Learning Fairer Representations with FairVIC

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

[1]*Department of Computer Science, University of York, York, United Kingdom*

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 based on predefined criteria, FairVIC abstracts fairness to minimise dependency on protected attributes. 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 significantly compromising accuracy ($\approx -5\%$), thus offering a robust, generalisable solution for fair deep learning across diverse tasks and datasets.

Keywords

Machine Learning Bias Mitigation, Fairness In-Processing, Fairness–Accuracy Trade-off, Ethical Basis for Trustworthy AI

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 [1], finance [2], and recruitment [3]. 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 real-world harms. For instance, several studies have shown how bias in facial recognition technologies disproportionately misidentifies individuals of certain ethnic backgrounds [4, 5], leading to potential discrimination in law enforcement and hiring practices.

This highlights the urgent need to address AI bias. 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 [6], and only eliminate considered biases. Current in-processing methods frequently lead to unstable models and often rely upon arbitrary definitions of fairness [7]. 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 [8].

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.

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

[*] Corresponding author.

charmaine.barker@york.ac.uk (C. Barker); daniel.bethell@york.ac.uk (D. Bethell); dimitar.kazakov@york.ac.uk (D. Kazakov)

0000-0002-2337-043X (C. Barker); 0000-0002-0685-5312 (D. Bethell); 0000-0002-0637-8106 (D. Kazakov)

This function is designed to minimise the correlation between binary decisions and binary 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-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, text, and image;
- 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 [7]. 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 [9, 10], reweighing the data such that different groups influence the model in a representative way [11, 12], or generating synthetic data to balance out the representation of each group [13]. Another set of approaches utilises causal methods to delineate relationships between sensitive attributes and the target variables within the data [14, 15, 16]. 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 [6]. 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. [17] and Agarwal et al. [18] utilise a chosen fairness metric and perform constraint optimisation during training. This requires choosing a fairness metric, introducing human bias and limiting generalisation [7]. Therefore, fairness cannot be achieved across multiple definitions in this way [7]. Another approach involves incorporating an adversarial component during model training that penalises the model if protected characteristics can be predicted from its outputs [19, 20, 21]. These methods are often effective but their main shortcoming is seen in their instability [22]. 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 [23] or by incorporating fairness adjustments into the boosting process [24]. 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 [25]. Alternatively, labels near the decision boundary can be altered to favour less biased outcomes [26, 27]. Calibration [28, 29] adjusts the predictions of the model directly so that the proportion of positive instances is equal across each sub-group. These methods oversimplify fairness and ignore model-level bias. 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 [8], 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) [30] 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 has mostly been confined to self-supervised learning, with little exploration beyond. To this extent, FairVIC reworks the VIC components completely to serve its application in supervised learning for bias mitigation, a principle that has remained so far unexplored. 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 $FPR = \frac{FP}{FP+TN}$ [25]. 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 [31].

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 [32].

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 [15]. 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. This broad, generalised approach to defining bias 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* in their adapted forms: each is based on a classical statistical concept but used to express a specific fairness objective within the model. We consider binary classification tasks, where the model output $\hat{y} \in [0, 1]$ represents the predicted probability of the favourable class, obtained via a sigmoid activation.

Variance: This term promotes diversity in the latent representations by penalising low variance across features in the bottleneck embeddings, denoted $z \in \mathbb{R}^d$, where z is the output of the encoder for each input. It ensures the embeddings capture sufficient information, rather than collapsing to a trivial relationship such as the protected characteristic.

$$L_{\text{var}} = \frac{1}{d} \sum_{j=1}^{d} \max(0, \gamma - \sigma_j(z)) \tag{1}$$

where $\sigma_j(z)$ represents the standard deviation of the j-th feature across the batch, we set γ to 1.0 as a margin parameter to encourage unit-scale variability in the latent representation, and d is the number of features in the embedding.

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 penalise any contribution of the protected attribute to the decision process of the classifier, ensuring that predictions are not systematically skewed for the privileged group. By doing so, it promotes group fairness. The loss function is designed to minimise this covariance, as defined by the following equation:

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

where $\hat{y} \in (0, 1)$ is the model's softmax output, $P_i \in \{0, 1\}$ is the binary protected attribute with $P_i = 1$ denoting the privileged group, and N is the number of samples. In general, $\mathbb{E}[\hat{y}]$ reflects the empirical average of predictions across the batch, though it may approach 0.5 in a balanced and well-calibrated setting. During training, this loss reduces the deviation of privileged group predictions from the batch mean. In practice, it tends to soften overly confident predictions for the privileged group without necessarily changing the predicted class. This term does not directly affect accuracy, which is controlled by a separate loss, but its interaction with the accuracy objective discourages reliance on group membership in decision-making, thereby promoting group fairness.

Together, alongside a suitable accuracy loss, FairVIC jointly optimises for the total loss equation seen in Equation 4.

$$L_{\text{total}} = \lambda_{\text{acc}} L_{\text{acc}} + \lambda_{\text{var}} L_{\text{var}} + \lambda_{\text{inv}} L_{\text{inv}} + \lambda_{\text{cov}} L_{\text{cov}} \tag{4}$$

where each λ is a non-negative weighting coefficient balancing the contribution of its corresponding term, subject to the normalisation constraint $\lambda_{\text{acc}} + \lambda_{\text{var}} + \lambda_{\text{inv}} + \lambda_{\text{cov}} = 1$.

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 seven widely used bias mitigation benchmarks across tabular, text, and image modalities. These datasets contain known demographic disparities, allowing us to assess FairVIC's generalisability. For each, we designate one attribute as the protected characteristic on which fairness is to be improved.

