MIMM-X: Disentangling Spurious Correlations for Medical Image Analysis

Louisa Fay, Hajer Reguigui, Bin Yang, Sergios Gatidis, Thomas Küstner

University Hospital of Tübingen, University of Stuttgart, and Stanford University

Fairness of AI in Medical Imaging
MICCAI 2025 Workshop

The Problem: Shortcut Learning

- ➔ Deep Learning (DL) has transformed medical imaging due its power of identifying patterns in the image that are not immediately visible to the human eye, resulting in remarkable success for segmentation, disease detection, and diagnosis.

- ➔ Medical datasets are inherently heterogeneous:

  scanners, acquisition protocols, and patient demographics.

The MIMM-X Framework

- The goal is to disentangle causal features from multiple spurious correlations.\
  - Use of true causal relationships instead of dataset-specific shortcuts.
  - minimize the Mutual Information (MI) between the primary task features and the spurious factor features.
  - Extension of MIMM model (Fay et. al 2023) : only handled a *single* spurious factor.

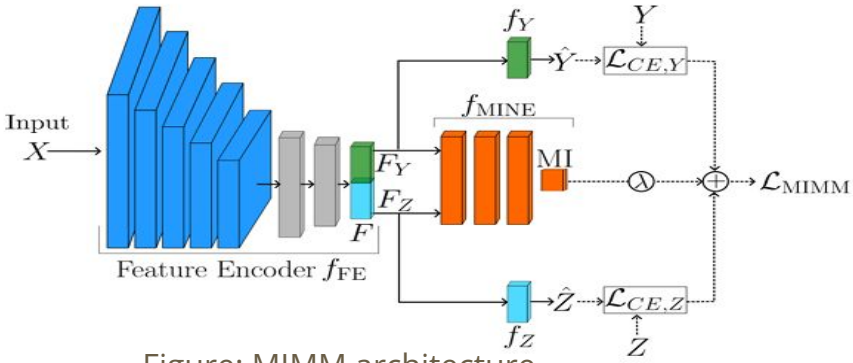
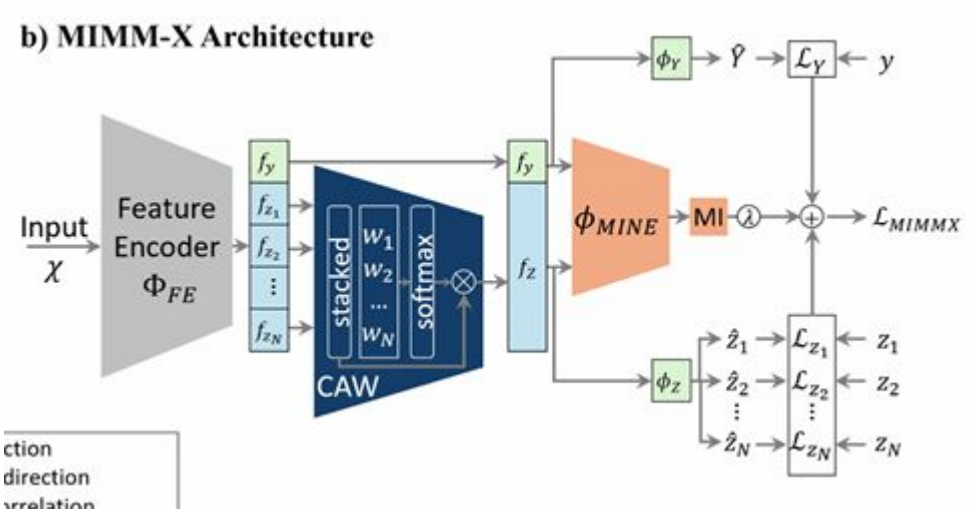


Figure: MIMM architecture

The MIMM-X Framework: Key Components

1. Feature encoder(ϕ_{FE}): Maps image X to f (f_y and f_Z)
2. MI Estimation (ϕ_{MINE}) (Belghazi et al.2018) : MI value between f_y and f_Z
3. Confounder Attention Weighter (CAW): Assigns adaptive attention weights to each spurious feature f_Z to emphasize relevant factors.



