Mutual Information Regularization for Fairness-aware Deep Imaging Representations

Amir Reza Sadri, Thomas DeSilvio, and Satish E. Viswanath

Case Western Reserve University, Cleveland, OH 44106, USA
sev21@case.edu

Abstract. Fairness in medical imaging ML models is dependent on ensuring they are not impacted by sensitive attributes such as such as race and gender. Building on popularly considered in-processing fairness mitigation strategies, we present a novel approach to leveraging mutual information (MI) regularization to learn fairness-aware deep imaging representations. Based on analytical and theoretical justification, we develop a unique gradient-based mutual information penalty which bypasses the need for MI estimation within our Fairness-aware MI (FaMI) framework which avoids unstable approximations and scales effectively to large datasets. FaMI was implemented in conjunction with popular DenseNet and Vision Transformer architectures and evaluated against nine alternative fairness-aware alternatives as well as alternative MI estimators. Experiments on multi-institutional retinal OCT and rectal cancer MRI cohorts demonstrate that FaMI-ViT achieves the highest overall classification AUC (0.83 in distinguishing glaucoma vs non-glaucoma, 0.81 in distinguishing responders vs non-responders) while also improving fairness-related metrics across disparity subgroups, increasing EOM up to 0.84 and reducing EOdd by up to 0.85. These results highlight the potential of fairness-aware MI constraints in developing robust and equitable imaging-based ML models.

Keywords: Fairness · Mutual Information · Regularization

1 Introduction

Attributes such as demographics or ethnicity are known to be embedded in medical imaging data, and can pose significant challenges in ensuring machine learning (ML) model fairness, leading to biased predictions that compromise both the reliability and equity of ML models [1]. In clinical applications, such biases can have profound ethical implications, undermining the trustworthiness of deep learning (DL) models [2]. These *sensitive* attributes are known to be present in multiple modalities, including MRI [3] and OCT [4], resulting in challenges for achieving unbiased algorithmic performance across subgroups [5].

2 Previous Work and Novel Contributions

Addressing fairness challenges in medical imaging can broadly be categorized [5] into either *fairness evaluation* or *unfairness mitigation* strategies. Fairness

evaluation involves *post hoc* identification of subgroup disparities using metrics such as demographic parity and equalized odds, providing insights into the extent and nature of biases [5].

Unfairness mitigation aims to reduce bias using three main strategies: *pre-processing*, *in-processing*, and *post-processing* [5]. Pre-processing modifies data to balance subgroups or remove sensitive information [6,7], while post-processing adjusts model outputs, such as through threshold calibration [8]. In-processing, the most effective approach, integrates fairness constraints into training to learn subgroup-invariant representations without sacrificing performance [5,9]. Recent methods include adversarial learning [10], disentanglement [11], and contrastive learning [12], though they can introduce architectural complexity or training instability.

A popularly used fairness constraint is Mutual Information (MI), which is used to quantify and minimize the dependence between ML representations and sensitive attributes [13]. However, the most significant challenge with calculating MI in this context is that it requires knowledge of the underlying data distribution, which may be typically unknown with real-world imaging data [13, 14]. While deep learning has enabled the estimation of MI using variational bounds or neural estimators, these methods often rely on approximations that can be noisy and unstable, particularly in high-dimensional settings [15].

We propose a novel MI gradient penalty that avoids estimation, enhancing stability and enabling fairness-aware imaging without variational bounds or adversarial branches. Our Fairness-aware Mutual Information Regularization (FaMI) framework was implemented in two DL models—DenseNet [16] and Vision Transformer [17]—and evaluated on two clinical tasks ($n \approx 3500$): (i) classifying glaucoma vs. non-glaucoma from 3D OCT, focusing on racial disparities [4]; and (ii) predicting response to neoadjuvant therapy in rectal cancer from pre-treatment MRI, focusing on sex disparities.