Tabular datasets. We use Adult Income [33], COMPAS [34], and German Credit [35], all binary classification tasks with known biases. Adult Income predicts whether income exceeds \$50K and exhibits gender and racial bias. COMPAS predicts recidivism risk and is notorious for racial bias. German Credit classifies creditworthiness, with biases related to age and gender [36].

Language datasets. CivilComments-WILDS [37] and BiasBios [38] are used to assess FairVIC on text data. We sample 50,000 stratified examples per dataset to ensure balance. CivilComments classifies online comments as toxic or non-toxic, using race as the protected attribute. BiasBios consists of professional biographies classified into favourable vs. unfavourable occupations (e.g., surgeon vs. nurse), with gender as the protected feature.

Image datasets. CelebA [39] contains celebrity portraits labelled for attributes such as blond hair; we predict whether a person has blond hair or not, using gender as the protected attribute. UTKFace [40] includes facial images labelled with age and race; we predict whether a subject is above or below 30, with race as the protected attribute.

Detailed metadata for each dataset, including our selections for protected groups and classification goals, can be found in Appendix A.2.

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 [19].

[1] The code for our FairVIC implementation is provided at: https://github.com/CharmaineBarker/FairVIC.

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 [18].
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 [17].
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 [23].
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 [24].

A baseline neural network trained with binary cross-entropy was also implemented to reflect dataset biases. Model architecture and hyperparameters for both baseline and FairVIC are detailed in Appendix A.3.

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 3, 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.

Table 1
Accuracy and fairness results of FairVIC versus the baseline and five in-processing methods on three tabular datasets. **Bold** indicates the best overall trade-off between fairness and accuracy, reflecting our aim of holistic improvement (see Figure 3). Underlined values show the best score for each individual metric.

		Performance Metrics		Fairness Metrics			
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	**0.8284 ± 0.0088**	**0.5314 ± 0.0509**	**0.2993 ± 0.0683**	**0.1637 ± 0.0371**	**-0.0088 ± 0.0249**	**0.9803 ± 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	**0.6501 ± 0.0173**	**0.5934 ± 0.0357**	**0.0976 ± 0.0375**	**0.0719 ± 0.0305**	**-0.0602 ± 0.0678**	**0.9139 ± 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	**0.7250 ± 0.0239**	**0.8108 ± 0.0237**	**0.1443 ± 0.0796**	**0.1017 ± 0.0464**	**0.0016 ± 0.0604**	**1.0037 ± 0.0764**

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

[2] See Table 6, Appendix B.3 for our FairVIC lambdas, and discussion on these selections.

could misleadingly suggest that the model is fair, when in reality, the bias may only become evident when captured through a different perspective. Single-metric approaches like EGR often fail to address significant bias.

Overall, FairVIC outperforms all other comparable methods by demonstrating consistent improvements in both fairness and accuracy retention. As seen in Figure 1, 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 seen in Figure 3, Appendix B. This further highlights FairVIC's ability to generalise across datasets.

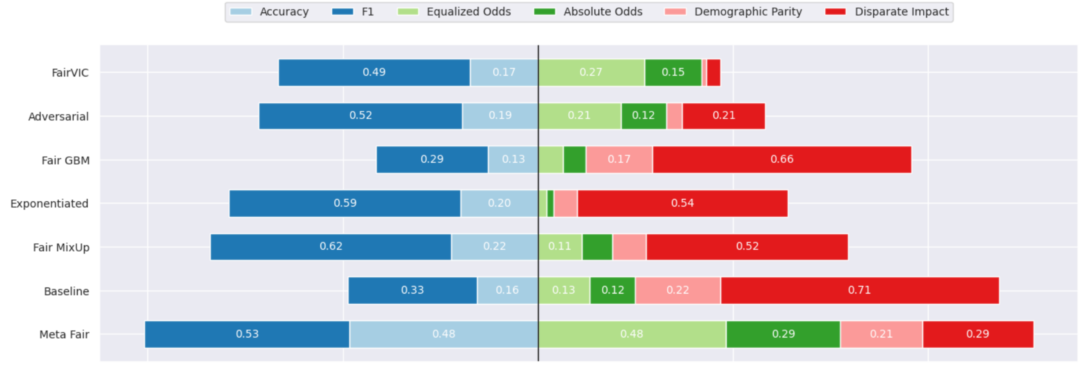


Figure 1: A qualitative analysis of the absolute differences from the ideal value (e.g., perfect accuracy and fairness) in performance (left) and fairness (right) metrics of comparable techniques on the Adult Income dataset.

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, FairMixUp, 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. Exponentiated Gradient Reduction (EGR) sees similar struggles. While performing the best in terms of each individual fairness metric in the COMPAS dataset, it does this at the expense of accuracy, where it sees a drop of 15.83% in accuracy. Further to this, EGR then performs poorly on the Adult Income dataset, only achieving a disparate impact of 0.4602, suggesting inability to generalise across datasets.

Following this paper's objective to create an approach that performs well across multiple fairness metrics without significantly compromising accuracy, we find that FairVIC demonstrates a consistent ability to balance fairness and accuracy across diverse datasets. Its adaptability, strong performance on all fairness metrics, and robustness to dataset shifts position it as the most effective method overall. This is further supported by its consistently low cumulative absolute error across both fairness and performance measures, highlighting its advantage over existing in-processing techniques.

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, 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.