The MIMM-X Framework: Training

Overall Loss Function:

-Combines cross-entropy terms L_0 for the primary task (y) and spurious factors (z) with a Mutual Information penalty.

Dynamic Loss Scaling (DLS):

-Adjusts the loss scaling factors (γ_i) for each task.

$$\mathcal{L}_{\text{MIMM-X}} = \gamma_0 \mathcal{L}_0 + \sum_{i=1}^{N} \gamma_i \mathcal{L}_i + \lambda \phi_{MINE}(f_y, f_Z)$$

The MIMM-X Framework: Training

DLS function:

$$\gamma_i = \left(\frac{\mathcal{L}_i}{\overline{\mathcal{L}}}\right)^{\alpha_i}, \quad \text{where } \overline{\mathcal{L}} = \frac{1}{N+1}\sum_{i=0}^{N}\mathcal{L}_i, \quad \text{and } \alpha_i = \begin{cases} \alpha_Y, & \text{if } i = 0, \\ \alpha_Z, & \text{if } i = 1, \ldots, N \end{cases} \tag{2}$$

with dynamic α_Y for the primary task:

$$\alpha_Y = \alpha_{Y,\text{initial}} + \left(\frac{\text{epoch}}{N_{\text{epoch}}} \cdot \beta_Y\right), \tag{3}$$

where β_Y controls the increasing emphasis on y over training epochs.

Experiments

Exp	Modality	Dataset	Primary Task (y)	Spurious z_1	Spurious z_2	Spurious z_3
1	Brain MRI	NAKO/UKB	Age (young/old)	Sex (f/m)	Dataset (NAKO/UKB)	–
2	Chest X-Ray	CheXpert	Pleural Effusion (y/n)	Sex (f/m)	Age (young/old)	–
3	Chest X-Ray	CheXpert	Pleural Effusion (y/n)	**Sex (induced)**	**Age (natural)**	**CoD (natural)**

Experiments

1. **Experiment 1 (Brain MRI): Predict age as binary group(Y:<51, O:>57)**

   2D axial slices from 3D brain MRI (NAKO/UKB), resized to 256 × 256 and z-score normalized

   z_1: sex (female/male) and **z_2:** dataset (NAKO/UKB)

   Feature encoder(ϕ_{FE}): same feature encoder as in (Fay et al.2023) based on four convolutional layers. batch size:150 and N_B : 6.

   DLS hyperparameters : α_Y =0.3, α_Z =0.8, β_Y =0.01, and λ =1.5.

2. **Experiment 2 (Chest X-Ray): pleural effusion (yes/no)**

   chest X-rays from CheXpert(Irvin et el. 2019) downsampled to 96 × 96

   z_1: sex (female/male) and **z_2:** age (young: <50 years/old: > 60 years).

   Feature encoder(ϕ_{FE}): DenseNet-121 (Huang et al.2017)

   300 epochs , learning rate of 10−5 , batch size of 100 using N_B = 5.

   DLS hyperparameters : α_Y =0.3, α_Z =0.7, β_Y =0.01, and λ =1.5.

