop in accuracy, compared to previous comparable methods which typically aim to improve only upon a single metric. This balance showcases FairVIC's strength in providing a robust and effective solution applicable across various tasks and datasets. Future work would look to extend FairVIC to consider multiple protected characteristics simultaneously and expand its utility to image datasets.

8. Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

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

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

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Abeba Birhane. The unseen Black faces of AI algorithms. *Nature*, 610:451–452, 2022.

Toon Calders and Indrė Žliobaitė. *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, page 43–57. Springer, Berlin, Heidelberg, 2013. ISBN 978-3-642-30487-3. doi: 10.1007/978-3-642-30487-3_3. URL https://doi.org/10.1007/978-3-642-30487-3_3.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2020. URL https://api.semanticscholar.org/CorpusID:222208640.

Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE transactions on biometrics, behavior, and identity science*, 3(1):101–111, 2020.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

Silvia Chiappa and William S Isaac. A causal bayesian networks viewpoint on fairness. *Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data: 13th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Vienna, Austria, August 20-24, 2018, Revised Selected Papers 13*, pages 3–20, 2019.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021.

André Cruz, Catarina G Belém, João Bravo, Pedro Saleiro, and Pedro Bizarro. Fairgbm: Gradient boosting with fairness constraints. In *The Eleventh International Conference on Learning Representations*, 2023.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

Matthew Dixon, Diego Klabjan, and Jin Hoon Bang. Classification-based financial markets prediction using deep neural networks. *Algorithmic Finance*, 6(3-4):67–77, 2017.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

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

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

Taeuk Jang, Feng Zheng, and Xiaoqian Wang. Constructing a fair classifier with generated fair data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7908–7916, 2021.

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

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 10 2012. ISSN 0219-3116. doi: 10.1007/s10115-011-0463-8.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining*, pages 924–929. IEEE, 2012.

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

Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/c8dfece5cc68249206e4690fc4737a8d-Abstract.html.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

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

Alejandro Noriega-Campero, Michiel A. Bakker, Bernardo Garcia-Bulle, and Alex "Sandy" Pentland. Active fairness in algorithmic decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 77–83, New York, NY, USA, 1 2019. Association for Computing Machinery. ISBN 978-1-4503-6324-2. doi: 10.1145/3306618.3314277. URL https://dl.acm.org/doi/10.1145/3306618.3314277.

Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. *Advances in neural information processing systems*, 30, 2017.

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810, 2019.

Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. Adaptive sampling for minimax fair classification. In *Advances in Neural Information Processing Systems*, volume 34, page 24535–24544. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cd7c230fc5deb01ff5f7b1be1acef9cf-Abstract.html.

Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *Proceedings of the 36th International Conference on Machine Learning*, page 6373–6382. PMLR, 5 2019. URL https://proceedings.mlr.press/v97/ustun19a.html.

Pelin Vardarlier and Cem Zafer. Use of artificial intelligence as business strategy in recruitment process and social perspective. *Digital business strategies in blockchain ecosystems: Transformational design and future of global business*, pages 355–373, 2020.

Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018.

Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

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

A. Experiment Details

A.1. Dataset Metadata

Detailed metadata for each dataset, including our selection of privileged group, can be found in Table 3. Note that for the language datasets, the number of features is obtained by combining the protected characteristic and the toxicity label with the 50 tokenised text features. For BiasBios, we take architect, attorney, dentist, physician, professor, software engineer, surgeon as the favourable professions, and interior designer, journalist, model, nurse, poet, teacher, and yoga teacher as the unfavourable professions for our binary classification task.

Table 3. Metadata on all four experimental datasets.

Dataset	Adult Income	COMPAS	German Credit	CivilComments-WILDS	BiasBios
Data modality	Tabular	Tabular	Tabular	Text	Text
No. of Features	11	8	20	52	52
No. of Rows	48,842	5,278	1,000	50,000	50,000
Target Variable	income	two_year_recid	credit	toxicity	profession
Favourable Label	>50K (1)	False (0)	Good (1)	Non-Toxic (0)	Favourable (1)
Unfavourable Label	<=50K (0)	True (1)	Bad (0)	Toxic (1)	Unfavourable (0)
Protected Characteristic	sex	race	age	race	gender
Privileged Group	male (1)	Caucasian (1)	>25 (1)	white (1)	Male (0)
Unprivileged Group	female (0)	African-American (0)	<=25 (0)	non-white (0)	Female (1)

A.2. Neural Network Configuration

The configurations for the neural networks utilised for both the tabular and language data can be seen in Table 4. To obtain results, each model was run 10 times over random seeds, with a randomised train/test split each time. The averages and standard deviations were then outputted from across all 10 of the runs.

Table 4. Experimental model setup and parameters.

Parameter	Tabular Datasets	Language Datasets
Neural Network Architecture	Dense(128, 64, 32, 2, 32, 64, 128)	BiLSTM(64,32), Dense(64, 2, 64)
No. of Epochs	200	50
Batch Size	256	256
Optimiser	Adam	AdamW
Learning Rate	5e-2	5e-5
Dropout Rate	0.25	0.50
Regularisation	L1(1e-4)L2(1e-3)	L1(1e-4)L2(1e-3)

A visualisation for our neural network architecture for tabular data is seen in Figure 2, alongside our loss terms to illustrate where FairVIC components are applied.

All models were run with minimal and consistent data preprocessing. While some models, such as MetaFair, may underperform due to their reliance on specific sampling techniques, all comparable methods are treated uniformly as in-processing techniques. This allows them to be applied to any dataset, ensuring a fair evaluation across models.