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

		Performance Metrics		Fairness Metrics			
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
	Baseline 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	0.8668 ± 0.0070	0.8633 ± 0.0087	0.0970 ± 0.0167	0.0830 ± 0.0084	-0.1220 ± 0.0110	0.7744 ± 0.0207
	FairVIC CF	0.8601 ± 0.0069	0.8582 ± 0.0059	0.1464 ± 0.0222	0.1181 ± 0.0112	0.0862 ± 0.0117	1.1945 ± 0.0304
	FairVIC AD	0.0066 ± 0.0063	0.0051 ± 0.0070	0.0493 ± 0.0240	0.0350 ± 0.0146	0.2082 ± 0.0132	0.4201 ± 0.0295
CelebA	Baseline	0.8765 ± 0.0048	0.8776 ± 0.0057	0.2019 ± 0.0516	0.1268 ± 0.0249	-0.4672 ± 0.0124	0.4353 ± 0.0208
	Baseline CF	0.8747 ± 0.0063	0.8724 ± 0.0106	0.2286 ± 0.0495	0.1396 ± 0.0232	0.4676 ± 0.0096	2.2250 ± 0.1288
	Baseline AD	0.0018 ± 0.0055	0.0052 ± 0.0073	0.0267 ± 0.0510	0.0127 ± 0.0246	0.9348 ± 0.0176	1.7897 ± 0.1319
	FairVIC	0.7571 ± 0.0286	0.7811 ± 0.0249	0.2788 ± 0.0869	0.1543 ± 0.0349	-0.1018 ± 0.0815	0.8259 ± 0.1596
	FairVIC CF	0.7442 ± 0.0140	0.7804 ± 0.0180	0.3189 ± 0.0372	0.1686 ± 0.0148	0.0612 ± 0.0318	1.1956 ± 0.0909
	FairVIC AD	0.0129 ± 0.0142	0.0006 ± 0.0239	0.0401 ± 0.0584	0.0142 ± 0.0214	0.1630 ± 0.0958	0.3697 ± 0.1949
UTKFace	Baseline	0.6827 ± 0.0120	0.6839 ± 0.0225	0.0887 ± 0.0322	0.0631 ± 0.0260	0.1261 ± 0.0265	1.3136 ± 0.1024
	Baseline CF	0.6723 ± 0.0230	0.6655 ± 0.0767	0.0988 ± 0.0218	0.0741 ± 0.0224	-0.1328 ± 0.0233	0.7416 ± 0.0868
	Baseline AD	0.0104 ± 0.0198	0.0184 ± 0.0521	0.0100 ± 0.0106	0.0110 ± 0.0101	0.2589 ± 0.0450	0.5720 ± 0.1806
	FairVIC	0.6308 ± 0.0288	0.5901 ± 0.1011	0.0647 ± 0.0335	0.0509 ± 0.0313	-0.0035 ± 0.0332	1.0076 ± 0.0770
	FairVIC CF	0.6231 ± 0.0247	0.5713 ± 0.1176	0.0796 ± 0.0367	0.0645 ± 0.0325	0.0146 ± 0.0385	1.0097 ± 0.1332
	FairVIC AD	0.0077 ± 0.0250	0.0188 ± 0.1106	0.0149 ± 0.0317	0.0136 ± 0.0251	0.0181 ± 0.0334	0.0021 ± 0.0894

6.3. Language and Image Dataset Results

To demonstrate FairVIC's versatility, we apply it to CivilComments-WILDS and BiasBios. Results are shown in Table 2; our selected lambda configurations are given in Table 6. Across all fairness metrics, FairVIC improves upon the baseline, consistent with trends seen in the tabular datasets. Notably, Disparate Impact improves from 1.9390 to 1.1344 (CivilComments-WILDS), 0.6038 to 0.7817 (BiasBios), 0.4353 to 0.8259 (CelebA), and 1.3136 to 1.0076 (UTKFace). For individual fairness, the mean absolute difference between regular and counterfactual models is reduced substantially: from 0.3303 to 0.0503 on CivilComments-WILDS, 0.2563 to 0.1207 on BiasBios, 0.4618 to 0.1001 on CelebA, and 0.1468 to 0.0125 on UTKFace. These results confirm FairVIC's effectiveness across modalities and architectures, improving both group and individual fairness with minimal accuracy trade-off.

6.4. Lambda Ablation Study Analysis

Finally, to better understand the internal dynamics of FairVIC, we conduct an ablation study on the λ weightings of FairVIC's loss terms to analyse their individual contributions. 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 trade-off. 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.3, 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 7, Appendix C, and visualisations for the absolute difference in performance and fairness from ideal values for each run are visualised in Figure 7, Appendix C.

In Figure 7 (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 the effect of each individual VIC term 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 8.

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 8, offering stability to FairVIC. The covariance term makes the greatest contribution to group fairness, as seen in Table 8. 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.

Based on these findings, we can offer our recommendations for default lambda values to achieve effective results. We suggest beginning with a configuration of $\lambda_{\text{acc}} = 0.1$, $\lambda_{\text{var}} = 0.1$, $\lambda_{\text{inv}} = 0.1$, and $\lambda_{\text{cov}} = 0.7$, which consistently performs well across datasets. This configuration provides a strong default, with further tuning advised based on the relative importance of group fairness, individual fairness, or accuracy in a given application. A detailed discussion of dataset-specific configurations and practical guidance for adjusting these weights is provided in Appendix B.3.

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 will extend FairVIC to multiple protected attributes and targets.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118.

[2] M. Dixon, D. Klabjan, J. H. Bang, Classification-based financial markets prediction using deep neural networks, Algorithmic Finance 6 (2017) 67–77.

[3] P. Vardarlier, C. 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 (2020) 355–373.

[4] A. Birhane, The unseen Black faces of AI algorithms, Nature 610 (2022) 451–452.

