IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 June 2023, accepted 11 June 2023, date of publication 26 June 2023, date of current version 29 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3289397

RESEARCH ARTICLE

Avoiding Shortcut-Learning by Mutual Information Minimization in Deep Learning-Based Image Processing

LOUISA FAY[1,2], ERICK COBOS[1,3], BIN YANG[2], (Senior Member, IEEE), SERGIOS GATIDIS[1,3], AND THOMAS KÜSTNER[1], (Member, IEEE)

[1]Medical Image and Data Analysis (MIDAS.lab), Department of Diagnostic and Interventional Radiology, University Hospital of Tuebingen, 72076 Tübingen, Germany
[2]Institute of Signal Processing and System Theory, University of Stuttgart, 70550 Stuttgart, Germany
[3]Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany

Corresponding author: Louisa Fay (louisa.fay@med.uni-tuebingen.de)

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant 428219130, in part by DFG through Germany's Excellence Strategy—EXC 2064/1 under Grant 390727645, in part by the German National Cohort (GNC) (www.nako.de) through the Federal Ministry of Education and Research (BMBF) under Grant 01ER1301A/B/C and Grant 01ER1511D, in part by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health) under Grant U01 AG024904, and in part by the Department of Defense (DOD) ADNI under Award W81XWH-12-2-0012.

ABSTRACT Deep learning models are increasingly being used in detecting patterns and correlations in medical imaging data such as magnetic resonance imaging. However, conventional methods are incapable of considering the real underlying causal relationships. In the presence of confounders, spurious correlations between data, imaging process, content, and output can occur that allow the network to learn shortcuts instead of the desired causal relationship. This effect is even more prominent in new environments or when using out-of-distribution data since the learning process is primarily focused on correlations and patterns within the data. Hence, wrong conclusions or false diagnoses can be obtained from such confounded models. In this paper, we propose a novel framework, denoted as Mutual Information Minimization Model (MIMM), that predicts the desired causal outcome while simultaneously reducing the influence of present spurious correlations. The input imaging data is encoded into a feature vector that is split into two components to predict the primary task and the presumed spuriously correlated factor separately. We hypothesize that learned mutual information between both feature vector components can be reduced to achieve independence, i.e., confounder-free task prediction. The proposed approach is investigated on five databases: two non-medical benchmark databases (Morpho-MNIST and Fashion-MNIST) to verify the hypothesis and three medical databases (German National Cohort, UK Biobank, and ADNI). The results show that our proposed framework serves as a solution to address the limitations of conventional deep learning models in medical image analysis. By explicitly considering and minimizing spurious correlations, it learns causal relationships which result in more accurate and reliable predictions. The novel contributions in this work are: 1) the separation of features into the prediction of the primary task and the spuriously correlated factor; 2) MIMM targets the preservation of invariance to counterfactuals, prevents shortcut learning, and enables confounder-free network training; and 3) the mutual information minimization addresses heterogeneous data cohorts as usually encountered in the medical domain.

INDEX TERMS Causality, deep learning, medical image analysis, mutual information, shortcut learning.

I. INTRODUCTION

The tremendous development of deep learning (DL) algorithms and their promising results have the potential to

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

fundamentally transform clinical workflows in the forthcoming decades [1], [2]. DL models are already able to not only reconstruct medical images [3], [4], [5], [6] but also diagnose diseases [7], [8], [9], [10], suggest treatments [11], [12], [13], [14], or segment different tissues in the body [15], [16], [17], [18].

In general, the predictions of DL algorithms are based on the detection of patterns and correlations in given training data. Trained DL models are evaluated on test data from the same distribution [19]. As soon as the distribution of test data, e.g., in real-world scenarios, deviates from the distribution of data utilized during training, it is likely that the model will exhibit poor performance since its predictions are based on learned correlations between the input samples and their corresponding labels [20], [21]. However, as Fig. 1A shows, hidden confounding within the training data bears the risk to generate spurious correlations that reflect biases or selection effects within the training data [22], [23], [24], [25]. Consequently, the networks may have been trained on these shortcuts instead of on the intended causal relation [26]. DL models can therefore fail to correctly predict outputs in a new environment when the spurious correlation changes and thus, the shortcut deviates from the causal relationship [27], [28], [29].

Especially in medical imaging, multiple factors such as scanners, acquisition sites, scan conditions (e.g., patient positioning, imaging protocol, etc.), and patient compliance, etc. influence the acquisition of medical images like magnetic resonance imaging (MRI), computed tomography (CT), or positron emission tomography (PET) [27], [30]. DL models that are aimed to operate on a heterogeneous or confounded imaging cohort, tend to learn shortcuts based on spurious correlations instead of task-specific features [27], [31]. While these confounded networks may perform reasonably well within the same data distribution (as training), out-of-distributional generalization and objective image analysis are impaired [32], [33]. The usage of these methods within a clinical context could mean that wrong or overly optimistic predictions are obtained which could be life-threatening as the real causal relationship is missed [27], [33]. For example, a hospital has two MR scanners from the same manufacturer. Scanner A is older and produces noisier images compared to Scanner B. While Scanner A scans mainly healthy patients, Scanner B primarily scans patients with Alzheimer's disease (AD). When training a DL model to predict AD with such data, learning the shortcut created by the spurious correlation between AD and noise level is easier than learning the desired but hidden causal relationship between AD and anatomical features. Thus, an AD patient scanned on Scanner A will most likely be wrongly categorized as healthy. For the successful application and support of DL models in clinical workflows, it is essential to enhance the model's robustness to be able to make correct predictions in every environment even if the predictor is trained on an unbalanced and spuriously correlated dataset.

This paper addresses this challenge by proposing the Mutual Information Minimization Model (MIMM). The aim is to train a counterfactual invariant predictor robust to any changes of known spurious correlations within the training data. Our framework predicts the task-specific output (e.g., AD patient) while simultaneously reducing the influence of spuriously correlated factors (e.g., type of scanner) in the dataset. The rationale of this work is to split the information obtained from the actual task and the confounding, which allows us to train models that avoid shortcut learning and are more robust to distribution shifts.

The proposed method is based on a DL architecture, which comprises two major parts. In the first part, a feature encoder embeds the input images to a feature vector. The resulting feature vector is divided into two parts, one part is used for the prediction of the primary task, and the other part is for the prediction of the spuriously correlated factors. In order to make these two feature subvectors independent from each other, the framework aims to reduce the mutual information (MI) between them. Therefore, in the second part, the MI is estimated with a Mutual Information Neural Estimator (MINE) [34] to achieve independent features and train a counterfactual invariant predictor. Enforcing MI-based feature independence during training enables the proposed MIMM to generate causal and counterfactual invariant predictions. The main contributions of this work are:

1) MIMM is a deep learning-based approach that is capable of predicting the primary task and the given spurious correlation.
2) Our model is able to reduce the influence of spuriously correlated factors on the features of the primary task and thus, avoids shortcut learning over spurious correlations.
3) We show that MIMM exhibits robustness against varying spurious correlations and counterfactual samples.
4) The proposed pipeline can be easily adapted to different tasks and applications. MIMM was investigated on five different databases including large-scale epidemiological cohort studies with medical imaging data.

Our work is organized as follows. Firstly, in Section II-*Related Work*, we provide a review of previous work related to our problem. Section III-*Causal Background* introduces causality and its challenges with spurious correlation and confounding. The subsequent Section IV-*Methods* gives a detailed introduction to MIMM, its architecture, and its training process. In Section V-*Experiments and Materials*, we introduce our experiments and the materials used. We then present the *Results and Discussion* in Section VI. Finally, in Section VII-*Conclusion*, we present our concluding remarks.