3 Methodology

As illustrated in Fig. 1, consider a dataset with N data samples, each represented as a triplet in the form $\{(\mathbf{x}, u, y)\}$, where the input $\mathbf{x} \in \mathcal{X}$, $u \in \mathcal{U}$ is the sensitive attribute, and $y \in \mathcal{Y}$ is the target label. The model $f_\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by learnable parameters θ, is trained to predict the target label y, while we aim to improve its fairness by mitigating bias. The model produces a predicted output given by $\mathbf{v}(\theta) = f_\theta(\mathbf{x})$. The goal is to ensure that no information about u is used to predict y; in other words, we aim for:

Definition 1 (Fair Prediction Rule [18]). *A prediction* $\mathbf{v}$ *is fair with respect to the sensitive attribute* u *if and only if* $u \perp \mathbf{v}$*, meaning* u *is independent of the learned representation* $\mathbf{v}$*.*

To achieve this independence, we incorporate a penalty term into the model loss function that minimizes the mutual information (MI) between u and $\mathbf{v}$. Reducing MI—ideally to zero—promotes independence between the sensitive

attribute and the learned representation. This objective function can be defined as:

$$L(\theta) \;=\; L_{\mathrm{CE}}(\theta) + \mu\, L_{\mathrm{MI}}(\theta) = L_{\mathrm{CE}}(\theta) + \mu\, I\big(u; \mathbf{v}(\theta)\big) \tag{1}$$

where $L_{\mathrm{CE}}(\theta)$ is the cross-entropy loss, $\mu \geq 0$ is a regularization parameter, and $L_{\mathrm{MI}}(\theta)$ is the mutual-information penalty term. Here, calculating MI directly is notoriously challenging, as it depends on knowing the underlying data distributions [19]. Instead, Equation (1) can be minimized by descending its stochastic gradient. The gradient of $L_{\mathrm{CE}}(\theta)$ can be computed by standard backpropagation, whereas $L_{\mathrm{MI}}(\theta)$ will utilize the chain rule:

$$\nabla_\theta L_{\mathrm{MI}}(\theta) = \mathbf{J}_{\mathbf{v}}^T \times \nabla_{\mathbf{v}} I(u; \mathbf{v}) \quad \mathbf{J}_{\mathbf{v}} = \nabla_\theta \mathbf{v}(\theta) \tag{2}$$

where $\mathbf{J}_{\mathbf{v}}$ is the Jacobian of $\mathbf{v}$ with respect to θ, which can be obtained via automatic differentiation. By the Mutual Information Difference theorem [20],

$$\nabla_{\mathbf{v}} I(u; \mathbf{v}) = \mathbb{E}\big\{\boldsymbol{\lambda}(u, \mathbf{v})\big\}, \quad \boldsymbol{\lambda}(u, \mathbf{v}) = \nabla_{\mathbf{v}} \ln p(u|\mathbf{v}) \tag{3}$$

where $\boldsymbol{\lambda}(u, \mathbf{v})$ is called *score function difference (SFD)*. The SFD serves as a critical gradient for mutual information minimization, facilitating model training by adjusting $\mathbf{v}$ to achieve independence from u while preserving task accuracy [20]. Theorem 1 demonstrates that even though backpropagation is not used on the MI term directly, it will still converge under a suitable learning-rate schedule.

Theorem 1. *If the Jacobian of model predictions is bounded, i.e.,* $\|\mathbf{J}_{\mathbf{v}}\| \leq M$*, and the cross-entropy loss converges during backpropagation, then the sequence* $\{\theta_k\}$ *in* $\theta_{k+1} = \theta_k - \eta\Big(\nabla_\theta L_{\mathrm{CE}}(\theta_k) + \mu \nabla_\theta L_{\mathrm{MI}}(\theta_k)\Big)$ *converges to a stationary point of* $L(\theta)$*. Convergence is ensured if the learning rate* η *satisfies* $\eta \geq \frac{1}{\mu M^2}$*.*

Proof. At each iteration k, let $\mathbf{v}_k$ denote the model representation and define the intermediate update $\mathbf{v}_{k+1} = \mathbf{v}_k + \mathbf{h}_k$, where $\mathbf{h}_k$ is chosen to reduce $I(u; \mathbf{v}_k)$, minimizing dependency on sensitive attributes. A schematic of this concept is shown in Fig. 1 (left panel). Setting $\mathbf{h}_k = -\boldsymbol{\lambda}(u, \mathbf{v}_k)$, we have:

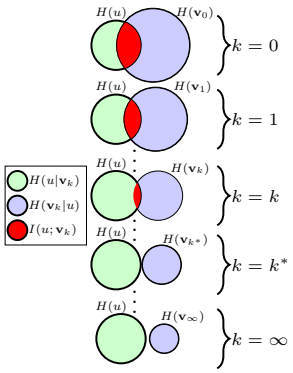
$$I(u; \mathbf{v}_{k+1}) - I(u; \mathbf{v}_k) \leq -\mathbb{E}\big\|\boldsymbol{\lambda}(u, \mathbf{v}_k)\big\|^2. \tag{4}$$

Thus, $\{I(u; \mathbf{v}_k)\}$ is monotonically decreasing.

Summing (4) gives $\mathbb{E}\left\{\sum_{k=0}^{\infty} \|\mathbf{h}_k\|^2\right\} \leq I(u; \mathbf{v}_0)$. Since $I(u; \mathbf{v}_\infty) = 0$, it follows that $\mathbf{h}_k \to 0$, $\mathbf{v}_{k+1} \to \mathbf{v}_k$, and $\{\theta_k\}$ converges. Given $\nabla_{\mathbf{v}} L_{\mathrm{MI}}(\theta_k) = \mathbf{v}_{k+1} - \mathbf{v}_k$, we obtain $\nabla_\theta L_{\mathrm{MI}}(\theta_k) \to 0$. Furthermore:

$$\begin{aligned} \|\nabla_{\mathbf{v}} L_{\mathrm{MI}}(\theta_k)\| = \|\mathbf{v}_{k+1} - \mathbf{v}_k\| &\leq M \,\|\theta_{k+1} - \theta_k\| \\ &\leq \eta M \,\|\nabla_\theta L_{\mathrm{CE}}(\theta_k)\| + \eta\mu M \,\|\nabla_\theta L_{\mathrm{MI}}(\theta_k)\| \\ &\leq \eta\mu M^2 \,\|\nabla_{\mathbf{v}} L_{\mathrm{MI}}(\theta_k)\| \end{aligned} \tag{5}$$

which completes the proof.



Algorithm 1 FaMI Optimization

Input: $\{(\mathbf{x}_i, u_i, y_i)\}_{i=1}^N$, μ, η ▷ (dataset with N samples)
Initialization: θ_0 ▷ (parameters)
for $k = 0, 1, 2, \cdots$ (until convergence) **do**
  Sample a mini-batch $b \subset \{1, \cdots, N\}$
  for $i \in b$ **do**
    $\mathbf{v}_i = f_{\theta_k}(\mathbf{x}_i)$ ▷ (model output)
    $\mathbf{p}_i = \text{softmax}(\mathbf{v}_i)$ ▷ (model prediction)
    $\boldsymbol{\lambda}(u_i, \mathbf{v}_i)$ ▷ (estimate SFD)
  end for
  $\nabla_\theta L_{\text{CE}} = \frac{1}{|b|} \sum_i \nabla_\theta \text{CE}(y_i, \mathbf{p}_i)$ ▷ (cross-entropy gradient)
  $\nabla_\theta \mathbf{v}_i = \mathbf{J}_{\mathbf{v}_i}$ ▷ (Jacobian)
  $\nabla_\theta L_{\text{MI}} = \frac{1}{|b|} \sum_i \mathbf{J}_{\mathbf{v}_i}^T \boldsymbol{\lambda}(u_i, \mathbf{v}_i)$ ▷ (MI gradient)
  $\theta_{k+1} = \theta_k - \eta(\nabla_\theta L_{\text{CE}} + \mu \nabla_\theta L_{\text{MI}})$ ▷ (parameters update)
end for
Output: θ

Fig. 1. Illustration of the proposed FaMI optimization framework. Left panel is the Venn diagram illustrating gradual modification of model outputs to minimize MI with respect to the sensitive attribute ($k = 1$ to $k = k^*$).

This sequence can be solved in a minibatch setting with stochastic gradient descent via Algorithm 1, to ensure $\mathbf{v}(\theta)$ is independent of u.