[5] J. G. Cavazos, P. J. Phillips, C. D. Castillo, A. 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 (2020) 101–111.

[6] B. Salimi, L. Rodriguez, B. Howe, D. Suciu, Interventional fairness: Causal database repair for algorithmic fairness, in: Proceedings of the 2019 International Conference on Management of Data, 2019, pp. 793–810.

[7] S. Caton, C. Haas, Fairness in machine learning: A survey, ACM Computing Surveys (2020). URL: https://api.semanticscholar.org/CorpusID:222208640.

[8] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, A. Roth, A convex framework for fair regression, arXiv preprint arXiv:1706.02409 (2017).

[9] S. Shekhar, G. Fields, M. Ghavamzadeh, T. Javidi, Adaptive sampling for minimax fair classification, in: Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, p. 24535–24544. URL: https://proceedings.neurips.cc/paper/2021/hash/cd7c230fc5deb01ff5f7b1be1acef9cf-Abstract.html.

[10] B. Ustun, Y. Liu, D. Parkes, Fairness without harm: Decoupled classifiers with preference guarantees, in: Proceedings of the 36th International Conference on Machine Learning, PMLR, 2019, p. 6373–6382. URL: https://proceedings.mlr.press/v97/ustun19a.html.

[11] T. Calders, I. Žliobaitė, Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures, Springer, Berlin, Heidelberg, 2013, p. 43–57. URL: https://doi.org/10.1007/978-3-642-30487-3_3. doi:10.1007/978-3-642-30487-3_3.

[12] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and Information Systems 33 (2012) 1–33. doi:10.1007/s10115-011-0463-8.

[13] T. Jang, F. Zheng, X. Wang, Constructing a fair classifier with generated fair data, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 7908–7916.

[14] S. Chiappa, W. 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 (2019) 3–20.

[15] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, Advances in neural information processing systems 30 (2017).

[16] C. Russell, M. J. Kusner, J. Loftus, R. Silva, When worlds collide: integrating different counterfactual assumptions in fairness, Advances in neural information processing systems 30 (2017).

[17] L. E. Celis, L. Huang, V. Keswani, N. K. Vishnoi, Classification with fairness constraints: A meta-algorithm with provable guarantees, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 319–328.

[18] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, H. Wallach, A reductions approach to fair classification, in: International conference on machine learning, PMLR, 2018, pp. 60–69.

[19] B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 335–340.

[20] C. Wadsworth, F. Vera, C. Piech, Achieving fairness through adversarial learning: an application to recidivism prediction, arXiv preprint arXiv:1807.00199 (2018).

[21] D. Xu, Y. Wu, S. Yuan, L. Zhang, X. Wu, Achieving causal fairness through generative adversarial

networks, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019.

[22] X. Han, J. Chi, Y. Chen, Q. Wang, H. Zhao, N. Zou, X. Hu, Ffb: A fair fairness benchmark for in-processing group fairness methods, arXiv preprint arXiv:2306.09468 (2023).

[23] C.-Y. Chuang, Y. Mroueh, Fair mixup: Fairness via interpolation, in: International Conference on Learning Representations, 2021.

[24] A. Cruz, C. G. Belém, J. Bravo, P. Saleiro, P. Bizarro, Fairgbm: Gradient boosting with fairness constraints, in: The Eleventh International Conference on Learning Representations, 2023.

[25] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems 29 (2016).

[26] F. Kamiran, A. Karim, X. Zhang, Decision theory for discrimination-aware classification, in: 2012 IEEE 12th international conference on data mining, IEEE, 2012, pp. 924–929.

[27] F. Kamiran, S. Mansha, A. Karim, X. Zhang, Exploiting reject option in classification for social discrimination control, Information Sciences 425 (2018) 18–33.

[28] M. Kim, O. Reingold, G. 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.

[29] A. Noriega-Campero, M. A. Bakker, B. Garcia-Bulle, A. Pentland, Active fairness in algorithmic decision making, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 77–83. URL: https://dl.acm.org/doi/10.1145/3306618.3314277. doi:10.1145/3306618.3314277.

[30] A. Bardes, J. Ponce, Y. LeCun, VICReg: Variance-Invariance-Covariance Regularization for self-supervised learning, arXiv preprint arXiv:2105.04906 (2021).

[31] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd innovations in theoretical computer science conference, 2012, pp. 214–226.

[32] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 259–268.

[33] B. Becker, R. Kohavi, Adult, UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[34] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, in: Ethics of data and analytics, Auerbach Publications, 2022, pp. 254–264.

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

[36] F. Kamiran, T. Calders, Classifying without discriminating, in: 2009 2nd international conference on computer, control and communication, IEEE, 2009, pp. 1–6.

[37] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al., Wilds: A benchmark of in-the-wild distribution shifts, in: International conference on machine learning, PMLR, 2021, pp. 5637–5664.

[38] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. T. 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, 2019, pp. 120–128.

[39] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision (ICCV), 2015.

[40] Z. Zhang, Y. Song, H. Qi, Age progression/regression by conditional adversarial autoencoder, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5810–5818.

[41] J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores, arXiv preprint arXiv:1609.05807 (2016).

A. Experiment Details

A.1. FairVIC Training Algorithm

During training, the model iterates over epochs E, with the data shuffled into batches. For each batch, the model produces predictions $\hat{Y}$, which are compared with the true labels Y using a suitable accuracy loss function (e.g., binary cross-entropy, hinge loss, or Huber loss). The resulting loss is then minimised by an optimiser.