B. Full Training Results

In addition to the results and analysis presented in Section 6, this section provides supplementary experiments and figures.

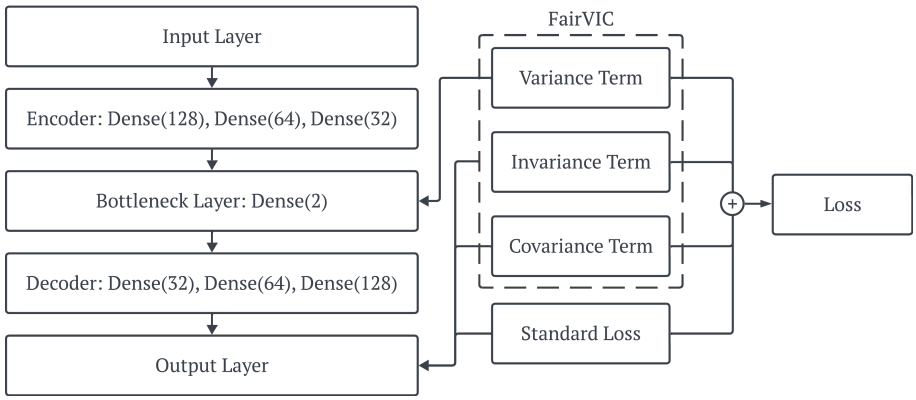


Figure 2. Network architecture for tabular data, with FairVIC loss components applied at relevant stages.

B.1. Feature Importances

Figure 3 shows the feature importance of the baseline and FairVIC models across the three tabular datasets. In all baseline models, the protected attributes show some importance to the decision-making process, such as in the COMPAS dataset, where *race* is a dominant feature. Combined with the results presented in Section 6.1, this suggests that the baseline models are prone to using the protected attribute to propagate bias. Additionally, proxy variables (highlighted with their importance in black), which are strongly correlated with the protected attributes, further show how bias can be perpetuated in the baseline model. For example, in the Adult Income dataset, *relationship* has a mean feature importance of 0.0124. This indicates that even though the model appears to have limited reliance on the protected attribute *sex* (which is among the least used features), it may still propagate bias through proxies such as *relationship*.

In contrast, the FairVIC models for all three datasets demonstrate a strong reduction in the mean importance of protected attributes and proxy variables. This reduction is due to the three additional terms used in FairVIC- variance, invariance, and covariance. We can see that the covariance term exactly minimises the model's dependency on the protected characteristic, which, in combination with results in Section 6.1, suggests a fairer decision-making process. The reduction in proxy variables should also be noted. Not only does FairVIC successfully reduce the reliance on the protected attribute, but it can also reduce the reliance on any features strongly correlated to the protected attribute. For example, in the Adult Income dataset, *sex* and *relationship* have a strong negative correlation (-0.58) meaning a model cannot only propagate bias through the use of *sex* but also through the use of *relationship* which we see the baseline model rely upon. The FairVIC model sees the mean feature importance of *relationship* drop by approximately a third and the importance of *sex* drop by half. This shows FairVIC's ability to mitigate both direct and indirect biases, leading to more equitable outcomes.

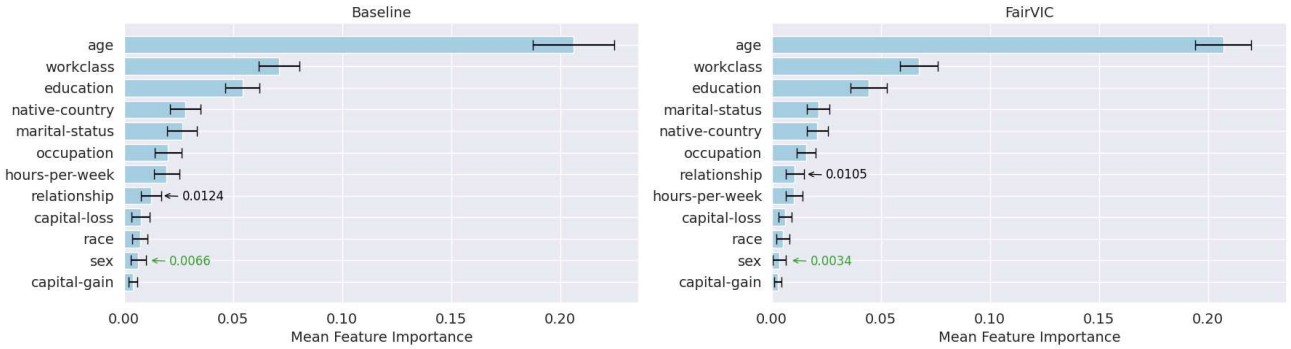
B.2. Individual Fairness Results

Following the analysis found in Section 6.2, Table 5 shows the individual fairness on both the baseline and FairVIC with our recommended lambdas, and FairVIC Invariance ($\lambda_{\text{acc}} = 0.1, \lambda_{\text{inv}} = 0.9, \lambda_{\text{var, cov}} = 0.0$) models using their absolute differences to their counterfactual model results. In the Adult Income dataset, the mean absolute difference across all six metrics combined for the baseline model is ≈ 0.0094, while for FairVIC invariance it is ≈ 0.0055. In the COMPAS dataset, the mean absolute difference for the baseline model is ≈ 0.0285, while for FairVIC Invariance it is ≈ 0.0050. Finally, for the German Credit dataset the mean absolute difference for the baseline model is ≈ 0.0277, while for FairVIC Invariance it is ≈ 0.0108. FairVIC's invariance term, designed to enhance individual fairness, proves to be effective. The FairVIC invariance model consistently achieves significantly absolute differences, demonstrating the success of the approach. In our selection of FairVIC terms, we prioritize group fairness by weighting invariance lower, yet the model still maintains low counterfactual absolute differences.