SFD Estimation: The gradient of the conditional probability can be estimated directly through polynomial kernel estimation [20, 21], eliminating the need for precise knowledge of the underlying probability density function. Unlike approaches that approximate the joint probability or employ empirical Bayes smoothing and density estimation [13,22], this method leverages polynomial kernels to provide a robust and efficient framework for score function estimation.

4 Experimental Design

4.1 Data Description

For all experiments, cohorts segregated into train (70% per class) and test sets, ensuring the class distribution was preserved (see Table 1).

Table 1. Multi-institutional cohorts with distribution of sensitive attributes. Class and sensitive attribute (e.g., race and gender) counts are reported independently and may overlap within each cohort.

Split	$\mathbf{C_1}$ (Eye Diseases, OCT)					$\mathbf{C_2}$ (Rectal Cancer, MRI)			
	Classes		*Race*			*Classes*		*Gender*	
	Glaucoma	Non-glaucoma	Asian	Black	White	pCR	Non-pCR	Male	Female
Train	1083	1017	700	700	700	20	110	95	35
Test	665	535	400	400	400	11	45	34	22

C_1 (Eye Diseases, OCT) comprises a publicly available cohort of 3,300 retinal nerve fiber layer thickness maps and 3D OCT B-scan images. This dataset, curated for fairness learning, included balanced racial groups (Asian, Black, or White) as the sensitive attribute, with glaucoma and non-glaucoma as the outcome classes. For further details, see [4].
C_2 (Rectal Cancer, pre-CRT MRI) comprised 186 pre-treatment T2w MRIs from rectal cancer patients, retrospectively collected from four institutions prior to neoadjuvant therapy and surgery. The objective was to predict pathologic complete response (pCR) to therapy, defined as histopathologic tumor regression grade (TRG) 0 (ypTRG0 or 0% viable tumor cells) [23]. Sex (male or female) was designated as the sensitive attribute. Prior to analysis, MRI scans were resampled to an isotropic voxel resolution of $1 \times 1 \times 1$mm using trilinear interpolation. Rectal tumors on each scan were manually annotated by expert radiologists, ensuring precise delineation of relevant structures. Bounding boxes were generated around the annotated regions using connected component analysis, excluding regions smaller than 20 pixels to avoid artifacts and irrelevant components. These bounding boxes were standardized to dimensions of $79 \times 79 \times 40$ voxels, reflecting the mean bounding box size across cohorts and was used in all experiments.

4.2 Model Implementation

Execution: To evaluate the impact of fairness regularization, two FaMI variants were developed using the loss function defined in (1), where the MI penalty term was weighted by $\mu = 0.5$ (selected based on prior work and further validated in the current study):
• **FaMI-DN:** Based on DenseNet-121, leveraging densely connected convolutional layers to enhance feature propagation, encourage parameter efficiency, and reduce redundancy. Transition layers were included for downsampling, followed by global average pooling, and a softmax layer for classification.
• **FaMI-ViT:** The Vision Transformer architecture splits images into patches, which are processed by transformer layers employing multi-head self-attention and positional encoding. A classification token was added and passed through a fully connected output layer to complete the architecture.
• **Baselines:** Non-fairness aware DenseNet-121 (DN) and ViT were included.
• **Fairness-aware Alternatives:** The following state-of-the-art fairness-aware approaches were considered: FIN [4] which normalizes logits per identity group to balance feature importance, FairBatch [24] which dynamically adjusts minibatch compositions to enforce fairness constraints during training, and FairViT [25] which modifies ViT with adaptive masking and distance loss to ensure equitable attention. FIN and FairBatch were used in conjunction with either model, while FairViT is a ViT-specific variant.
• **Fairness-aware MI-Based Alternatives:** To evaluate the advantages of our MI estimation approach, we implemented two alternative methods: Mutual Information Neural Estimator (MINE) [15], which estimates mutual information through gradient descent, and Mutual Information Gradient Estimator (MIGE) [13], which directly estimates MI gradients. Both MINE and MIGE are

general approaches applicable to any architecture. We adapted them for fairness optimization in DN (MIGE-DN, MINE-DN) and ViT (MIGE-ViT, MINE-ViT).