The pseudocode below summarises the full training procedure for FairVIC. It integrates the accuracy loss alongside the proposed fairness-promoting objectives- variance, invariance, and covariance- as defined in Section 4.1. The total loss is computed as a weighted combination of these terms. Subsequently, gradients are computed, and the optimiser adjusts the model parameters with respect to this combined loss.

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 $\{k \cdot \lambda_{\text{acc}}, k \cdot \lambda_{\text{var}}, k \cdot \lambda_{\text{inv}}, k \cdot \lambda_{\text{cov}}\},\ k \in \mathbb{R}$.

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 $M \leftarrow M - \alpha \nabla L_{\text{total}}$
```

A.2. 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 target label with the 50 tokenised text features and for image datasets, this is the pixels in an image plus the protected attribute and target label. 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 seven experimental datasets.

Dataset	Adult Income	COMPAS	German Credit	CivilComments-WILDS	BiasBios	CelebA	UTKFace
Data Modality	Tabular	Tabular	Tabular	Text	Text	Image	Image
No. of Features	11	8	20	52	52	128x128 +2	128x128 +2
No. of Rows	48,842	5,278	1,000	50,000	50,000	50,000	23,705
Target Variable	income	two_year_recid	credit	toxicity	profession	Blond_Hair	age
Favourable Label	>50K (1)	False (0)	Good (1)	Non-Toxic (0)	Favourable (1)	Not Blond (-1)	<30 (1)
Unfavourable Label	<=50K (0)	True (1)	Bad (0)	Toxic (1)	Unfavourable (0)	Blond (1)	>=30 (0)
Protected Characteristic	sex	race	age	race	gender	Male	race
Privileged Group	male (1)	Caucasian (1)	>25 (1)	white (1)	Male (0)	Male (1)	White (1)
Unprivileged Group	female (0)	African-American (0)	<=25 (0)	non-white (0)	Female (1)	Female (0)	Non-White (0)

A.3. Neural Network Configuration

The configurations for the neural networks utilised for the tabular, language, and image 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	Image Datasets
Network Architecture	Dense(128, 64, 32, 2, 32, 64, 128)	BiLSTM(64,32) + Dense(64, 2, 64)	Conv2D(32, 64)+ Dense(128, 64, 32, 2, 32, 64, 128)
No. of Epochs	200	50	50
Batch Size	256	256	256
Optimiser	Adam	AdamW	AdamW
Learning Rate	5e-2	5e-5	5e-5
Dropout Rate	0.25	0.50	0.25
Regularisation	L1(1e-4)L2(1e-3)	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.

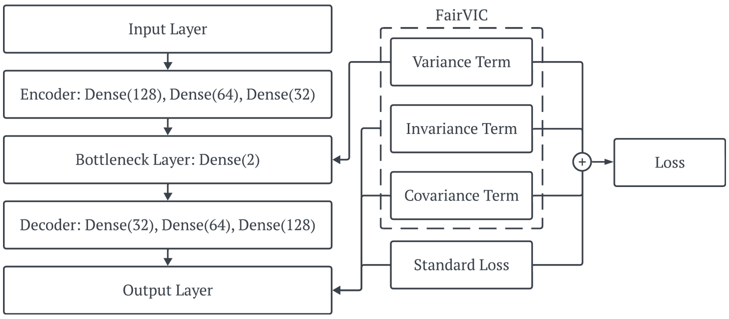
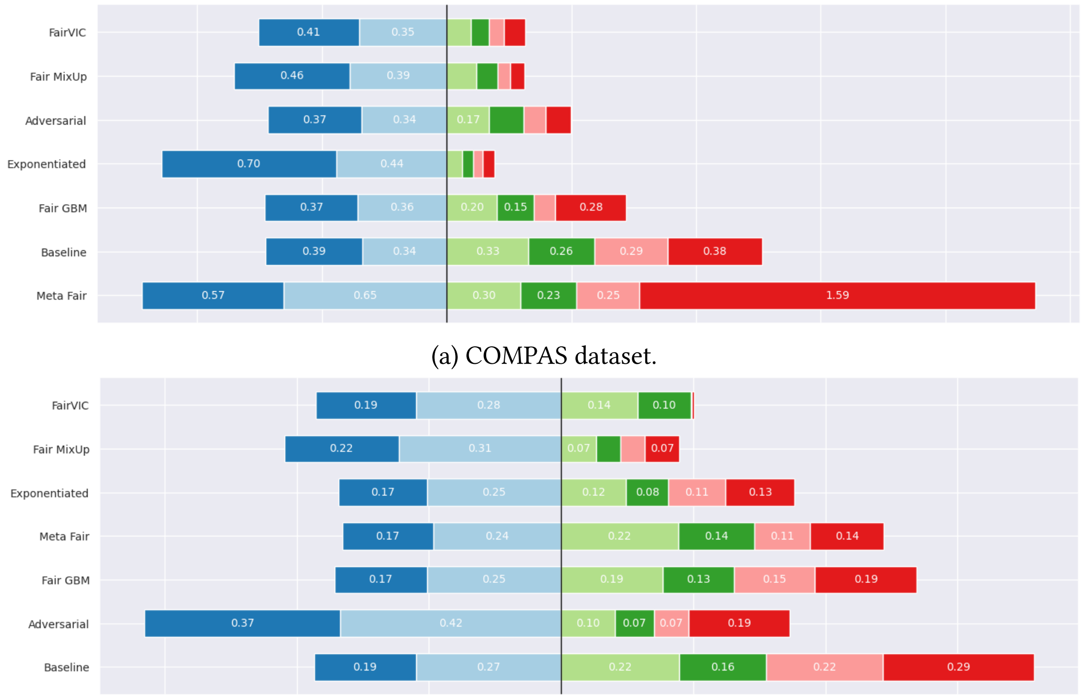

Input Layer
Encoder: Dense(128), Dense(64), Dense(32)
Bottleneck Layer: Dense(2)
Decoder: Dense(32), Dense(64), Dense(128)
Output Layer
FairVIC
Variance Term
Invariance Term
Covariance Term
Standard Loss
+
Loss

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

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. First, the qualitative visualisations for the COMPAS and German Credit can be seen in Figure 3, following the analysis of the Adult Income dataset in Figure 1. Discussion on these results can be seen in Section 6.1.



(a) COMPAS dataset.

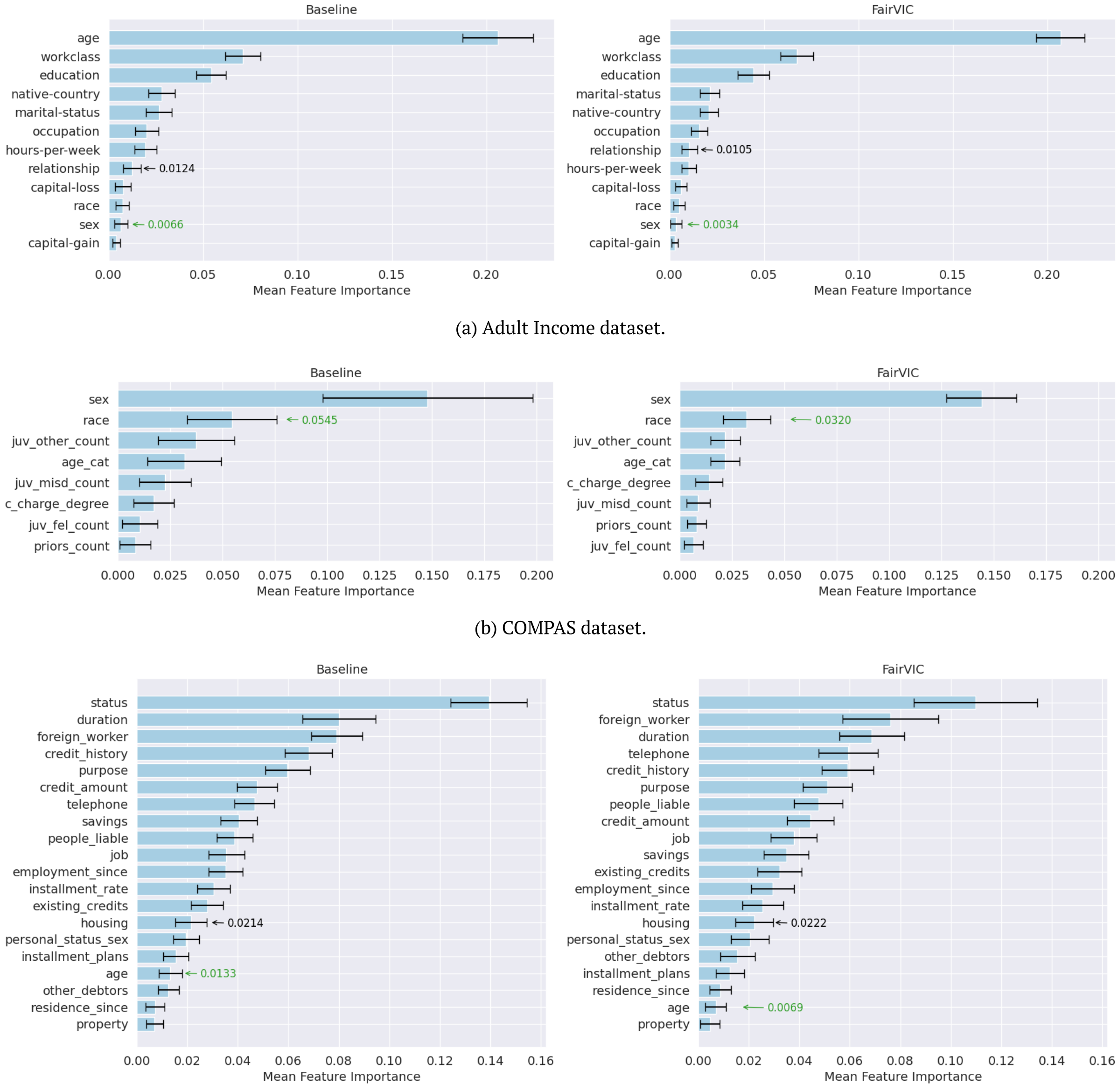
(b) German Credit dataset.

Figure 3: A qualitative analysis of the absolute differences from the ideal value (e.g., perfect accuracy and fairness) in performance (left) and fairness (right) metrics of comparable techniques on the COMPAS and German Credit datasets.

B.1. Feature Importances