II. RELATED WORK

In the field of deep learning, causality has become an increasingly important and relevant topic with the aim to understand the relationship between input and output

variables [35], [36], [37]. The main idea is causal representation learning which combines concepts from representation learning and causal inference by predicting outputs from patterns but with the incorporation of causal dependencies [38]. Research into the area of causality in deep learning has focused on a range of topics, including causal inference, domain adaptation, and shifts, as well as counterfactual analysis [38], [39], [40], [41].

Important aspects for causal relationship learning were raised in [28] and [38] in which the differences between causal settings and predictive direction were defined. Thereby, data $\boldsymbol{X}$ is composed of its causal component Y, the real label of the primary task, and the spuriously correlated factor Z. Our work focuses on the predictive direction, which is the inverse of the causal direction [42] since the cause in the given input images $\boldsymbol{X}$ are the effect of the ground truth Y and all latent spuriously correlated factors Z [38], [43].

Multiple techniques have been developed to mitigate the learning of biases in databases. One approach involves debiasing the training dataset through methods, like data augmentation [44], to weaken spurious correlations by creating counterfactual samples. However, data augmentation requires specific domain knowledge. Another approach is to address the class imbalance by oversampling the minority class with replacement [45] or by incorporating data from different sources to improve generalization [46]. Nevertheless, obtaining data from diverse sources is often challenging, particularly in the medical domain.

Another line of work adds new layers to the DL model that filter biases during training. For instance, [47], [48] utilize plug-in layers to remove the influence of metadata (confounding) after a DL layer. However, this approach suffers from computational inefficiency, as a closed-form solution needs to be calculated for each feature at every iteration.

Other approaches aim to disentangle spuriously correlated features from the primary task features using similarity measurements such as matrix correlation [49], Hilbert-Schmidt independence criterion (HSIC) [50], [51], cosine similarity [52], [53] as well as maximum mean discrepancy (MMD) [54] as additional regularization terms after hidden layers in the model.

Several approaches have been investigated to mitigate the learning of biases in databases through adversarial learning [50], [55], [56]. These methods tackle the problem by formulating a min-max optimization framework, where the objective is to simultaneously maximize the prediction loss associated with the spuriously correlated task and minimize the loss related to the primary task. Previous studies have shown the effectiveness of incorporating constraints into the learning problem. [57]. However, these approaches faced the challenge of minimizing the mutual information since the calculation of MI is computationally complex, and the MI is non-differentiable [58]. We tackle this problem by estimating the MI with a MINE model [34]. In the scope of this work, we follow the approach of [28]. While [28] rely on

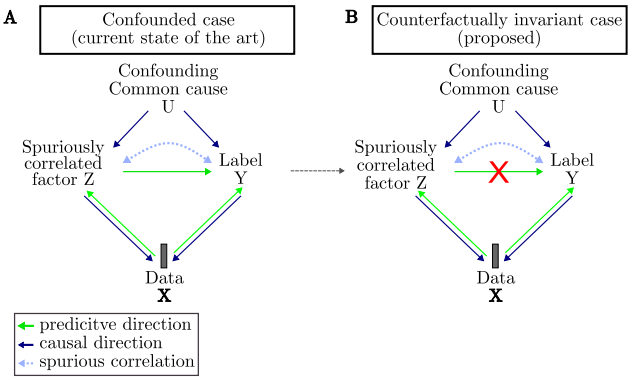


FIGURE 1. Visualization of the causal direction, the spurious correlation, and possible DL-based prediction paths when confounding is present in data. (A) Current methods: Confounding causes a spurious correlation, which bears the risk that a DL model learns a shortcut over a spuriously correlated factor (e.g., writing style, see Fig. 2) instead of the real causal relationship (e.g., shape) between input data X and ground truth label Y. (B) Proposed method: By interrupting the shortcut, learning a spurious correlation is avoided. Hence, the DL model is able to learn the real causal relationship.

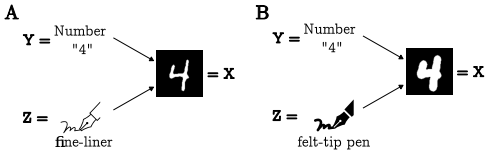


FIGURE 2. Example of a simplified causal image generation of counterfactuals. (A) An image is generated by the variables Y and Z, where Y is the idea of the number 4 and Z is the writing style with a fine-liner. (B) Changing the writing style Z from a fine-liner to a felt-tip pen only changes the writing style but does not change the shape of the number.

MMD as a regularization term in their loss term, we use the estimated MI between the features relevant to the actual task and the features affected by confounding factors as a penalty term. This approach specifically targets the preservation of invariance to counterfactuals and the prevention of shortcut learning, which have a positive impact on the robustness of the model against changes in data distribution or domain shifts [28].

In the context of causal representation learning, the interest in fair representation learning is growing due to the discovery that many widely known databases, such as Adience [59], IARPA Janus Benchmark A [60], and ImageNet [61] contain significant biases and under-representation of certain groups, e.g., darker-skinned subjects who are represented in less than 20% of the samples. Similar unintentional biases exist in medical databases [62]. Especially the influence of scanners for medical imaging adds bias to the images due to individual scanner variations [63]. Moreover, selection effects, for instance, due to the choice of subjects for a study, can also lead to unintentional under-representation or oblivion of specific groups [62]. These biases and selection effects tremendously influence the hazard of spurious correlation within databases and hence, reinforce shortcut learning.

Models trained on such databases tend to make discriminatory and racially biased predictions and moreover, bear the risk to make wrong conclusions, as demonstrated in [64] and [65]. Therefore, several studies [56], [66], [67], [68], [69], [70], [71], [72] focused on fair and invariant representation learning with respect to protected and sensitive attributes. In line with these works, we aim to provide a bias-free learning strategy.

III. CAUSAL BACKGROUND

A common objective of a DL model is to predict a ground truth label Y from an input $\boldsymbol{X}$ by learning a general mapping function f that approximates the inverse of an underlying causal mechanism [42], [73]. In such a scenario, input data $\boldsymbol{X}$ are caused by the label Y and multiple latent variables Z, which we also refer to as spuriously correlated factors. For the case of simplicity, we consider Y as one variable and summarize Z as another variable. As exemplarily depicted in Fig. 2A, the image $\boldsymbol{X}$ is caused by the digit $Y = 4$ and the writing style $Z =$ ''thin''.

In an ideal scenario, Y and Z are independent, and a DL model is able to learn a mapping f that predicts Y from $\boldsymbol{X}$ without being confounded by Z as formally stated [28]:

$$f(\boldsymbol{X}) \perp\!\!\!\perp Z|Y \tag{1}$$

As introduced in [28], a predictor f that fulfills the criteria in (1) is counterfactually invariant. This means the predictions $f(\boldsymbol{X})$ are invariant to perturbations Z^* of Z on the input $\boldsymbol{X}$ and hence,

$$f(\boldsymbol{X}(Y, Z)) = f(\boldsymbol{X}(Y, Z^*)). \tag{2}$$

Regarding the example of Fig. 2, by changing the writing style $Z =$ ''thin'' to $Z =$ ''thick'', the prediction of a counterfactual invariant predictor remains the same. This fulfills the criteria of (1) as well as of (2): $f(\boldsymbol{X}(Y = 4, Z =$ ''thin''$)) = f(\boldsymbol{X}(Y = 4, Z =$ ''thick''$))$.

However, in practice, this mapping function is rarely able to be an exact inverse of the causal mechanism due to biases, confounding, selection effects, or other factors within the training data. These kinds of factors induce, as shown in Fig. 1A, spurious correlations between Y and Z, which confound the causal relationship between Y and $\boldsymbol{X}$ as well as Z and $\boldsymbol{X}$. A state-of-the-art predictor trained on the confounded data may not be counterfactually invariant since it is not able to differentiate between causal relationships and spurious correlations. A predictor (unintentionally) trained on spurious correlations fails as soon as the spuriously correlated factor Z changes. Under these circumstances, (1) and (2) are no longer valid.