In total 11 models were evaluated using the same train and test cohorts (Table 1). All models were implemented in 3D, with parameters optimized using a ten-fold cross-test strategy on the training cohort using 150 epochs with a batch size of 64, a learning rate of $1e-4$, a cross-entropy loss function, and an SGD optimizer with a weight decay of $1e-2$. Gradient clipping with a maximum norm of 1.0 was applied to stabilize training and prevent exploding gradients. While SFD estimation may be dependent on batch size and embedding dimension [21, 26], stable performance was observed in our experiments. All models were implemented in Python 3.9.18 and executed on a high-performance desktop computer featuring an Intel(R) Core(TM) i5-8279U CPU (2.40 GHz), 16 GB RAM, 64-bit Windows 11 Pro (version 23H2), and an NVIDIA GeForce RTX 3080 GPU with 10 GB GDDR6X memory.

4.3 Model Evaluation and Statistical Analysis

The performance of all eleven models (DN, ViT, FairBatch, FIN, FairViT, MINE-DN, MINE-ViT, MIGE-DN, MIGE-ViT, FaMI-DN, and FaMI-ViT) was evaluated on both the train and holdout test cohorts. Metrics included (i) area under the ROC curve (AUC) to measure overall classification performance, (ii) Equality of Opportunity across Multiple Subclasses (EOM) as a fairness metric assessing the balance of true positive rates across subgroups [27], and (iii) Equalized Odds (EOdd) which evaluates disparities in both true positive and false positive rates across sensitive subgroups [5]. Pairwise Wilcoxon testing, corrected using the Bonferroni method ($p = 0.005$), was utilized to determine significant differences in model performance between approaches. To compare the effectiveness of our MI gradient computation against alternative approaches, loss curves were plotted for MI-based models with fluctuations quantified using Total Variation (TV) [28] and Root Mean Square Error (RMSE) [29], providing a numerical measure of stability and smoothness in the optimization process.

5 Results and Discussion

5.1 FaMI-based DL models compared to fairness-aware alternatives

Fig. 2 presents the classification AUCs for five fairness-aware models based on the vision transformer architecture: ViT, FairBatch-ViT, FIN-ViT, FairViT, and FaMI-ViT. Across both C_1 and C_2, FaMI-ViT can be seen to consistently achieve the highest overall AUC, with statistically significant improvements over other models ($p < 0.005$). FaMI-ViT exhibits significantly higher EOM and lower EOdd compared to FairBatch-ViT and FairViT, demonstrating its superior ability to mitigate fairness disparities while maintaining high classification performance. Overall, FaMI-ViT reduces EOdd by over 0.6 compared to alternative