Figure 4 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. On the COMPAS dataset, while *race* remains as the second most important feature, its actual importance dropped by $\approx 41\%$.



(a) Adult Income dataset.

(b) COMPAS dataset.

(c) German Credit dataset.

Figure 4: Mean feature importances derived from SHAP values for the baseline and FairVIC models across the three tabular datasets. The protected attribute (green) and its strong proxy variables (black) are annotated with their exact feature importance values.

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.

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.

		Performance Metrics		Fairness Metrics			
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

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. We have outlined the weights we used in our evaluation in Table 6.

To help users configure FairVIC effectively, we provide a recommended starting point and guidance for adapting the loss term weights to suit different fairness and performance goals. While our paper focuses on generalisable defaults, in a real-world deployment, one could envisage a tuning process to identify optimal results for a given application or domain.

We recommend starting with the configuration $\lambda_{\text{acc}} = 0.1$, $\lambda_{\text{var}} = 0.1$, $\lambda_{\text{inv}} = 0.1$, and $\lambda_{\text{cov}} = 0.7$, which provided strong results across multiple datasets including COMPAS, German Credit, and CivilComments-WILDS. This balanced setting encourages individual fairness, supports diverse representations, and strongly targets group fairness without significantly compromising accuracy. 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

Table 6
Our final lambda selection and recommendations for each of the seven datasets evaluated upon.

	Adult Income	Compas	German Credit	CivilComments- WILDS	BiasBios	CelebA	UTKFace
λ_{acc}	0.20	0.10	0.10	0.10	0.05	0.10	0.20
λ_{var}	0.10	0.10	0.10	0.10	0.05	0.05	0.10
λ_{inv}	0.10	0.10	0.10	0.10	0.05	0.05	0.10
λ_{cov}	0.60	0.70	0.70	0.70	0.85	0.80	0.60

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.

To adjust for different dataset characteristics or application goals, users should consider the intended role of each loss term:

If accuracy needs improvement while fairness is already strong, increasing the accuracy term weight can help the model prioritise predictive performance. For instance, in the Adult Income and UTKFace datasets, increasing λ_{acc} to 0.2 (while slightly lowering λ_{cov}) led to better trade-offs. This is likely because both datasets have relatively high-quality features and well-separated class distributions, allowing the model to benefit from a greater emphasis on discriminative capacity once basic fairness has been addressed.

When group fairness is a higher priority, the covariance term should be weighted more heavily. This term explicitly penalises statistical dependence between predictions and group membership, and is most impactful when datasets show strong baseline disparities between protected and privileged groups. For example, image-based datasets such as CelebA often contain visually encoded group cues that strongly correlate with target labels, while text-based datasets like BiasBios may exhibit linguistic patterns linked to social or demographic attributes. In such cases, increasing λ_{cov} (e.g. to 0.8 or 0.85) helps reduce spurious correlations and promotes more group-independent predictions.