In general, spurious correlations emerge due to a confounding induced by an unobserved common cause. The common cause principle of Reichenbach [74] states that there exists an unobserved common cause U that explains the spurious correlation between two statistically dependent variables Z and Y but makes them independent when conditioning on the common cause U. For instance, one physician working in Room 1 with scanner A always performs MRIs on healthy patients, while another physician working in Room 2 with scanner B, an older version of scanner A, is in charge of scanning patients with Alzheimer's Disease. This creates a spurious correlation between scanner A and healthy patients, as well as scanner B and patients with Alzheimer's Disease. However, the correlation between the scanner and the Alzheimer's Disease of a patient is spurious since there is no causal explanation. Neither the scanner causes Alzheimer's Disease nor the patient causes the type of scanner. The present common cause is the physician and having this information leads to $Y =$ ''type of disease'' and $Z =$ ''scanner'' being independent.

As depicted in Fig. 1B, the aim of our proposed approach is to cut the predictive connection between Z and Y which is generated by the spurious correlation, so that the real underlying causal structure between $\boldsymbol{X}$ and Y is learned by the predictor.

IV. METHODS

This section introduces the proposed architectural design of MIMM. Section IV-A gives a general overview of the proposed architecture MIMM, while Section IV-B and Section IV-C describe the feature encoder and MINE [34] in more detail, respectively.

A. MUTUAL INFORMATION MINIMIZATION MODEL (MIMM)

The architecture of our proposed Mutual Information Minimization Model referred to as MIMM, is designed to predict both the primary task (Y) and the spuriously correlated factor (Z) simultaneously as depicted in Fig. 3. Generally, the model comprises four parts, the feature encoder, two classification heads, and the MINE model.

B. FEATURE ENCODER MODEL AND CLASSIFICATION HEAD

The feature encoder $f_{FE}(\boldsymbol{X}, \theta_{\text{FE}})$ with the parameter θ_{FE} transforms an input sample $\boldsymbol{X}$ to a feature vector $\boldsymbol{F}$. This vector is split into two parts. The upper part $\boldsymbol{F_Y}$ is responsible for the prediction of the primary task with the ground truth Y while the lower part $\boldsymbol{F_Z}$ is used to predict the known spuriously correlated factor with the ground truth Z. In the scope of our work, the subvectors are of equal size and we set the length of $\boldsymbol{F}$ to be the sum of the classes of Y and Z. Each subvector is fed separately into a classification head, $f_y(\boldsymbol{F_Y}, \theta_Y)$ and $f_z(\boldsymbol{F_Z}, \theta_Z)$, which is in our case a log-softmax. Thereby, the classification head f_Y predicts Y from the feature subvector $\boldsymbol{F_Y}$, while the spuriously correlated factor Z is predicted by the classification head f_Z with $\boldsymbol{F_Z}$. The explicit architecture of the feature encoder model is task-specific. It is therefore described in Section V when introducing our experiments.

However, having this network architecture composed of these two parts bears the risk that the vectors $\boldsymbol{F_Y}$ and $\boldsymbol{F_Z}$ are correlated and share mutual information. This implies that the primary task predictor $f(\boldsymbol{X}) = f_Y(\boldsymbol{F_Y})$ does not comply

with (1) and thus, is not counterfactually invariant. As stated above, this results in the problem that the prediction of Y might be partially based on the spuriously correlated factor Z and therefore, would fail in a new environment when Z has changed. To avoid the problem of learning Y over the shortcut of the spuriously correlated factor Z, we propose to minimize the mutual information (MI) between $\boldsymbol{F_Y}$ and $\boldsymbol{F_Z}$ in order to obtain independent feature subvectors.

C. MUTUAL INFORMATION NEURAL ESTIMATION (MINE)

To estimate the MI between the upper part $\boldsymbol{F_Y}$ and the lower part $\boldsymbol{F_Z}$ of the feature vector, both feature subvectors serve as input to a MINE model, which is able to estimate the MI. MINE was introduced in [34]. The estimated MI serves as a penalty term in the loss function of the feature encoder as shown in (6). The penalization forces the feature encoder to generate independent vectors $\boldsymbol{F_Y}$ and $\boldsymbol{F_Z}$.

MI is a measure to quantify the dependence between two variables. It is also described as the difference between the entropy of a marginal variable $H(\boldsymbol{F_Y})$ and the conditional entropy $H(\boldsymbol{F_Y}|\boldsymbol{F_Z})$. In general, the estimation of MI for discrete random vectors $\boldsymbol{F_Y}$ and $\boldsymbol{F_Z}$ is defined as

$$I(\boldsymbol{F_Y};\boldsymbol{F_Z}) = \sum_{\boldsymbol{F_Y},\boldsymbol{F_Z}} p(\boldsymbol{F_Y},\boldsymbol{F_Z}) log\left(\frac{p(\boldsymbol{F_Y},\boldsymbol{F_Z})}{p(\boldsymbol{F_Y})p(\boldsymbol{F_Z})}\right), \quad (3)$$

where $p(\boldsymbol{F_Y},\boldsymbol{F_Z})$ is the joint probability mass function (PMF) and $p(\boldsymbol{F_Y})$ and $p(\boldsymbol{F_Z})$ are the marginal PMFs. As (3) reveals, MI can also be represented as the Kullback-Leibler (KL-) divergence $D_{KL}(p(\boldsymbol{F_Y},\boldsymbol{F_Z})||p(\boldsymbol{F_Y})p(\boldsymbol{F_Z}))$ between the joint distribution $p(\boldsymbol{F_Y},\boldsymbol{F_Z})$ and the product of the marginal distributions, $p(\boldsymbol{F_Y})$ and $p(\boldsymbol{F_Z})$, since the KL-divergences is described as

$$D_{KL}(P||Q) = E_{p(\boldsymbol{X})}\left[log\left(\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}\right)\right]. \quad (4)$$

The vectors $\boldsymbol{F_Y}$ and $\boldsymbol{F_Z}$ are independent if and only if $I(\boldsymbol{F_Y};\boldsymbol{F_Z}) = 0$.

The idea of MINE is to parameterize a set of functions $\mathcal{F} = \{T_\theta\}_{\theta\in\Theta}$ by a neural network, called the statistic network f_{MINE}. It returns the output of one parameterized function $T_\theta(\boldsymbol{F_Y},\boldsymbol{F_Z})$. Therefore, the dual representation of KL-divergence is introduced and described by the lower bound in 5.

$$\mathcal{D}_{KL}(\mathbb{P}||\mathbb{Q}) \geq \sup_{T\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[T] - log(\mathbb{E}_{\mathbb{Q}}[e^T]) \quad (5)$$

The architecture of the statistic network f_{MINE} is composed of four fully-connected layers, where the first three comprise 400 units and the last layer returns a single output with the estimated MI. The output of the MINE model is computed and maximized with (5) and serves as a lower bound approximation for the KL divergence or MI, respectively, see [34] for more details.

D. TRAINING PROCESS

The feature encoder model f_{FE} is minimized based on the loss function $\mathcal{L}_{\text{MIMM}}$ in (6). It is composed of the cross-entropy loss of the primary task Y and its estimation $\hat{Y} = f_Y(\boldsymbol{F_Y})$ in (7) as well as of the cross-entropy loss of the spuriously correlated factor Z and its estimation $\hat{Z} = f_Z(\boldsymbol{F_Z})$ in (8), and is additionally penalized by the estimated MI. Even though, the ideal value of MI would be 0, in practice this value of MI $= 0$ is not achieved, since the MI regularization term is a soft constraint.