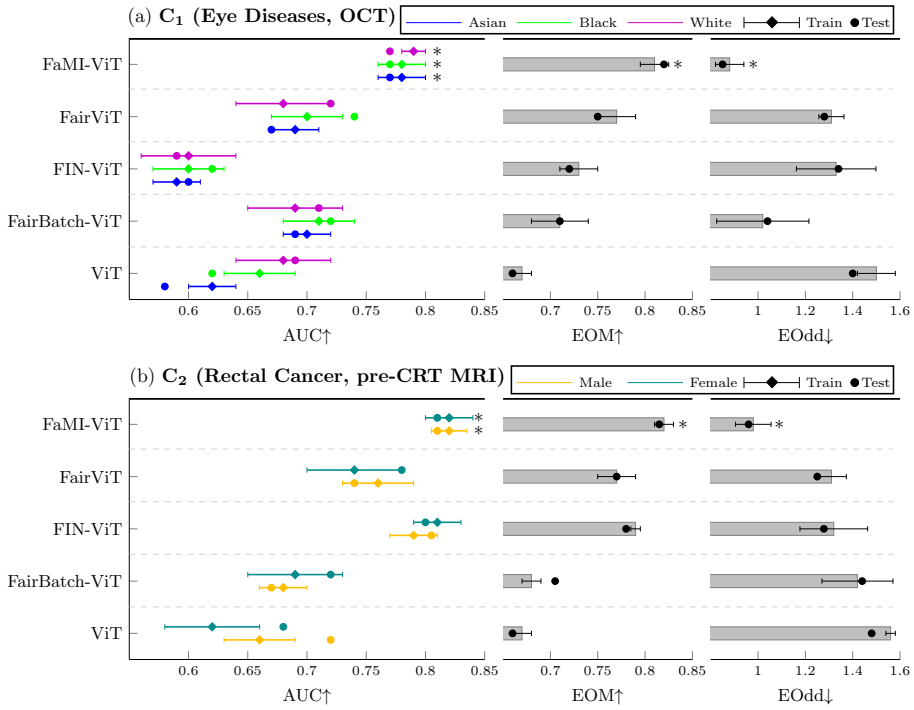


Fig. 2. Forest and bar plots of AUC, EOM, and EOdd for ViT models when evaluated on (a) C_1 and (b) C_2 datasets. In each plot, the left panel presents the forest plot of AUC results (subgroups in different colors), while the right panel displays the overall AUC, EOM, and EOdd as bar plots. Results are presented for both train and test sets. Arrows next to the metrics indicate whether a higher or lower value is considered better. Statistical significance (* $p < 0.005$) is indicated for comparisons of FaMI-ViT against alternative classifiers using pairwise Wilcoxon testing.

ViT-based fairness models and 0.5 versus the baseline ViT model, suggesting it can effectively minimize error rate disparities across groups.

Table 2 summarizes the overall performance of FaMI-DN compared to alternative DenseNet-based methods. FaMI-DN can be seen to achieve the highest AUC among DN-based models (0.83 in C_1, 0.82 in C_2), in addition to significantly improving fairness metrics of EOM and EOdd. Results for FaMI-ViT have been included in Table 2 to highlight differences compared to FaMI-DN (but which were not statistically significant).

5.2 Comparison of FaMI to alternative fairness-aware MI estimators

Fig. 3 depicts a line plot of the MI penalty term loss during training for each of FaMI, MIGE, and MINE within C_1 and C_2, when using the DN model. FaMI can be seen to achieve a more stable and smooth convergence, as well as consistently

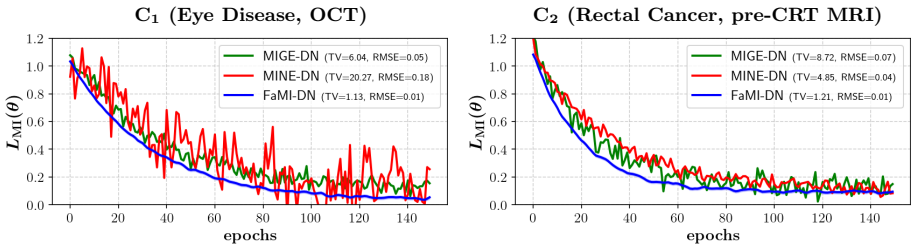


Fig. 3. Line plots visualizing smoother convergence of MI penalty term loss for FaMI-DN (blue) compared to MIGE-DN (green), and MINE-DN (red) across C_1 and C_2.

Table 2. AUC, EOM, and EOdd performance in distinguishing between patient groupings in each of C_1 and C_2, with the best model in bold. * indicates $p < 0.005$ in pairwise Wilcoxon ranksum testing between the best FaMI model and the closest alternative.