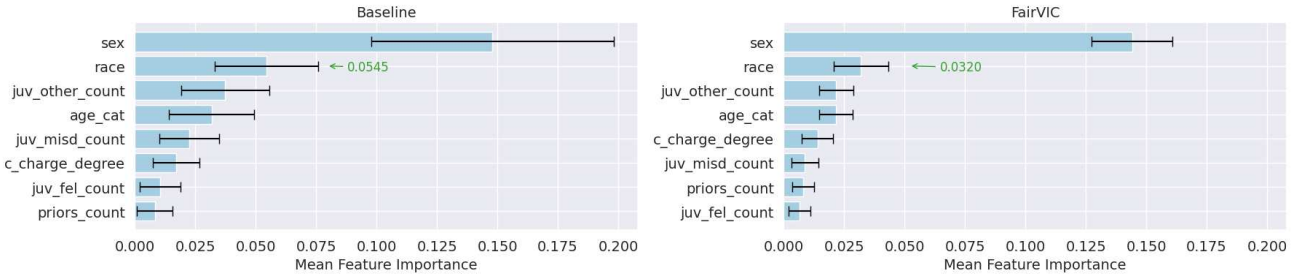
For discussion on the FairVIC Invariance model individual fairness results, see Section 6.2.

B.3. Hyperparameter Recommendations

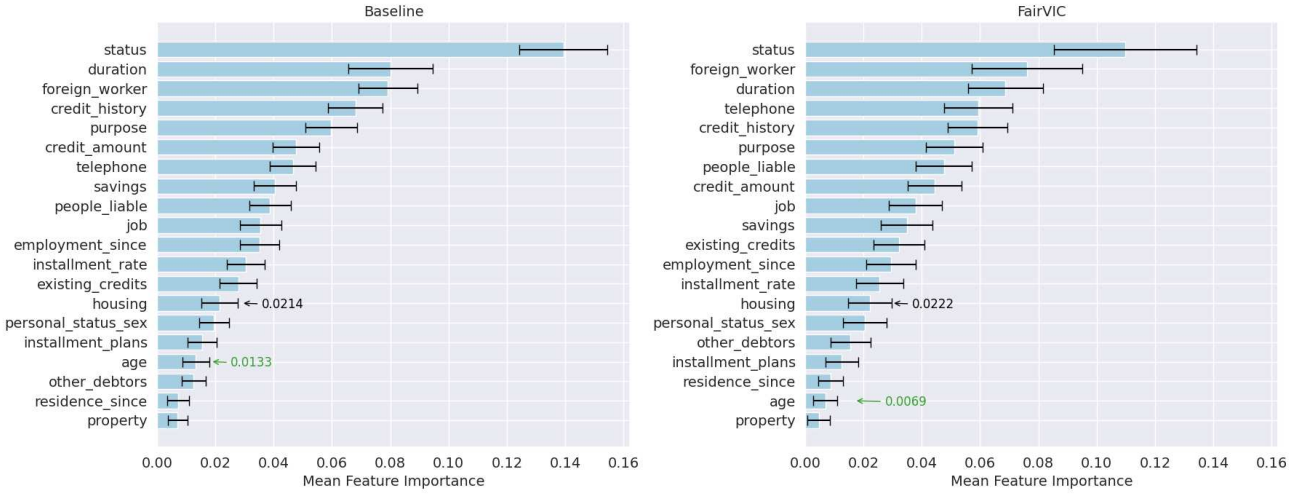
The weights for the loss terms in FairVIC ($\lambda_{\text{acc}}, \lambda_{\text{var}}, \lambda_{\text{inv}}, \lambda_{\text{cov}}$) were chosen based on insights from our ablation studies.



(a) Adult Income dataset.



(b) COMPAS dataset.



(c) German Credit dataset.

Figure 3. Mean feature importances for the baseline and FairVIC models across three tabular datasets. The protected attribute (green) and strong proxy variables to the protected attribute (black) are annotated for their exact feature importance.

For the COMPAS and German Credit datasets, the weights were set to $\lambda_{\text{acc, var, inv}} = 0.1$ and $\lambda_{\text{cov}} = 0.7$. The decision to use a relatively low weight for accuracy ($\lambda_{\text{acc}} = 0.1$) stems from the equal ablation study results, which demonstrated that this value achieves the best fairness-accuracy trade-off for these datasets. Group fairness is given significant emphasis, as shown by the higher weight assigned to the covariance term ($\lambda_{\text{cov}} = 0.7$), which plays a key role in minimizing disparities across protected groups. Meanwhile, the variance (λ_{var}) and invariance (λ_{inv}) terms were assigned a weight of 0.1, as this value still allowed for their individual fairness aims to be achieved effectively, thus balancing all fairness and accuracy objectives. Therefore, our default lambda recommendations would be $\lambda_{\text{acc}} = 0.1$, $\lambda_{\text{var}} = 0.1$, $\lambda_{\text{inv}} = 0.1$, and $\lambda_{\text{cov}} = 0.7$. These weights were also found to be effective for the CivilComments-Wilds dataset, and thus were utilised here as well.