Since the input to MINE changes after each update of the feature encoder model, the training process of MIMM is performed in an alternating fashion. Thereby, the feature encoder is trained for 1 batch followed by $(N_B - 1)$ batch updates of the MINE model, where N_B is a hyperparameter.

$$\mathcal{L}_{\text{MIMM}} = \mathcal{L}_{\text{CE, Y}}(\boldsymbol{X}, Y) + \mathcal{L}_{\text{CE, Z}}(\boldsymbol{X}, Z) + \lambda \cdot \text{MI}(\boldsymbol{X}) \quad (6)$$

$$\mathcal{L}_{\text{CE, Y}} = -Y^T \log f_Y(\boldsymbol{F_Y}) \quad (7)$$

$$\mathcal{L}_{\text{CE, Z}} = -Z^T \log f_Z(\boldsymbol{F_Z}) \quad (8)$$

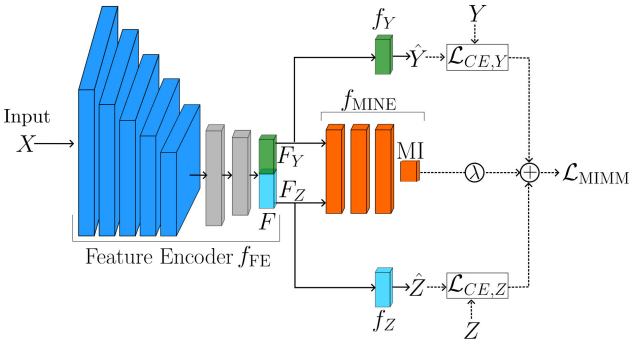


FIGURE 3. Architecture of the proposed MIMM. Feature encoder f_{FE} encodes an input image X, resulting in a feature vector (F) which is split into two parts. The upper part F_Y predicts the primary task Y with the classification head f_Y, while the lower part F_Z predicts the spuriously correlated factor Z with the classification head f_Z. The MI between F_Y and F_Z is estimated by a MINE model f_{MINE}. The aim is to minimize the MI between the two parts of F. For this purpose, MI is added as penalty to the loss function $\mathcal{L}_{\text{MIMM}}$ when updating the f_{FE}. The f_{MINE} itself is updated with its estimated MI value.

TABLE 1. Summary of used datasets, primary tasks, and spurious correlation task of all experiments.

Experiment	Dataset	Primary task Y	Spuriously correlated factor Z
#1	Morpho-MNIST	2 classes small/high digits	2 classes thin/thick writing style
#2	Adapted Fashion-MNIST	10 classes fashion items	10 classes boundary bars
#3	Brain MRI UKB/NAKO	2 classes male/female	2 classes UKB/NAKO
#4	Brain MRI UKB/NAKO	2 classes young/old	2 classes male/female
#5	Brain MRI ADNI	2 classes Alzheimer's/ healthy	2 classes GE/Siemens

TABLE 2. Number of training samples for the primary and spurious correlation task in all experiments. (In experiment 2, both Y and Z are 10-class variables. The short notation $Y = 1$ ($Z = 1$) represents the union of the remaining 9 classes with respect to the selected class $Y = 0$ ($Z = 0$)).

	Experiment									
	#1		#2		#3		#4		#5	
Z	Y		Y		Y		Y		Y	
	0	1	0	1	0	1	0	1	0	1
0	9264	487	5700	297	3747	416	2404	266	2322	255
1	487	9264	297	5700	416	3747	266	2404	255	2322

V. EXPERIMENTS AND MATERIALS

Our proposed MIMM approach is investigated on five different databases. This chapter describes our experimental design including the evaluation steps and general training setting, followed by a description of the applied datasets and the methods used for comparison. Our implementation is available at www.github.com/lab-midas/MIMM.

A. EXPERIMENTAL DESIGN

1) TASKS AND DATASETS

We conduct five experiments using five different datasets. All experiments are designed in a similar way. Four experiments perform a two-class primary classification task Y and a two-class spuriously correlated factor classification task Z. The other experiment deals with a ten-class primary and a ten-class spuriously correlated factor task. Table 1 summarizes both tasks and the applied dataset in all five experiments, see below for more details about the tasks and datasets.

2) DATASET DISTRIBUTION

Table 2 summarizes the number of training samples for all (Y, Z) combinations in each experiment. Clearly, the primary classes Y are all balanced, i.e. each primary task Y has the same number of training samples. In each primary class Y, however, the training samples originate from two or ten different spuriously correlated factors Z with a ratio of roughly 95:5 for the first two (non-medical) experiments and 90:10 for the remaining three medical experiments. This means 95% (90%) of the training samples of one primary class are associated with one value of the spuriously correlated factor while the remaining 5% (10%) training samples are associated with the other values of the spuriously correlated factor. This causes a spurious correlation between the primary task Y and the spurious correlated factor task Z which could lead to shortcut learning. As a representative example, Fig. 4a) depicts the class distribution of Experiment 4 during training, where the subjects are divided into two age groups: young ($\leq$51 years) and old ($\geq$57 years). These groups are spuriously correlated by sex. Hence, during training 90% of the young subjects are female and only 10% of the old subjects are female. It is vice versa for the male subjects.

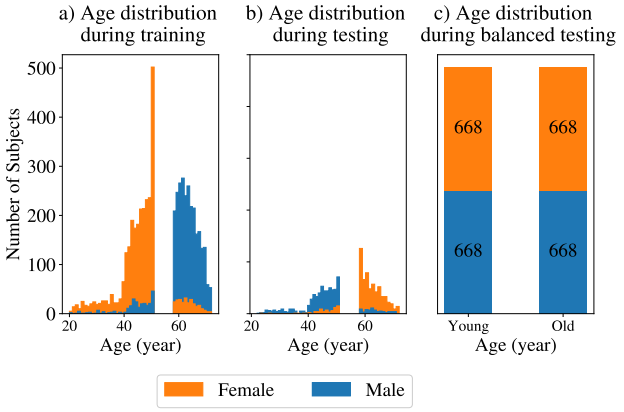


FIGURE 4. UKB/NAKO age group and sex (#4): Data distribution for the classification of age groups where group young is under the age of 51 and group old is above the age of 57. For a distinct separation of the age groups, subjects between 51 and 57 years are excluded from the experiment. The data samples are equally drawn from UK Biobank and NAKO. (a) Confounded training dataset; young subjects: 90% female, 10% male, old subjects: 90% male, 10% female. (b) Confounded test dataset with unseen samples and inverted distribution compared to the training distribution; young subjects: 10% female, 90% male, old subjects: 10% male, 90% female. (c) Balanced test dataset with unseen samples and balanced distribution; young subjects: 50% female, 50% male, old subjects: 50% male, 50% female.

a: EVALUATION

To evaluate the trained models and to demonstrate the suppressed shortcut learning, the classification accuracy is computed for a) new unseen samples of a validation dataset with the same distribution as the training set (e.g., Fig. 4a), b) a test set with a flipped distribution (e.g., Fig. 4b) and 4c) a dataset with balanced classes.

By carrying out the evaluation on these three datasets, we aim to test the performance of our model using different distributions. We expect that MIMM is able to avoid performance degradation when shifting from validation to test distribution. To show that $\boldsymbol{F_Y}$ and $\boldsymbol{F_Z}$ are independent and do not share information, $\boldsymbol{F_Y}$ is additionally used to predict Z (denoted as $\boldsymbol{F_Y} \rightarrow Z$), while $\boldsymbol{F_Z}$ predicts Y (denoted as $\boldsymbol{F_Z} \rightarrow Y$). For this evaluation step, the balanced dataset c) is applied. Thereby, an accuracy close to a random guess is desired.

b: TRAINING SETTINGS

All our models are trained with 5-fold-cross validation using ADAM optimizer with a learning rate of 10^{-4}. The training of each MIMM model is performed asynchronously, meaning the feature encoder and the classification heads are jointly trained for one batch followed by $N_B - 1$ updates of MINE. The estimated MI is added to the loss function multiplied by the hyperparameter λ as shown in (6).

B. EXPERIMENTS ON BENCHMARK DATASETS

The first two experiments are based on two non-medical benchmark datasets.