Approach	C_1 (Eye Diseases, OCT)						C_2 (Rectal Cancer, MRI)					
	Train			*Test*			*Train*			*Test*		
	AUC	EOM	EOdd	AUC	EOM	EOdd	AUC	EOM	EOdd	AUC	EOM	EOdd
DN	0.70 ± 0.02	0.60 ± 0.04	1.70 ± 0.08	0.68	0.59	1.65	0.71 ± 0.02	0.61 ± 0.04	1.68 ± 0.08	0.69	0.60	1.66
FairBatch-DN	0.72 ± 0.02	0.71 ± 0.03	1.02 ± 0.2	0.70	0.71	1.04	0.73 ± 0.02	0.68 ± 0.01	1.42 ± 0.15	0.71	0.70	1.44
FIN-DN	0.79 ± 0.02	0.73 ± 0.02	1.33 ± 0.17	0.81	0.72	1.34	0.79 ± 0.01	0.79 ± 0.05	1.32 ± 0.14	0.79	0.78	1.27
FaMI-DN	$\mathbf{0.83 \pm 0.01}^*$	0.80 ± 0.01	0.92 ± 0.02	0.79	0.81	0.97	$\mathbf{0.83 \pm 0.01}^*$	0.81 ± 0.01	0.95 ± 0.02	0.80	0.79	1.01
FaMI-ViT	0.82 ± 0.01	$\mathbf{0.81 \pm 0.01}^*$	$\mathbf{0.88 \pm 0.06}^*$	**0.81**	**0.82**	**0.85**	0.82 ± 0.01	$\mathbf{0.82 \pm 0.01}^*$	$\mathbf{0.98 \pm 0.07}^*$	**0.81**	**0.84**	**0.96**
MINE-DN	0.74 ± 0.02	0.72 ± 0.03	1.28 ± 0.05	0.72	0.71	1.30	0.75 ± 0.02	0.73 ± 0.03	1.27 ± 0.05	0.73	0.72	1.28
MINE-ViT	0.75 ± 0.02	0.73 ± 0.03	1.25 ± 0.05	0.74	0.72	1.24	0.75 ± 0.02	0.73 ± 0.03	1.25 ± 0.05	0.74	0.72	1.24
MIGE-DN	0.78 ± 0.02	0.78 ± 0.02	1.10 ± 0.03	0.76	0.77	1.12	0.78 ± 0.02	0.79 ± 0.02	1.09 ± 0.03	0.77	0.78	1.11
MIGE-ViT	0.79 ± 0.02	0.79 ± 0.02	1.08 ± 0.03	0.78	0.78	1.10	0.79 ± 0.02	0.80 ± 0.02	1.07 ± 0.03	0.78	0.79	1.09

exhibiting the lowest TV and RMSE, with TV values of 1.13 (C_1) and 1.21 (C_2) and RMSE of 0.01 for both datasets. By contrast, MINE-DN resulted in the highest TV (20.27 C_1, 4.85 C_2) as well as RMSE (0.18 C_1, 0.04 C_2), reflecting much more marked fluctuations. Our FaMI regulization and gradient optimization thus reduces MI constraints more effectively to ensure a more reliable model training process.

6 Concluding Remarks

In this work, we introduced a novel approach to leveraging mutual information regularization to learn fairness-aware deep imaging representations. We provide rigorous analytical justification to develop a gradient-based fairness-aware MI (FaMI) penalty term to effectively mitigate biases while preserving classification accuracy. Validation of FaMI variants for the popular DenseNet and Vision Transformer architectures demonstrated significant improvements in fairness and performance metrics compared to existing methods, including across racial and sex-based subgroups. Future work will optimize efficiency, extend to multi-class tasks, and assess generalization to unseen subgroups.

Acknowledgments. Research reported in this publication was supported by the National Cancer Institute (1R01CA280981-01A1), the National Institute of Nursing Research (1R01NR019585-01A1), the National Heart, Lung, and Blood Institute (1R01HL165218-01A1), the National Science Foundation (Award #2320952), the Veterans Affairs Biomedical Laboratory Research and Development Service (1I01BX006439-01), the DOD Peer Reviewed Cancer Research Program (W81XWH-21-1-0725), the Ohio Third Frontier Technology Validation Fund, the JobsOhio Program, and the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering at Case Western Reserve University. This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the U.S. Department of Veterans Affairs, the Department of Defense, or the United States Government.

Disclosure of Interests

The authors declare that they have no competing interests.

References

1. Yang, Y., Zhang, H., Gichoya, J.W., Katabi, D., Ghassemi, M.: The limits of fair medical imaging ai in real-world generalization. Nature Medicine 30(10), 2838–2848 (2024)
2. Sarridis, I., Koutlis, C., Papadopoulos, S., Diou, C.: Flac: Fairness-aware representation learning by suppressing attribute-class associations. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
3. Stanley, E.A., Wilms, M., Mouches, P., Forkert, N.D.: Fairness-related performance and explainability effects in deep learning models for brain image analysis. Journal of Medical Imaging 9(6), 061102–061102 (2022)
4. Luo, Y., Tian, Y., Shi, M., Pasquale, L.R., Shen, L.Q., Zebardast, N., Elze, T., Wang, M.: Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization. IEEE Transactions on Medical Imaging (2024)
5. Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M., Zhou, S.K.: Addressing fairness issues in deep learning-based medical image analysis: a systematic review. npj Digital Medicine 7(1), 286 (2024)
6. Puyol-Antón, E., Ruijsink, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., King, A.P.: Fairness in cardiac mr image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 413–423. Springer (2021)
7. Bissoto, A., Fornaciali, M., Valle, E., Avila, S.: (de) constructing bias on skin lesion datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
8. Oguguo, T., Zamzmi, G., Rajaraman, S., Yang, F., Xue, Z., Antani, S.: A comparative study of fairness in medical machine learning. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)