For the Adult Income dataset, the weights were set to $\lambda_{\text{acc}} = 0.2$, $\lambda_{\text{var}} = 0.1$, $\lambda_{\text{inv}} = 0.1$, and $\lambda_{\text{cov}} = 0.6$. A slightly

Table 5. Counterfactual (CF) model results and absolute differences (ADs) for the baseline, FairVIC ($\lambda_{\text{acc, var, inv}} = 0.1$, $\lambda_{\text{cov}} = 0.7$), and FairVIC Invariance ($\lambda_{\text{acc}} = 0.1$, $\lambda_{\text{inv}} = 0.9$, $\lambda_{\text{var, cov}} = 0.0$) models.

Dataset	Model	Accuracy	F1 Score	Equalized Odds	Absolute Odds	Statistical Parity	Disparate Impact
Adult Income	Baseline	0.8444 ± 0.0065	0.6685 ± 0.0118	0.1330 ± 0.0317	0.1172 ± 0.0289	-0.2173 ± 0.0291	0.2853 ± 0.0329
	Baseline CF	0.8444 ± 0.0059	0.6649 ± 0.0114	0.1208 ± 0.0286	0.1026 ± 0.0285	-0.2069 ± 0.0290	0.3006 ± 0.0347
	Baseline AD	0.0000 ± 0.0032	0.0036 ± 0.0089	0.0123 ± 0.0329	0.0147 ± 0.0211	0.0104 ± 0.0147	0.0152 ± 0.0193
	FairVIC Invariance	0.8150 ± 0.0053	0.4281 ± 0.0938	0.0437 ± 0.0363	0.0342 ± 0.0330	-0.0811 ± 0.0516	0.3199 ± 0.0383
	FairVIC Invariance CF	0.8147 ± 0.0067	0.4242 ± 0.0954	0.0438 ± 0.0332	0.0303 ± 0.0286	-0.0752 ± 0.0479	0.3388 ± 0.0541
	FairVIC Invariance AD	0.0003 ± 0.0039	0.0039 ± 0.0472	0.0002 ± 0.0169	0.0040 ± 0.0145	0.0059 ± 0.0228	0.0189 ± 0.0445
	FairVIC	0.8284 ± 0.0088	0.5314 ± 0.0509	0.2993 ± 0.0683	0.1637 ± 0.0371	-0.0088 ± 0.0249	0.9803 ± 0.2220
	FairVIC CF	0.8310 ± 0.0075	0.5430 ± 0.0382	0.2793 ± 0.0563	0.1524 ± 0.0326	-0.0166 ± 0.0235	0.9015 ± 0.1450
	FairVIC AD	0.0007 ± 0.0055	0.0243 ± 0.0368	0.0240 ± 0.0397	0.0152 ± 0.0243	0.0022 ± 0.0130	0.0313 ± 0.1266
COMPAS	Baseline	0.6622 ± 0.0150	0.6118 ± 0.0252	0.3281 ± 0.0574	0.2635 ± 0.0452	-0.2941 ± 0.0459	0.6223 ± 0.0504
	Baseline CF	0.6651 ± 0.0183	0.6285 ± 0.0389	0.2707 ± 0.0599	0.2237 ± 0.0608	-0.2588 ± 0.0585	0.6415 ± 0.0763
	Baseline AD	0.0028 ± 0.0054	0.0167 ± 0.0190	0.0575 ± 0.0516	0.0398 ± 0.0334	0.0353 ± 0.0309	0.0192 ± 0.0329
	FairVIC Invariance	0.6571 ± 0.0121	0.6232 ± 0.0384	0.2618 ± 0.0412	0.2101 ± 0.0266	-0.2435 ± 0.0264	0.6530 ± 0.0551
	FairVIC Invariance CF	0.6564 ± 0.0109	0.6130 ± 0.0366	0.2689 ± 0.0341	0.2117 ± 0.0333	-0.2438 ± 0.0324	0.6633 ± 0.0642
	FairVIC Invariance AD	0.0007 ± 0.0072	0.0102 ± 0.0232	0.0071 ± 0.0332	0.0016 ± 0.0166	0.0003 ± 0.0159	0.0103 ± 0.0256
	FairVIC	0.6501 ± 0.0173	0.5934 ± 0.0357	0.0976 ± 0.0375	0.0719 ± 0.0305	-0.0602 ± 0.0678	0.9139 ± 0.1135
	FairVIC CF	0.6295 ± 0.0392	0.5154 ± 0.1767	0.0771 ± 0.0532	0.0506 ± 0.0353	-0.0394 ± 0.0609	0.9489 ± 0.1057
	FairVIC AD	0.0205 ± 0.0419	0.0780 ± 0.1874	0.0204 ± 0.0368	0.0213 ± 0.0336	0.0209 ± 0.0450	0.0350 ± 0.0723
German Credit	Baseline	0.7255 ± 0.0284	0.8077 ± 0.0275	0.2234 ± 0.0974	0.1641 ± 0.0936	-0.2218 ± 0.0901	0.7140 ± 0.1203
	Baseline CF	0.7010 ± 0.0371	0.7889 ± 0.0388	0.2222 ± 0.0979	0.1677 ± 0.0830	-0.1678 ± 0.1171	0.7782 ± 0.1495
	Baseline AD	0.0245 ± 0.0294	0.0189 ± 0.0257	0.0012 ± 0.0576	0.0036 ± 0.0454	0.0540 ± 0.0520	0.0641 ± 0.0700
	FairVIC Invariance	0.7165 ± 0.0356	0.7917 ± 0.0319	0.1367 ± 0.0798	0.0964 ± 0.0521	-0.0600 ± 0.1090	0.9113 ± 0.1625
	FairVIC Invariance CF	0.7250 ± 0.0356	0.7961 ± 0.0412	0.1257 ± 0.0793	0.0927 ± 0.0603	-0.0759 ± 0.0724	0.8902 ± 0.0972
	FairVIC Invariance AD	0.0085 ± 0.0272	0.0044 ± 0.0316	0.0109 ± 0.0843	0.0036 ± 0.0548	0.0159 ± 0.0689	0.0211 ± 0.0856
	FairVIC	0.7250 ± 0.0239	0.8108 ± 0.0237	0.1443 ± 0.0796	0.1017 ± 0.0464	0.0016 ± 0.0604	1.0037 ± 0.0764
	FairVIC CF	0.7380 ± 0.0223	0.8248 ± 0.0134	0.1466 ± 0.0943	0.1002 ± 0.0507	-0.0017 ± 0.0768	1.0000 ± 0.0979
	FairVIC AD	0.0130 ± 0.0198	0.0140 ± 0.0239	0.0023 ± 0.0292	0.0014 ± 0.0181	0.0033 ± 0.0584	0.0036 ± 0.0759