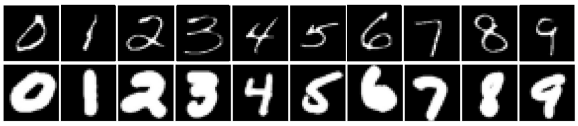


FIGURE 5. Morpho-MNIST (#1): Selection of thinly and thickly written digits from the Morpho-MNIST dataset. The different writing styles are the spuriously correlated factor attributed to the confounding influence of the writer in the scope of this experiment.

1) MORPHO-MNIST (#1)

The first experiment aims to predict small digits (0-4) and high digits (5-9) on the Morpho-MNIST dataset [75] which is the primary binary classification task. This dataset is a morphometric extension of the well-known MNIST dataset. In the scope of this work, we use samples that are thinned and thickened. A selection of morphometric digits is presented in Fig. 5. From this dataset, we create a confounded dataset with the different writing styles, thin and thick. To give a real-world example, two writers create training data. One writer writes mostly small digits between 0-4 with a fine liner. The other writer has a felt-tip pen and writes primarily high digits between 5-9. This setup bears the risk that a typical neural network trained on this data does not learn the digit group based on the shape of the number but rather by the thin or thick writing style. However, our aim is to overcome this challenge and force the model to predict the correct digit group based on the shape of the number.

Settings: The numbers of training samples for this experiment are shown in Table 2. Both digit groups are balanced with the same number of $9264 + 487 = 9751$ samples. However, in each digit group, the ratio of training data amount for thin/thick writing style is approximately 95:5 for small and 5:95 for high digits. This causes a spurious correlation between the target primary class Y and writing style Z.

Likewise the original MNIST dataset, each sample is two-dimensional with a resolution of 28×28 pixels. The architecture of the feature encoder is composed of three of 3×3 convolutional layers with 6 and two times 16 channels. After each convolutional layer, ReLU activation and 2×2 max-pooling are operated on the output. This is followed by two fully-connected layers of size 256 and 4. Training is performed on a batch size of 1000 samples with $N_B \in \{3, 5\}$ and $\lambda \in \{0.05n \mid n \in \mathbb{Z}, 0 \leq n \leq 20\}$, where n is a grid search parameter to optimize λ.

2) FASHION-MNIST (#2)

The second experiment applies the MIMM model to the Fashion-MNIST dataset. The original Fashion-MNIST [76] dataset consists of 28×28 gray-scale images grouped in ten classes of fashion items. To demonstrate the challenge of confounders in this dataset, we created an artificial confounding by adding different types of boundary bars to each image. Fig. 6 shows the ten different classes with the most often occurring bar per class during training. During training, 95% of the image samples of each class consist of the bar shown in Fig. 6. The remaining 5% samples consist equally of the bar types of the other nine classes.

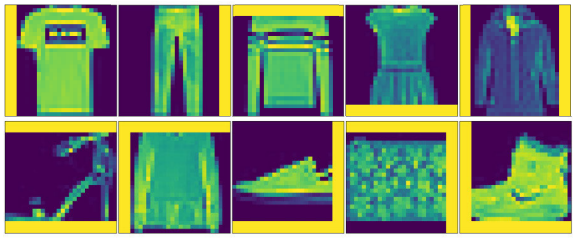


FIGURE 6. Adapted Fashion-MNIST (#2): Selection of Fashion-MNIST samples with the confounding boundary bar of the major classes during training. Images are gray-scale, but to emphasize the bars the visualization is colored.

Settings: The training set consists of a total of 5997 samples per class. The exact class distributions of Y and Z during training are given in Table 2 Experiment 2. MIMM is trained with a batch size of 1000 samples and with $N_B \in \{3, 5\}$ and $\lambda \in \{0.05n \mid n \in \mathbb{Z}, 0 \leq n \leq 20\}$. The architecture of the feature encoder comprises two convolutional layers with 32 and 64 channels each followed by batch normalization and ReLU activation function. Both convolutional layers use a kernel size of 3 and a max-pooling operation with a kernel size of 2. This is followed by three fully-connected layers with 600, 120, and 20 units. The feature vector is of size 20, hence $\boldsymbol{F_Y}$ and $\boldsymbol{F_Z}$ are of size 10. The aim of the primary task Y is to predict the correct fashion label, while Z represents the ten different types of bars.

C. EXPERIMENTS ON MEDICAL DATASETS

The following three experiments are based on medical datasets. All medical experiments and results are reviewed and validated by a radiologist (S.G.) with more than 10 years of clinical experience.

1) UK BIOBANK AND NAKO

The following two experiments are investigated on the UK Biobank (UKB) and the German National Cohort (NAKO). Both datasets collect a large cohort of brain MR images acquired with 3T Siemens Skyra MRI scanners. We used the T1-weighted 3D MPRAGE images with a resolution of $1 \times 1 \times 1$ mm^3 of the brain. In this study, we extract the middle slice of each image and create a two-dimensional input for our model with a matrix size of 256×256 voxels by cropping or padding the images to this dimension. All images are pre-processed by zero mean and unit variance scaling.

a: PREDICTION OF SEX FROM BRAIN MRI (#3)

The primary task Y of this experiment is to predict the sex of a given subject based on its anatomical features in brain MRI and without relying on the spurious correlation created by the scanner differences of the two acquisition sites UKB and NAKO.

Settings: The training dataset comprises 8326 samples from both acquisition sites. Thereby, the exact class

distribution between female/male samples from UKB/NAKO is given in Table 2 Experiment 3. The applied feature encoder architecture consists of four 3×3 convolutional layers with channel sizes of 16, 32, 64, and 32, respectively. Each convolution output is processed by batch normalization, ReLU activation max-pooling with a kernel of size 2. After the convolutional layers, three fully-connected layers with 600, 120, and 4 units are used to create the feature vector. The training of MIMM is performed with a batch size of 420 and with $N_B \in [3, 5]$ and $\lambda \in [0.3, 0.5]$.

b: PREDICTION OF AGE GROUP FROM BRAIN MRI (#4)

This experiment aims to predict the age group of subjects from the brain MRI. The patients are split into two groups, young ($\leq$51 years) and old ($\geq$57 years). Both groups comprise an equal amount of MRI data from UKB and NAKO.

Settings: In total, the training was conducted with 5340 subjects, the exact class distribution between young/old and female/male subjects is given in Table 2 Experiment 4. The architecture and training parameters remain consistent with those employed in the first medical experiment.

2) ALZHEIMER's DISEASE NEUROIMAGING INITIATIVE

a: PREDICTION of ALZHEIMER's DISEASE FROM BRAIN MRI (#5)

This experiment utilized the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and focused on analyzing the T1-weighted 3D brain MRI with a resolution of $1 \times 1 \times 1\text{mm}^3$ acquired by scanners of the manufacturers GE or Siemens. The aim of this experiment is the prediction of Alzheimer's Disease (AD) and healthy controls (HC). The data is spuriously correlated by the manufacturer which is either GE or Siemens. To generate samples for this experiment, pseudo-3D MRIs were created by cropping or padding each dimension of the six middle slices to a dimension of 256, resulting in a tensor shape of $6 \times 256 \times 256$. Each tensor is normalized by zero mean and unit variance scaling. To increase the number of training samples, data augmentation technique of random affine transformation with rotation degrees of -5 to 5, translation of 0.1 to 0.2 in both directions, and scaling factors ranging from 0.8 to 1.2 was applied to the dataset.

Settings: The original training dataset consists of 1718 samples. Including the augmented samples, the training dataset is increased to 5154 samples. The exact composition of the classes HC and AD confounded by the type of manufacturer GE and Siemens during training is given in Table 2 Experiment 5. It is composed of 2322 HC and 255 AD of Siemens scanners as well as 2322 AD and 255 HC of GE scanners. The feature encoder shares the same architecture as the feature encoder of the experiments on UKB and NAKO. However, due to limited computation capacity, the batch size was decreased to 280. The training of MIMM is conducted with $N_B \in [3, 5]$ and $\lambda \in [0.3, 0.5]$.