9. Pombo, G., Gray, R., Cardoso, M.J., Ourselin, S., Rees, G., Ashburner, J., Nachev, P.: Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3d deep generative models. Medical Image Analysis 84, 102723 (2023)
10. Li, X., Cui, Z., Wu, Y., Gu, L., Harada, T.: Estimating and improving fairness with adversarial learning. arXiv preprint arXiv:2103.04243 (2021)
11. Deng, W., Zhong, Y., Dou, Q., Li, X.: On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations. In: International Conference on Information Processing in Medical Imaging. pp. 158–169. Springer (2023)
12. Pakzad, A., Abhishek, K., Hamarneh, G.: Circle: Color invariant representation learning for unbiased classification of skin lesions. In: European Conference on Computer Vision. pp. 203–219. Springer (2022)
13. Wen, L., Zhou, Y., He, L., Zhou, M., Xu, Z.: Mutual information gradient estimation for representation learning. arXiv preprint arXiv:2005.01123 (2020)
14. Goceri, E.: Diagnosis of skin diseases in the era of deep learning and mobile technology. Computers in Biology and Medicine 134, 104458 (2021)
15. Belghazi, M.I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, R.D.: Mine: mutual information neural estimation. arXiv preprint arXiv:1801.04062 (2018)
16. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
18. Lum, K., Johndrow, J.: A statistical framework for fair predictive algorithms. arXiv preprint arXiv:1610.08077 (2016)
19. Do, H., Chang, Y., Cho, Y.S., Smyth, P., Zhong, J.: Fair survival time prediction via mutual information minimization. In: Machine Learning for Healthcare Conference. pp. 128–149. PMLR (2023)
20. Babaie-Zadeh, M., Jutten, C., Nayebi, K.: Differential of the mutual information. IEEE Signal Processing Letters 11(1), 48–51 (2004)
21. Babaie-Zadeh, M., Jutten, C.: A general approach for mutual information minimization and its application to blind source separation. Signal Processing 85(5), 975–995 (2005)
22. Wibisono, A., Wu, Y., Yang, K.Y.: Optimal score estimation via empirical bayes smoothing. arXiv preprint arXiv:2402.07747 (2024)
23. Antunes, J.T., Ofshteyn, A., Bera, K., Wang, E.Y., Brady, J.T., Willis, J.E., Friedman, K.A., Marderstein, E.L., Kalady, M.F., Stein, S.L., et al.: Radiomic features of primary rectal cancers on baseline t2-weighted mri are associated with pathologic complete response to neoadjuvant chemoradiation: a multisite study. Journal of Magnetic Resonance Imaging 52(5), 1531–1541 (2020)
24. Roh, Y., Lee, K., Whang, S.E., Suh, C.: Fairbatch: Batch selection for model fairness. arXiv preprint arXiv:2012.01696 (2020)
25. Tian, B., Du, R., Shen, Y.: Fairvit: Fair vision transformer via adaptive masking. In: European Conference on Computer Vision. pp. 451–466. Springer (2024)
26. Zhang, e.a.: Gradient estimation of information measures in deep learning. Information Sciences (2021)

27. Ghadiri, A., Pagnucco, M., Song, Y.: Xtranprune: explainability-aware transformer pruning for bias mitigation in dermatological disease classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 749–758. Springer (2024)
28. Liu, P., Lu, X.Y., He, K.: Real order total variation with applications to the loss functions in learning schemes. Communications in Contemporary Mathematics 26(07), 2350016 (2024)
29. Padhma, M.: A comprehensive introduction to evaluating regression models. Data Science Blogathon updated On October 31st (2023)