Experiments

1. **Experiment 3 (Chest X-Ray):**

-Same as Experiment 2 in primary task but sex is induced as a spurious correlation

-Simulate real world complexity and control for Age(z_2) and CoD(z_3) (often spuriously correlated with disease labels in clinical datasets), used with their natural distribution

-Evaluates if MIMM-X can disentangle y not only from known factor z_1, but also from naturally co-occurring variables that are not manually manipulated.

c) Experiment 3: Training distribution

Ratio of co-occurring labels

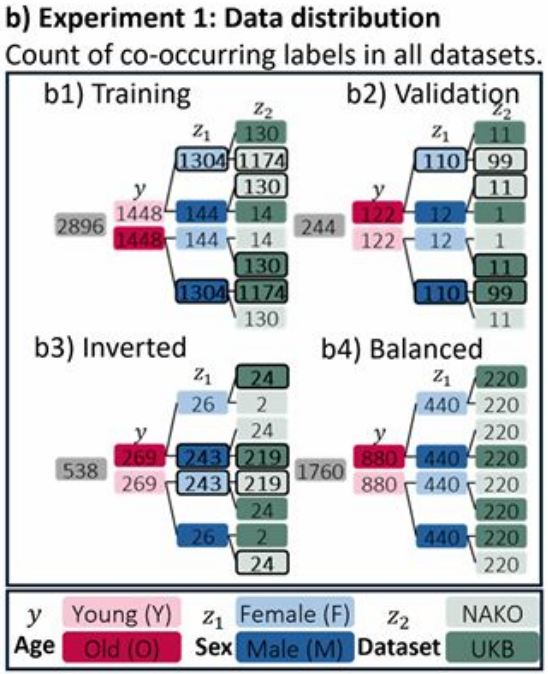
y	z_1		z_2		z_3	
Pleural	**Sex**		**Age**		**CoD**	
Effusion	0	1	0	1	0	1
No	0.45	0.05	0.19	0.31	0.29	0.21
Yes	0.05	0.45	0.12	0.38	0.14	0.36

Evaluation

- How to measure disentanglement?

  Test sets:

  1. Validation (Val.): Same correlations as training.
  2. Inverted (Inv.): Opposite correlations.
  3. Balanced (Bal.): No correlations



A disentangled model should have no significant performance drop when moving from the Validation set to the Inverted set.

Results: Experiment 1

- Lack of DLS: 60% decrease on Inv. set
- MIMM-X (Full) had only 1.4% drop

  best accuracy 82.8%
- Rebalanced is more robust(9.9% drop)
- High prediction performance of z1, z2

(a) **Classification accuracy** [%] across different evaluation sets (Val./Inv./Bal.).

			$f_y \to y$			$f_z \to z_1$			$f_z \to z_2$		
Method	**DLS**	**CAW**	**Val.**	**Inv.**	**Bal.**	**Val.**	**Inv.**	**Bal.**	**Val.**	**Inv.**	**Bal.**
Baseline	–	–	96.3	27.2	70.2	99.0	96.4	95.9	100.0	100.0	99.9
Baseline	✓	✓	89.6	42.9	66.6	94.2	91.1	90.5	100.0	100.0	99.8
Rebalance	✓	✓	91.8	81.9	87.3	97.5	96.7	97.3	100.0	100.0	100.0
dCor	–	–	96.2	41.7	73.4	98.8	95.7	96.8	100.0	100.0	99.7
MIMM	–	–	74.4	45.7	58.9	88.3	83.8	85.9	100.0	99.6	99.9
MIMM-X	–	–	46.3	54.0	53.5	89.2	88.3	71.6	99.2	98.3	95.8
MIMM-X	–	✓	94.2	33.2	69.0	63.3	57.5	66.5	54.6	44.5	57.3
MIMM-X	✓	–	85.6	74.2	79.8	96.8	96.3	89.4	98.2	98.4	95.4
MIMM-X	✓	✓	84.2	82.8	82.6	93.4	94.6	93.6	100	99.8	99.5

Results: Experiment 1

-Evaluation Disentanglement :

- predicting z_1/z_2 from f_y and

  y from f_{z1} /f_{z2} to assess residual shortcut information.

- MIMM-X and Rebalancing achieved near random guess
- Disadvantage of Rebalance

(b) **Disentanglement performance.** Accuracy in [%]. Ideally, near random (50%).