higher weight for accuracy ($\lambda_{\text{acc}} = 0.2$) was chosen compared to the COMPAS and German Credit datasets. This decision reflects the findings in Table 6, where a lower accuracy weight ($\lambda_{\text{acc}} = 0.1$) led to fairness metrics such as disparate impact exceeding one. While this reflects FairVIC's ability to actively address fairness concerns, the chosen weights create a careful balance between fairness and predictive performance. FairVIC leverages the increased accuracy weight to account for the complexity of the Adult dataset, characterised by its larger size, while maintaining strong fairness outcomes. The dominant weight assigned to the covariance term ($\lambda_{\text{cov}} = 0.6$) further ensures FairVIC prioritizes equitable outcomes across protected groups, achieving robust group fairness without compromising individual fairness or accuracy. The BiasBios dataset also require stronger interventions from the covariance term to improve group fairness, therefore we assign weights of $\lambda_{\text{acc, var, inv}} = 0.05$, and $\lambda_{\text{cov}} = 0.85$ to this dataset.

These weight configurations reflect the flexibility of FairVIC in balancing dataset-specific requirements for individual fairness, group fairness, and prediction accuracy. To effectively utilise FairVIC, we recommend users prioritise their specific fairness objectives—whether group fairness or individual fairness—based on the desired application context. If group fairness is the primary goal, assigning a higher weight to the covariance term (λ_{cov}) can help mitigate disparities across protected groups. Conversely, for tasks requiring equitable treatment at the individual level, increasing the weight of the invariance (λ_{inv}) term will enhance individual fairness. For users aiming to optimize specific performance metrics, such as accuracy or fairness, we could recommend conducting a grid search over the loss term weights ($\lambda_{\text{acc}}, \lambda_{\text{var}}, \lambda_{\text{inv}}, \lambda_{\text{cov}}$) to fine-tune the trade-offs for their dataset.

Additionally, users should consider the complexity and size of their dataset when configuring the weights. Larger datasets or those with greater feature diversity may require higher weights for accuracy (λ_{acc}) to maintain strong predictive performance. Fairness requirements may also vary depending on the level of societal or organisational impact, and we encourage users to carefully assess the implications of their choices in real-world deployments. Finally, we recommend evaluating FairVIC's performance using a range of fairness and accuracy metrics to ensure that the selected configuration aligns with the intended goals of the application.

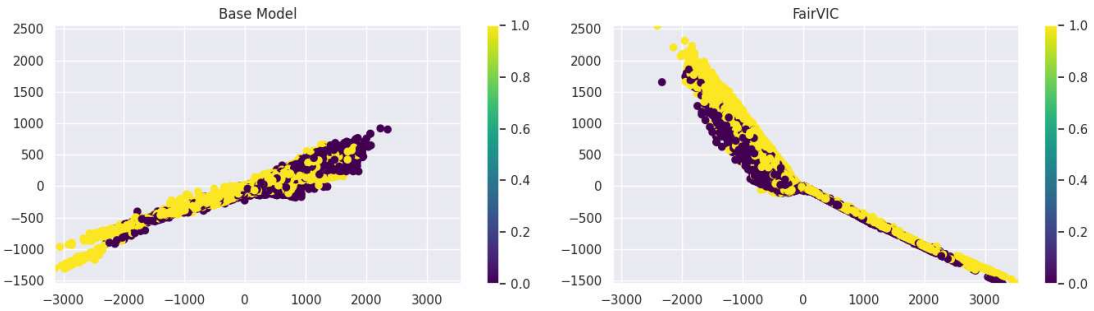


Figure 4. An example latent space visualization from one random seed of a baseline model and a FairVIC model on the Adult Income dataset. Subgroup (1) represents male individuals, and subgroup (0) represents female individuals.