To enhance individual fairness, a higher weight on the invariance term helps the model maintain consistency under counterfactual changes to the protected attribute. This is especially useful for datasets where protected features are not strongly entangled with the label, such that counterfactual consistency is a plausible fairness goal. Although our default already includes invariance, users with specific fairness requirements may find that increasing λ_{inv} offers more robust guarantees.

In practice, the ideal configuration depends on the dataset size, complexity, and the fairness-performance trade-off required by the application. Larger or more complex datasets may require a slightly higher weight on the accuracy term to avoid underfitting, while fairness-sensitive domains may justify higher weights on the fairness terms.

- **Default configuration:** $\lambda_{\text{acc}} = 0.1,\ \lambda_{\text{var}} = 0.1,\ \lambda_{\text{inv}} = 0.1,\ \lambda_{\text{cov}} = 0.7$
- **To improve accuracy:** Increase λ_{acc}
- **To emphasise group fairness:** Increase λ_{cov}
- **To emphasise individual fairness:** Increase λ_{inv}

If users have a specific optimisation goal—such as maximising a particular fairness or accuracy metric—a targeted grid search over the loss weights may be appropriate, though this falls outside the primary scope of our work.

B.4. Model Representation Analysis

An example latent space visualisation from the baseline model and FairVIC can be seen in Figure 5. 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, alongside results in Table 1 and feature importance in Figure 4, indicates 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.

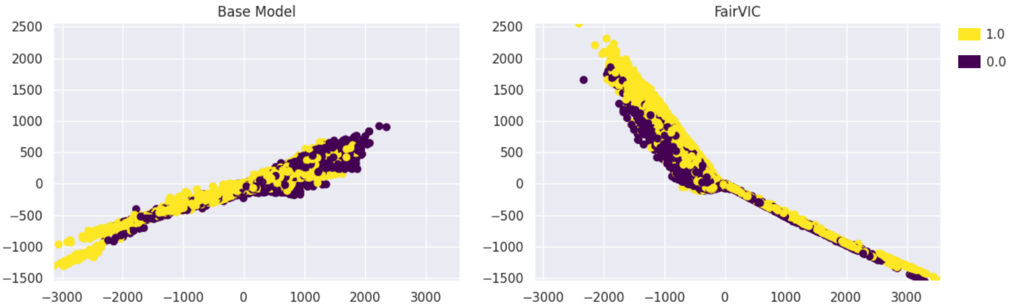


Figure 5: An example latent space visualization from one random seed of a baseline model and a FairVIC model on the Adult Income dataset. Subgroup (1) represents male individuals, and subgroup (0) represents female individuals.

B.5. Model Optimization Analysis

Figure 6 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.

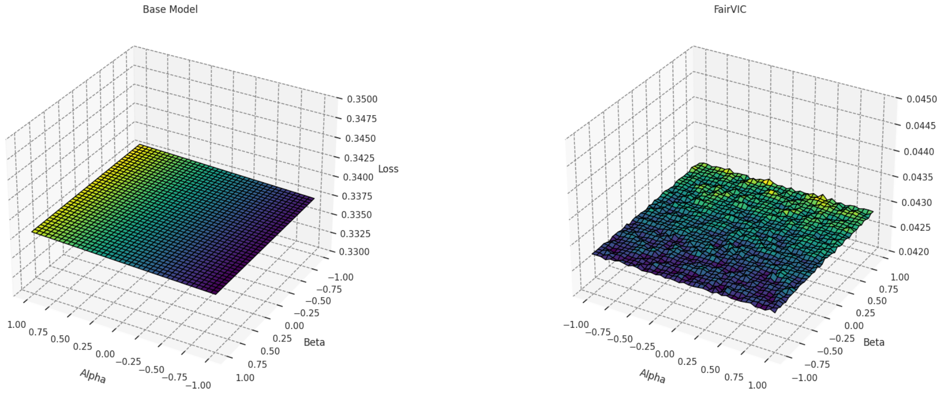


Figure 6: An example loss landscape visualisation from one random seed of a baseline model and a FairVIC model on the Adult Income dataset.

B.6. Theoretical Analysis and Discussion

In this section, we theoretically analyse FairVIC and show how each individual loss term is subdifferentiable.

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:

$$L_{\text{var}} = \frac{1}{d} \sum_{j=1}^{d} \max(0, \gamma - \sigma_j(z)) \tag{5}$$

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:

$$L_{\text{inv}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - \hat{y}_i^*)^2 \tag{6}$$

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:

$$L_{\text{cov}} = \frac{1}{N} \sqrt{\sum_{i=1}^{N} ((\hat{y}_i - \mathbb{E}[\hat{y}]) \cdot P_i)^2} \tag{7}$$

where $\sum_{i=1}^{N} ((\hat{y} - \mathbb{E}[\hat{y}]) \cdot P)^2$ 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.

□

B.7. Trade-Off Discussion

The design of FairVIC is grounded in the recognition that group fairness metrics such as Statistical Parity Difference, Equalized Odds Difference, and Disparate Impact capture distinct and often conflicting notions of fairness. Foundational theoretical work by Kleinberg et al. [41] has shown that, under realistic conditions- such as differing base rates between groups- it is impossible for any prediction system to simultaneously satisfy multiple fairness criteria, such as calibration and equalized odds, without introducing trade-offs.