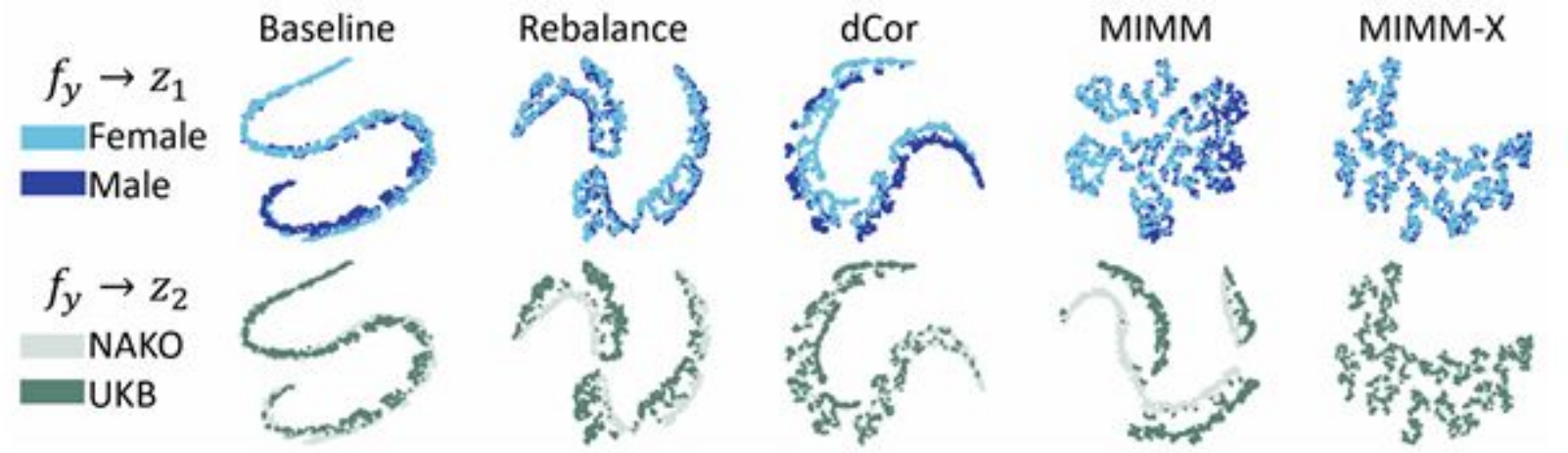
	Baseline			Rebalance			dCor			MIMM			MIMM-X		
	y	z_1	z_2	y	z_1	z_2	y	z_1	z_2	y	z_1	z_2	y	z_1	z_2
f_y	–	70.3	70.0	–	47.6	48.1	–	70.1	67.4	–	40.4	41.4	–	51.8	50.0
f_{z_1}	51.7	–	51.7	49.6	–	49.4	52.3	–	52.4	54.1	–	–	49.6	–	51.5
f_{z_2}	50.0	50.1	–	50.1	50.0	–	50.1	50.0	–	50.7	–	–	49.8	50.1	–

Results: Experiment 1

Visualizing Disentanglement (t-SNE)

Are the Features Independent?

Ideal case(disentanglement)-› No visual class separation should be seen.



Results: Experiment 2

- 2 Induced Correlations : sex and age
- Similar to the Brain MRI experiment, all comparison methods except Rebalancing failed on the inverted set (near 50% random-guess accuracy).
- MIMM-X generalized well and achieved the highest inverted accuracy (70.4%).
- Disentanglement performance: other methods leave info about f_Z in f_Y

(a) Classification accuracy [%] across different evaluation sets (Val./Inv./Bal.).