D. METHODS FOR COMPARISON

The proposed method is compared with the previously introduced approaches. The reference methods are shortly described in the following.

Baseline: We refer to baseline as a network that utilizes the same feature encoder as the proposed MIMM model but does not incorporate the MI estimation as loss penalty. The baseline model is capable of predicting both Y and Z.

Rebalancing: Instead of training the model on the imbalanced class, rebalancing generates balanced classes by sampling with replacement from the minor represented spuriously correlated factor class of each primary task class. The model architecture is equivalent to the baseline architecture.

Metadata Normalization (MDN): Metadata Normalization, a method that was previously described in [47], removes the metadata $M\beta$, which is the confounding, from the features $\boldsymbol{F}$ after each layer and outputs the corrected features R. Mathematically, it can be described as a general linear model as in (9), where β is an unknown set of linear parameters.

$$\boldsymbol{R} = \text{MDN}(\boldsymbol{F}; \boldsymbol{M}) = \boldsymbol{F} - \boldsymbol{M}\beta_M \tag{9}$$

This type of model is not able to predict Z, therefore, we only use it to compare the results of the primary task Y. See [47] for more details.

EnD: (Entangeling and Disentangeling) is a regularization strategy proposed in [49]. The strategy involves minimizing a joint loss function that incorporates both the standard cross-entropy loss and an entangling and disentangling loss. In this context, entangling refers to the correlation of extracted feature vectors belonging to the same primary task, whereas disentangling refers to the ability to separate the extracted feature vectors from the unwanted confounding. Similarly to MDN, this architecture does not allow the prediction of Z. Hence, we only compare the results of the primary task Y. See [49] for more details.

VI. RESULTS AND DISCUSSION

This chapter provides the results and analysis derived from our experiments. Firstly, we present the outcomes on the non-medical benchmark datasets, namely Morpho-MNIST and Fashion-MNIST, followed by the results obtained from our medical experiments. The chapter concludes with an overall discussion that synthesizes the key findings and highlights the implications and future work of our research. In addition, we conducted further experiments on the BiasedMNIST database [50]. These experiments aim to compare our proposed method with recently introduced methods for confounder-free learning. The complete details of the experiment, including the obtained results, can be found in the Supplementary Material.

A. EXPERIMENTS ON BENCHMARK DATASETS

1) MORPHO-MNIST

The models that obtained the best performance over the 5-fold cross-validation are presented in Table 3. For MIMM, the best performing model uses the hyperparameters

$N_B = 3$ and $\lambda = 0.55$. Although all reference methods exhibit a high increased accuracy on the validation set, which shares the same distribution as the training set, our method MIMM achieves the highest 94.0% for the more critical test set with an inverted distribution as the training set. MIMM demonstrates a minimal difference of 1% in accuracy between the validation and test set, whereas the rebalancing, MDN, EnD, and baseline methods experience a significant decrease of 12.7%, 34.1%, 7.9% and 16.4%, respectively. This signifies that these models lack the ability to generalize in the presence of spurious correlation to different writing styles. Moreover, to demonstrate the independence between $\boldsymbol{F_Y}$ and $\boldsymbol{F_Z}$, we conduct predictions on the balanced dataset using the opposed label, i.e., we predict Z given $\boldsymbol{F_Y}$ and Y given $\boldsymbol{F_Z}$. Thereby, an accuracy of 50% indicates that one subvector does not convey information about the other subvector. As shown in Table 3, the accuracy of $\boldsymbol{F_Z}$ predicting Y is close to random for all methods. This means, $\boldsymbol{F_Z}$ indeed does not contain any information about Y. However, our proposed model comes closest to randomness with an accuracy of about 50%. For the prediction of the writing style Z with $\boldsymbol{F_Y}$, MIMM reaches also an accuracy close to random with 49.9%, while the other methods are apart from random by 4.6% (Rebalancing), 19.1% (MDN), 4.3% (EnD), and 9.2% (Baseline). Thus, the subvectors $\boldsymbol{F_Y}$ of the reference methods remain dependent as they contain information about the writing style Z. The t-SNE plot in Fig. 7 shows the feature vector component $\boldsymbol{F_Y}$ colored by its labels of Z. The t-SNE plots of the reference models show clusters of the thin (red) and thick (blue) writing styles. This reveals that $\boldsymbol{F_Y}$ still inherits information of the writing style Z as already demonstrated by the prediction $\boldsymbol{F_Y} \rightarrow Z$ in Table 3. However, it is desired that $\boldsymbol{F_Y}$ is free from information about the writing style. This is given for MIMM, since $\boldsymbol{F_Y}$ is free from any writing style information, i.e. the Z classes, thin and thick, are not separable anymore.

To compare the actual estimated MI between both $\boldsymbol{F_Y}$ and $\boldsymbol{F_Z}$, we take the trained feature encoder of the rebalancing, baseline, and MIMM model, freeze it, and train a randomly initialized MINE model with the same architecture as in the MIMM model for 1000 epochs. As mentioned above, a smaller value of MI indicates less information sharing between the feature vector subvectors. The MINE estimation training is shown in Fig. 14a) of the AppendixVII. The MI of MIMM converges to the smallest value with approx. 0.45 while the baseline model converges above 0.9 and the rebalancing model reaches a value of approx. 1.1. This emphasizes the above results as it shows that MIMM is able to ignore the spurious correlation and focuses on task-relevant features.

To conclude, these results demonstrate that the MIMM model trained on the Morpho-MNIST dataset is robust to changes in the writing style Z and shows counterfactual invariant predictions.

TABLE 3. Morpho-MNIST (#1): Classification accuracy in percentage of digit group Y (small/high) and writing style Z (thin/thick). a) Val. refers to the validation dataset consisting of unseen samples with the same class distribution as the training dataset. b) Test refers to the test dataset consisting of unseen samples with a flipped distribution as the training dataset. c) Bal. refers to a balanced dataset with unseen samples. The accuracy of $F_Y \rightarrow Z$ (estimating writing style Z from F_Y) and $F_Z \rightarrow Y$ (estimating digit group Y from F_Z) are better if close to a random guess (50%).

	Y (small/high digits) Datasets			Z (thin/thick) Datasets		
Model	a)Val.	b)Test	c)Bal.	a)Val.	b)Test	c)Bal.
Rebalancing	98.3	85.6	92.3	**99.9**	99.6	99.7
MDN	97.1	63.0	79.9	-	-	-
EnD	98.9	91.0	**94.3**	-	-	-
Baseline	**98.3**	81.9	90.3	99.7	99.5	99.7
MIMM	93.0	**94.0**	93.2	99.8	**99.7**	**99.8**
Model	$F_Y \rightarrow Z$			$F_Z \rightarrow Y$		
Rebalancing	54.6			49.4		
MDN	69.1			-		
EnD	54.3			-		
Baseline	59.2			50.2		
MIMM	**49.9**			**50.0**		

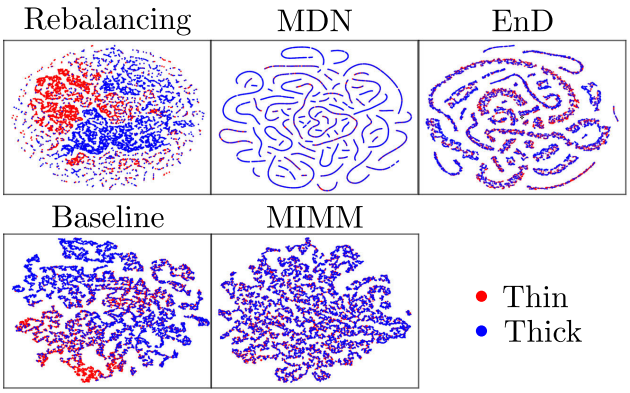


FIGURE 7. Morpho-MNIST (#1): t-SNE visualization of the feature vector component F_Y colored by its Z class labels, thin (red) and thick (blue). F_Y is the FV component that should contain only information about the digit group, small or high. It demonstrates the contained feature information of Z in F_Y. Hence, if the classes thin and thick are not separable anymore, the independence of F_Y from Z is demonstrated. While the t-SNE plots of the reference methods still allow a fuzzy separation between thin and thick (i.e. the confounded variable), the separation is impossible for our proposed model MIMM.