FairVIC does not attempt to satisfy these fairness criteria simultaneously in a formal sense. Rather than optimising for any specific fairness definition directly, FairVIC introduces inductive biases at the representation level through variance, invariance, and covariance regularisation. These objectives are agnostic to downstream fairness metrics and aim to induce feature representations that are disentangled from protected attributes while preserving predictive signal.

The improvements observed across multiple fairness metrics are therefore not the result of directly encoding incompatible fairness constraints, but an emergent property of the learned representations. As shown in Table 1, FairVIC achieves significant reductions in Statistical Parity Difference and Equalized Odds Difference across all datasets, and improves Disparate Impact in most cases. In the case of our experiments, these fairness improvements come with modest reductions in predictive performance-consistent with expected trade-offs. Sometimes, improvements to one fairness metric such as Disparate

Table 7
Performance and fairness results for FairVIC on the three tabular datasets, where the FairVIC terms are weighted equally, such that $\lambda_{\text{acc}} + \lambda_{\text{var}} + \lambda_{\text{inv}} + \lambda_{\text{cov}} = 1$.

			Performance Metrics		Fairness Metrics			
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

Impact- slightly affect the results for another fairness metric, highlighting that such trade-offs are context-dependent and not inevitable.

Importantly, FairVIC does not commit to a normative fairness definition during training. This design choice reflects a practical and ethical consideration: in many real-world settings, there may be no consensus on which fairness metric best captures the relevant notion of harm or justice. By focusing on representation-level properties rather than metric-specific constraints, FairVIC supports post hoc evaluation across multiple fairness metrics, encouraging empirical pluralism and transparency in fairness assessments.

In summary, FairVIC does not circumvent impossibility theorems, nor does it attempt to. Rather, it sidesteps the need to encode conflicting fairness notions directly, and instead fosters representation learning that generalises well across subgroups. The observed metric improvements - and their trade-offs - are empirical outcomes of this strategy, not theoretical contradictions.

C. Lambda Ablation Study Results

Tables 7 and 8 show the full results for each model when the weights on the FairVIC terms are adapted. Table 7 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 8 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 6.4.

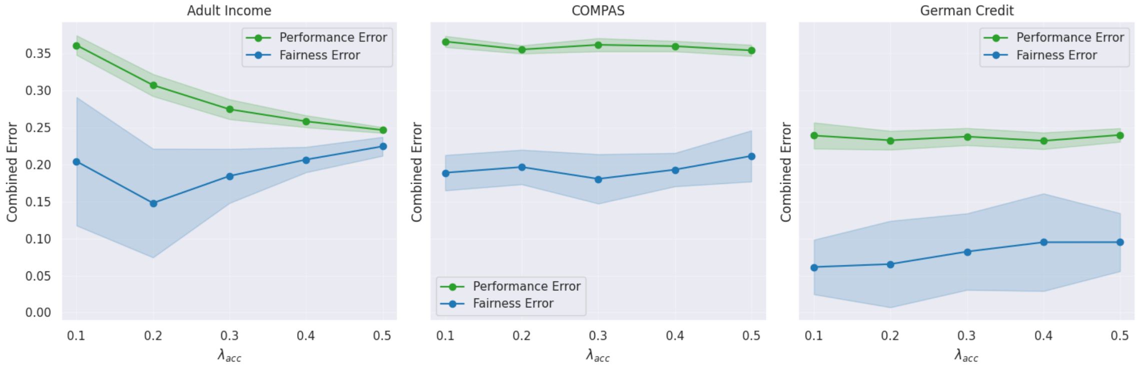


Figure 7: Absolute difference from the ideal value for performance (green) and fairness (blue) metrics of FairVIC with varying λ_{acc} values across all tabular datasets. The FairVIC terms are weighted equally, such that $\lambda_{\text{acc}} + \lambda_{\text{var}} + \lambda_{\text{inv}} + \lambda_{\text{cov}} = 1$.

Table 8
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.

					Performance Metrics		Fairness Metrics			
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