	$f_y \to y$			$f_z \to z_1$			$f_z \to z_2$		
Method	**Val.**	**Inv.**	**Bal.**	**Val.**	**Inv.**	**Bal.**	**Val.**	**Inv.**	**Bal.**
Baseline	82.6	42.6	64.0	79.7	74.8	73.6	74.4	70.5	57.4
Rebalance	77.7	70.3	76.4	70.8	69.3	71.2	76.5	71.9	74.8
dCor	87.8	42.2	69.0	80.0	62.3	65.1	83.7	63.8	64.5
MIMM	87.2	51.0	71.4	72.1	68.9	66.7	63.6	64.1	53.1
MIMM-X (ours)	86.3	70.4	76.5	79.1	76.2	77.8	78.2	74.8	77.2

(b) Disentanglement performance. Accuracy in [%]. Ideally, near random (50%).

	Baseline			**Rebalance**			**dCor**			**MIMM**			**MIMM-X**		
	y	z_1	z_2	y	z_1	z_2	y	z_1	z_2	y	z_1	z_2	y	z_1	z_2
f_y	–	64.5	62.5	–	54.9	53.9	–	66.4	64.0	–	61.1	61.0	–	54.8	54.3
f_{z_1}	57.9	–	55.5	48.8	–	49.4	56.8	–	58.7	49.8	–	–	50.0	–	50.0
f_{z_2}	52.9	62.6	–	52.7	50.5	–	65.8	65.0	–	50.4	–	–	49.6	50.1	–

Results: Experiment 3

(1 Induced z_1:sex, 2 Natural: age and CoD)

-MIMM-X again achieved the highest performance (70.2%) on the inverted set,

- Good generalization in both induced and natural correlations

(a) Classification accuracy [%] across different evaluation sets (Val./Inv./Bal.).

	$f_y \to y$			$f_z \to z_1$			$f_z \to z_2$			$f_z \to z_3$		
Method	**Val.**	**Inv.**	**Bal.**	**Val.**	**Inv.**	**Bal.**	**Val.**	**Inv.**	**Bal.**	**Val.**	**Inv.**	**Bal.**
Baseline	85.4	57.4	71.5	85.7	62.2	58.6	73.2	74.4	73.6	63.5	61.0	62.5
Rebalance	74.2	69.4	71.4	76.8	71.6	74.0	71.7	73.0	72.1	57.7	55.5	57.1
dCor	87.0	54.9	70.9	87.5	72.0	74.0	81.0	80.7	78.0	61.5	57.6	57.5
MIMM-X (ours)	81.7	70.2	72.9	81.5	72.7	74.2	81.4	78.3	79.4	58.9	57.5	58.1

(b) Disentanglement performance. Accuracy in [%]. Ideally, near random (50%).

	Baseline				**Rebalance**				**dCor**				**MIMM-X**			
	y	z_1	z_2	z_3	y	z_1	z_2	z_3	y	z_1	z_2	z_3	y	z_1	z_2	z_3
f_y	–	66.6	60.9	60.3	–	45.7	53.5	50.8	–	70.0	62.2	61.3	–	52.6	53.1	50.9
f_{z_1}	65.9	–	62.8	60.6	68.5	–	62.8	58.5	62.6	–	60.8	59.5	64.7	–	59.9	56.0
f_{z_2}	59.9	59.3	–	57.1	57.1	55.0	–	53.7	57.5	53.8	–	52.5	56.9	53.5	–	52.3
f_{z_3}	55.4	53.9	60.1	–	50.1	50.1	56.2	–	50.0	50.0	56.0	–	50.0	50.0	56.2	–

Limitation And Discussion

-MIMM-X currently relies on specifying the confounding factors in advance. The model must be told what to disentangle.

-While Rebalancing can perform well, it requires significantly larger, carefully designed training sets. This is often not feasible. MIMM-X works without modifying the data distribution.

Take home message:

- MIMM-X is a scalable method to mitigate multiple spurious correlations in medical imaging, improving robustness and feature disentanglement. It can be used not just for synthetic biases, but for naturally occurring correlations as well.

Future work:

- Extend the model to discover unknown spurious correlations automatically and handle more complex confounding structures.

Code:https://github.com/lab-midas/MIMM-X