2) FASHION-MNIST

This experiment aims to predict the fashion item Y as well as the type of the boundary bar Z as shown in the samples of Fig. 6. The results of the best performing models of this experiment are shown in Table 4. Predicting the type of bar is distinctively the easier task to solve as the prediction of Z achieves approx. 100% accuracy for all models and distribution shifts. Thus, this bears the risk that the model learns patterns based on the spurious correlation instead of the shape of the fashion item. The MIMM model trained with $N_B = 5$ and $\lambda = 0.3$ yields the highest test accuracy of the fashion items Y with 79.9%. The results of the MIMM's prediction

TABLE 4. Adapted Fashion-MNIST (#2): Comparison of classification accuracy of Y, fashion item, and Z, type of bar. Accuracy is shown in percentage. The accuracy of $F_Y \rightarrow Z$ (estimating type of bar from F_Z) and $F_Z \rightarrow Y$ (estimating fashion type from F_Y) are better if close to random guess (10%) since this tests the prediction of the opposed label.

	Y (fashion item) Datasets			Z (type of bar) Datasets		
Model	a)Val.	b)Test	c)Bal.	a)Val.	b)Test	c)Bal.
Rebalancing	84.2	78.5	80.6	99.8	100.0	100.0
MDN	95.8	33.1	45.7	-	-	-
EnD	**98.7**	78.5	79.5	-	-	-
Baseline	98.5	73.4	76.1	100.0	100.0	100.0
MIMM $\lambda = 0.3$	97.1	**79.9**	**82.1**	100.0	100.0	100.0
MIMM $\lambda = 0.5$	86.1	78.3	79.2	100.0	100.0	100.0

Model	$F_Y \rightarrow Z$	$F_Z \rightarrow Y$
Rebalancing	18.3	10.0
MDN	46.9	-
EnD	21.0	-
Baseline	24.4	10.0
MIMM $\lambda = 0.3$	16.6	10.0
MIMM $\lambda = 0.5$	**12.0**	10.0

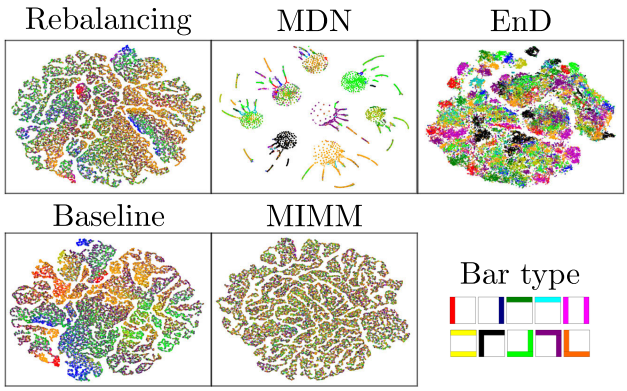


FIGURE 8. Adapted Fashion-MNIST (#2): t-SNE visualization of the feature vector component F_Y and colored by its bar type labels Z. F_Y should contain only information about the fashion item. Hence, if the Z labels (i.e. the confounded variable) are not separable anymore, the independence of F_Y from Z is demonstrated. The reference methods still allow some degree of separation of the bar types F_Y. Only the MIMM model is able to fully remove the influence of Z on F_Y and does not show any colored bar type clusters.

of Z using $\boldsymbol{F_Y}$ with $\lambda = 0.3$ deviates from random guess by 6.6%. This suggests that $\boldsymbol{F_Y}$ still contains some information related to the bar types Z. The MIMM model, which was trained with $N_B = 5$ and $\lambda = 0.5$, shows a decrease in test accuracy of 1.6%, but it is found to be with 12.0% closer to a random guess when predicting Z using $\boldsymbol{F_Y}$. This is logical because a larger value of lambda gives more emphasis to the MI regularisation term MI($\boldsymbol{X}$) in equation (6) and further reduces the dependence between $\boldsymbol{F_Y}$ and $\boldsymbol{F_Z}$. Among the reference methods, rebalancing and EnD demonstrated the highest accuracy of 78.5% on the test set. Their accuracy for $\boldsymbol{F_Y}$ in predicting Z is further from random guess than the two MIMM models, with a distance of 8.3% and 11.0% respectively. Likewise, the t-SNE plots in Fig. 8 which show the feature vector component $\boldsymbol{F_Y}$ colored by the bar type labels $\boldsymbol{F_Z}$, display a certain level of separable bar type clusters for rebalancing as well as for MDN, EnD, and the baseline. In the t-SNE plot of MIMM, distinct clusters are not discernible for the bar types. This demonstrates, as desired, that the feature vector component $\boldsymbol{F_Y}$ does not contain information of the bar type Z. Furthermore, in Fig. 14b) the estimated MI value increases tremendously even after 500 epochs of training for rebalancing and baseline, while the value of MI for MIMM is close to 0.

Based on these results, it can be inferred that MIMM mitigates the impact of the spurious correlation given by the bar types and performs better than the reference methods in predicting counterfactual samples.

B. EXPERIMENTS ON MEDICAL DATASETS

1) PREDICTION OF SEX IN BRAIN MRI FROM UK BIOBANK AND NAKO

The best results of MIMM for the prediction of sex and the different acquisition sites, UK Biobank and NAKO, from brain MRIs are achieved by training the model with $N_B = 5$ and $\lambda = 0.5$. As the accuracy of the methods in Table 5 reveal, it is easier for the model to learn to predict the acquisition site Z than the sex of the subjects since the accuracy of Z is close to 100% for validation and test distributions. This bears the risk to use spuriously correlated factors, i.e. acquisition site to predict sex rather than learning anatomical features that are causally linked to sex. The fact that the reference methods learn the spurious correlation rather than the causal relationship is reflected in Figure 10. While unseen samples from the major classes during training, female brains from UK Biobank and male brains from NAKO, are correctly predicted (Fig. 10a)), the samples from the underrepresented class (Fig. 10b)) are only correctly predicted by the proposed MIMM model. This demonstrates the counterfactual invariance of MIMM. The baseline and MDN models are unable to predict the subject's sex based on its anatomical features as the accuracy between validation and test set decreases by approx. 15%. While EnD achieves on the validation dataset the highest accuracy with 98.7%, our model achieves equal accuracy on the test set. EnD and rebalancing have a slightly higher accuracy on the prediction of Y on the balanced dataset. However, as the results of the opposed prediction, $\boldsymbol{F_Y}$ on Z and $\boldsymbol{F_Z}$ on Y, disclose, $\boldsymbol{F_Y}$ and $\boldsymbol{F_Z}$ are independent since the performances for MIMM are close to random guess with 50%. Compared to the other methods, MIMM is closest to random guess.

The t-SNE plots in Fig. 9 highlight the independence of $\boldsymbol{F_Y}$ from features related to the acquisition site for the proposed MIMM model. The feature vector subvector $\boldsymbol{F_Y}$ does not allow any separation between the acquisition sites, UKB (red) and NAKO (blue). The other reference methods still allow a separation of the acquisition sites. This means, $\boldsymbol{F_Y}$ still contains information about the acquisition site Z.