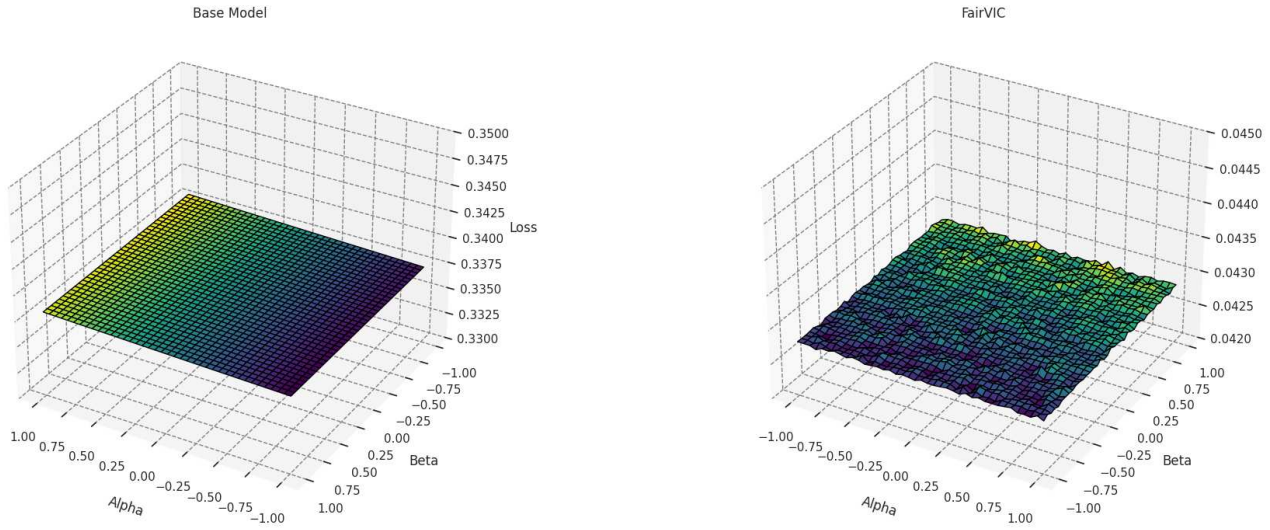


Figure 5. An example loss landscape visualisation from one random seed of a baseline model and a FairVIC model on the Adult Income dataset.

B.4. Model Representation Analysis

An example latent space visualisation from the baseline model and FairVIC can be seen in Figure 4. In the baseline model, we observe a separation between subgroups, where women (subgroup 0) are predominantly located in the upper region and men (subgroup 1) in the lower region of the latent space. This separation suggests that the baseline model's representations may be influenced by the protected attribute, leading to the biased decision-making reported in Table 1. In contrast, the FairVIC model shows a more condensed and overlapping distribution of both subgroups within the same latent space. This indicates, alongside results in Table 1 and feature importance in Figure 3a that FairVIC has successfully reduced the model's reliance on the protected characteristic and any proxy variables, thereby promoting more equitable representations. The overlapping and compact structure in the FairVIC latent space demonstrates that similar data points, regardless of their subgroup membership, are mapped closer together, ensuring that the model's predictions are not unfairly biased towards one group over the other.

B.5. Model Optimization Analysis

Figure 5 illustrates the loss landscapes of the baseline and FairVIC models on the Adult Income dataset. Both models exhibit smooth loss surfaces, indicating that they are relatively well-optimized. The baseline model (left) shows a stable loss landscape with a slight gradient. The FairVIC model (right), despite incorporating additional fairness constraints, maintains a similarly smooth surface albeit with tiny peaks in various places. This demonstrates that the inclusion of variance, invariance, and covariance terms in the loss function does not introduce instability or optimisation challenges.

B.6. Theoretical Analysis

In this section, we theoretically analyze FairVIC and show how each individual loss term is sub-differentiable.

B.6.1. THEOREM 1

Theorem 1. *Each individual term in FairVIC $L_{var}, L_{inv}, L_{cov}$ is sub-differentiable everywhere in the model's parameters θ.*

Proof. The variance term is defined as:

$$L_{\text{var}} = \frac{1}{N} \sum_{i=1}^{N} \max(0, \gamma - \sigma(z)) \tag{4}$$

where z is the latent embeddings for the input x. $z = g_\theta(x)$ is continuous in θ, where g_θ is the function/layer that maps input x to the latent embedding z. $\sigma(z)$ is the standard deviation of a continuous variable in a finite sample, which is continuous except at rare instances where all z_j are identical. Even in this degenerate case, $\sigma(\cdot)$ is sub-differentiable. The $\max(0, \cdot)$ operator is only non-differentiable at 0, where the sub-derivative set is $[0, 1]$. Hence $\max(0, \cdot)$ is sub-differentiable w.r.t θ. The invariance term is defined as:

$$L_{\text{inv}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - \hat{y}_i^*)^2 \tag{5}$$

where $\hat{y}_i$ is the model's predictions, and $\hat{y}_i^*$ is the model's predictions where the protected attribute is flipped. As $\hat{y}_i$ is differentiable in θ, then $(\hat{y}_i - \hat{y}_i^*)^2$ is differentiable as it is the composition of smooth functions. The covariance term is defined as:

$$L_{\text{cov}} = \frac{\sqrt{\sum_{i=1}^{N} \left((\hat{y}_i - \mathbb{E}[\hat{y}])^\top \cdot P_i \right)^2}}{N} \tag{6}$$

where $\sum_{i=1}^{N} \left((\hat{y} - \mathbb{E}[\hat{y}])^\top \cdot P \right)^2$ is a sum of squares, which is smooth and differentiable. The square root is differentiable for non-zero input and sub-differentiable at 0.

Each of the three terms is (sub-)differentiable everywhere in θ. Hence a gradient-based or subgradient-based method can be applied directly with FairVIC.

□

C. Lambda Ablation Study Results

Tables 6 and 7 show the full results for each model when the weights on the FairVIC terms are adapted. Table 6 shows the effect of changing λ_{acc} while keeping the FairVIC terms equal so that $\lambda_{\text{var, inv, cov}} = \frac{1-\lambda_{\text{acc}}}{3}$, where $0 < \lambda_{\text{acc}} < 1$, and Table 7 sets $\lambda_{\text{acc}} = 0.1$, and suppresses one or two FairVIC terms to explore the effect of only utilising one or two term(s) at a time. For full discussion and analysis of the results of the lambda ablation study, see Section 6.3.

Table 6. Performance and fairness results for FairVIC on the three tabular datasets, where the FairVIC terms are weighted equally, such that $\lambda_{\text{acc}} + \lambda_{\text{var}} + \lambda_{\text{inv}} + \lambda_{\text{cov}} = 1$.

Dataset	λ_{acc}	$\lambda_{\text{var,inv,cov}}$	Accuracy	F1 Score	Equalized Odds	Absolute Odds	Statistical Parity	Disparate Impact
Adult Income	0.10	0.30	0.8157 ± 0.0061	0.4622 ± 0.0412	0.3506 ± 0.0560	0.1929 ± 0.0328	0.0191 ± 0.0222	1.2542 ± 0.2653
	0.20	$0.2\overline{6}$	0.8358 ± 0.0084	0.5500 ± 0.0463	0.2321 ± 0.0784	0.1226 ± 0.0442	-0.0339 ± 0.0302	0.7970 ± 0.2118
	0.30	$0.2\overline{3}$	0.8448 ± 0.0053	0.6061 ± 0.0421	0.0918 ± 0.0507	0.0550 ± 0.0229	-0.0983 ± 0.0341	0.5073 ± 0.0952
	0.40	0.20	0.8481 ± 0.0033	0.6354 ± 0.0253	0.0560 ± 0.0137	0.0433 ± 0.0102	-0.1339 ± 0.0276	0.4069 ± 0.0434
	0.50	$0.1\overline{6}$	0.8506 ± 0.0052	0.6564 ± 0.0110	0.0630 ± 0.0117	0.0473 ± 0.0175	-0.1561 ± 0.0161	0.3686 ± 0.0301
COMPAS	0.10	0.30	0.6618 ± 0.0130	0.6061 ± 0.0197	0.1881 ± 0.0412	0.1451 ± 0.0317	-0.1754 ± 0.0324	0.7533 ± 0.0442
	0.20	$0.2\overline{6}$	0.6661 ± 0.0114	0.6661 ± 0.0114	0.2000 ± 0.0466	0.1448 ± 0.0339	-0.1805 ± 0.0306	0.7391 ± 0.0344
	0.30	$0.2\overline{3}$	0.6606 ± 0.0091	0.6162 ± 0.0266	0.1754 ± 0.0615	0.1326 ± 0.0485	-0.1687 ± 0.0426	0.7545 ± 0.0564
	0.40	0.20	0.6643 ± 0.0094	0.6162 ± 0.0202	0.1946 ± 0.0400	0.1433 ± 0.0345	-0.1797 ± 0.0313	0.7457 ± 0.0363
	0.50	$0.1\overline{6}$	0.6681 ± 0.0142	0.6239 ± 0.0192	0.2037 ± 0.0500	0.1654 ± 0.0527	-0.1988 ± 0.0526	0.7221 ± 0.0620
German Credit	0.10	0.30	0.7160 ± 0.0431	0.8059 ± 0.0351	0.1034 ± 0.0429	0.0704 ± 0.0300	-0.0298 ± 0.0612	0.9574 ± 0.0842
	0.20	$0.2\overline{6}$	-0.0298 ± 0.0612	0.8112 ± 0.0239	0.1305 ± 0.0754	0.0915 ± 0.0506	-0.0190 ± 0.0970	0.9791 ± 0.1282
	0.30	$0.2\overline{3}$	0.7205 ± 0.0286	0.8042 ± 0.0229	0.1189 ± 0.0593	0.0864 ± 0.0522	-0.0545 ± 0.0842	0.9305 ± 0.1154
	0.40	0.20	0.7265 ± 0.0270	0.8096 ± 0.0226	0.1222 ± 0.1240	0.0815 ± 0.0783	-0.0767 ± 0.0880	0.9004 ± 0.1189
	0.50	$0.1\overline{6}$	0.7175 ± 0.0211	0.8029 ± 0.0199	0.1073 ± 0.0549	0.0745 ± 0.0406	-0.0851 ± 0.0605	0.8866 ± 0.0839

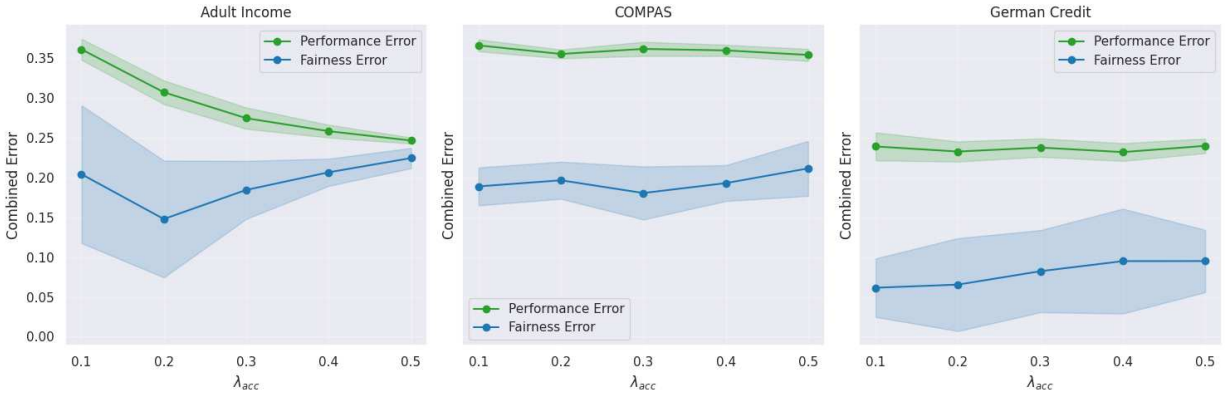


Figure 6. Absolute difference from the ideal value for performance (green) and fairness (blue) metrics of FairVIC with varying λ_{acc} values across all tabular datasets. The FairVIC terms are weighted equally, such that $\lambda_{\text{acc}} + \lambda_{\text{var}} + \lambda_{\text{inv}} + \lambda_{\text{cov}} = 1$.

Table 7. Performance and fairness results for FairVIC on the three tabular datasets, where only one or two FairVIC terms (λ_{var}, λ_{inv}, or λ_{cov}) are weighted at a time.

Dataset	λ_{acc}	λ_{var}	λ_{inv}	λ_{cov}	Accuracy	F1 Score	Equalized Odds	Absolute Odds	Statistical Parity	Disparate Impact
Adult Income	0.10	0.90	0.00	0.00	0.8481 ± 0.0038	0.6650 ± 0.0180	0.1018 ± 0.0270	0.0919 ± 0.0229	-0.1943 ± 0.0285	0.3079 ± 0.0370
	0.10	0.00	0.90	0.00	0.8150 ± 0.0053	0.4281 ± 0.0938	0.0437 ± 0.0363	0.0342 ± 0.0330	-0.0811 ± 0.0516	0.3199 ± 0.0383
	0.10	0.00	0.00	0.90	0.8106 ± 0.0109	0.4396 ± 0.0764	0.3535 ± 0.1011	0.1990 ± 0.0591	0.0282 ± 0.0291	1.4377 ± 0.6835
	0.10	0.45	0.45	0.00	0.8322 ± 0.0103	0.5382 ± 0.0747	0.0608 ± 0.0254	0.0481 ± 0.0210	-0.1112 ± 0.0332	0.3353 ± 0.0390
	0.10	0.45	0.00	0.45	0.8275 ± 0.0079	0.5535 ± 0.0524	0.2860 ± 0.0505	0.1597 ± 0.0294	-0.0126 ± 0.0239	0.9651 ± 0.2578
	0.10	0.00	0.45	0.45	0.8137 ± 0.0048	0.4614 ± 0.0504	0.3573 ± 0.0669	0.1975 ± 0.0380	0.0191 ± 0.0325	1.2790 ± 0.2316
COMPAS	0.10	0.90	0.00	0.00	0.6598 ± 0.0144	0.6250 ± 0.0283	0.2932 ± 0.1011	0.2490 ± 0.0746	-0.2834 ± 0.0733	0.6144 ± 0.0823
	0.10	0.00	0.90	0.00	0.6571 ± 0.0121	0.6232 ± 0.0384	0.2618 ± 0.0412	0.2101 ± 0.0266	-0.2435 ± 0.0264	0.6530 ± 0.0551
	0.10	0.00	0.00	0.90	0.6475 ± 0.0172	0.6018 ± 0.0405	0.0874 ± 0.0522	0.0606 ± 0.0427	-0.0146 ± 0.0686	1.0010 ± 0.1556
	0.10	0.45	0.45	0.00	0.6683 ± 0.0103	0.6424 ± 0.0156	0.2173 ± 0.0353	0.1809 ± 0.0239	-0.2223 ± 0.0223	0.6694 ± 0.0392
	0.10	0.45	0.00	0.45	0.6575 ± 0.0131	0.6147 ± 0.0280	0.1007 ± 0.0519	0.0730 ± 0.0471	-0.0540 ± 0.0795	0.9274 ± 0.1440
	0.10	0.00	0.45	0.45	0.6718 ± 0.0164	0.6358 ± 0.0326	0.2047 ± 0.0448	0.1635 ± 0.0404	-0.2018 ± 0.0454	0.7067 ± 0.0578
German Credit	0.10	0.90	0.00	0.00	0.7140 ± 0.0253	0.8011 ± 0.0233	0.1414 ± 0.0566	0.0951 ± 0.0412	-0.1049 ± 0.0412	0.8646 ± 0.0511
	0.10	0.00	0.90	0.00	0.7165 ± 0.0356	0.7917 ± 0.0319	0.1367 ± 0.0798	0.0964 ± 0.0521	-0.0600 ± 0.1090	0.9113 ± 0.1625
	0.10	0.00	0.00	0.90	0.7060 ± 0.0325	0.7911 ± 0.0358	0.1497 ± 0.0863	0.1005 ± 0.0599	0.0229 ± 0.0755	1.0225 ± 0.1128
	0.10	0.45	0.45	0.00	0.7485 ± 0.0436	0.8248 ± 0.0340	0.1675 ± 0.0866	0.1135 ± 0.0546	-0.1204 ± 0.0996	0.8469 ± 0.1271
	0.10	0.45	0.00	0.45	0.7110 ± 0.0311	0.7975 ± 0.0231	0.1219 ± 0.0620	0.0880 ± 0.0533	0.0175 ± 0.0635	1.0283 ± 0.0894
	0.10	0.00	0.45	0.45	0.7260 ± 0.0167	0.8091 ± 0.0145	0.1516 ± 0.0617	0.1190 ± 0.0481	-0.0739 ± 0.1093	0.9006 ± 0.1405