When training the MINE model after freezing the feature encoder, the resulting estimated MI between the feature vector components of the trained comparison models on sex and

TABLE 5. Brain MRI of UK Biobank and NAKO - sex and acquisition site (#3): Comparison of classification accuracy of sex Y and acquisition site (UK Biobank and NAKO) Z, on brain MRI. Accuracy is shown in percentage. The accuracy of $F_Y \rightarrow Z$ (estimating acquisition site from F_Y) and $F_Z \rightarrow Y$ (estimating sex from F_Z) are better if close to random guess (50%).

	Y (sex) Datasets			Z (acquisition site) Datasets		
Model	a)Val.	b)Test	c)Bal.	a)Val.	b)Test	c)Bal.
Rebalancing	95.0	**94.5**	94.8	**100**	99.8	99.9
MDN	98.3	86.9	92.7	-	-	-
EnD	**98.7**	**94.5**	**96.6**	-	-	-
Baseline	98.1	83.0	90.8	**100**	99.9	99.9
MIMM	94.4	**94.5**	94.5	**100**	**100**	**100**

Model	$F_Y \rightarrow Z$	$F_Z \rightarrow Y$
Rebalancing	55.6	55.1
MDN	56.9	-
EnD	52.3	-
Baseline	63.4	55.1
MIMM	**50.0**	**50.1**

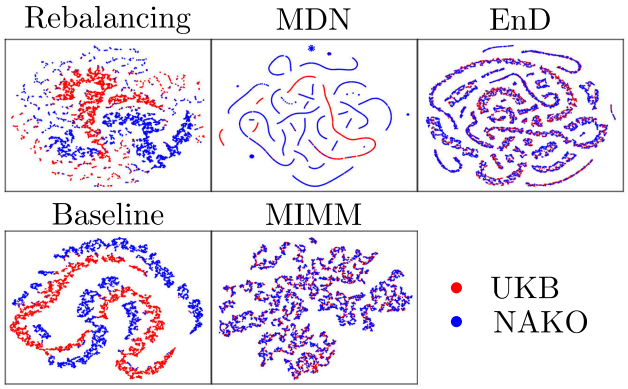


FIGURE 9. Brain MRI of UK Biobank and NAKO - sex and acquisition site (#3): t-SNE visualization of the feature vector component F_Y and colored by the labels of Z, UKB (red) and NAKO (blue). F_Y is the FV component that should only contain information about the anatomical features related to sex within the brain MRI. While the t-SNE plots of the reference methods still allow a separation between UKB and NAKO, for MIMM the separation is, as desired, impossible.

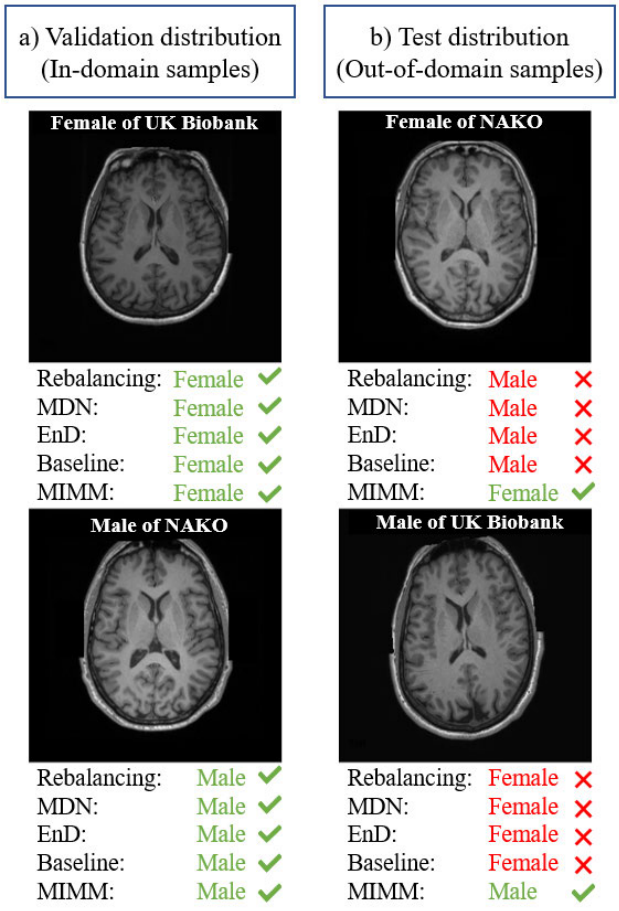


FIGURE 10. Prediction results Brain MRI of UK Biobank and NAKO - sex and acquisition site (#3): (Left) Unseen samples of the major classes of the validation set (female brain from UK Biobank, male brain from NAKO) are predicted correctly for all methods. (Right) Unseen samples of the major class of the test set (female brain from NAKO, male brain from UK Biobank) are only correctly predicted by MIMM.

acquisition site share more information than the feature vector parts of MIMM as MIMM's MI value converges around 0.4. In comparison, the baseline converges to a MI value of approx 1.2 and rebalancing to approx. 0.8. The training of MINE is shown in Fig. 14c) of the AppendixVII. To sum up the given results, our proposed model MIMM is robust to shifts in distribution and let us conclude that the trained MIMM model is counterfactual invariant.

2) PREDICTION OF AGE GROUPS IN BRAIN MRI FROM UK BIOBANK AND NAKO

Table 6 shows the best results of the trained models on the age group, young (≤51 years) and old (≥57 years), and sex. The prediction accuracy of sex Z yielded almost 100% for all methods which indicates that predicting the sex is the easier task. However, it also reveals that the prediction of the age group might be based on the spuriously correlated pattern created by sex, as this is easier to learn than the anatomical features that are causally related to the age group. While rebalancing achieves on the validation dataset the highest accuracy for both Y and Z, its accuracy decreases by 43% for Y and 2.1% for Z on the test set. MIMM achieves 88% on the validation set for Y and drops by only 22% on the test set. In case of the prediction of sex, MIMM has a slight drop in accuracy of 0.1%. The other methods of comparison, MDN, EnD, and Baseline, exhibit a significant decrease in accuracy of approx. 40% and higher when transitioning from validation to test distribution. This is a distinct indication, that MIMM performs best on distribution changes. On the balanced dataset, rebalancing and MIMM achieve both an accuracy of 97.4% for the predicted sex. Nevertheless, MIMM demonstrates superior performance in predicting the age group from brain MRI scans with a 5% improvement in accuracy in comparison to rebalancing. Additionally, the results of the inverted prediction of $\boldsymbol{F_Y}$ with Z labels and $\boldsymbol{F_Z}$ with Y labels demonstrate that the MIMM model discriminates the features for Y and Z more effectively compared to the other methods. This is highlighted in the t-SNE plots

TABLE 6. Brain MRI of UK Biobank and NAKO - age group and sex (#4): Comparison of classification accuracy of age group Y - young ($\leq$ 51 years) and old ($\geq$ 57 years) - and sex Z on brain MRI. Accuracy is shown in percentage. The accuracy of $F_Y \rightarrow Z$ (estimating sex from F_Y) and $F_Z \rightarrow Y$ (age group from F_Z) are better if close to random guess (50%) since this tests the prediction of the opposed label.

	Y (age group) Datasets			Z (sex) Datasets		
Model	a)Val.	b)Test	c)Bal.	a)Val.	b)Test	c)Bal.
Rebalancing	**95.2**	52.2	72.0	**98.5**	96.4	**97.4**
MDN	90.5	41.9	63.9	-	-	-
EnD	91.2	52.3	72.4	-	-	-
Baseline	91.1	46.9	68.8	97.8	94.2	95.0
MIMM	88.0	**66.3**	**77.1**	97.5	**97.4**	**97.4**

Model	$F_Y \rightarrow Z$	$F_Z \rightarrow Y$
Rebalancing	75.7	53.9
MDN	80.2	-
EnD	74.2	-
Baseline	77.2	52.1
MIMM	**63.5**	**50.0**

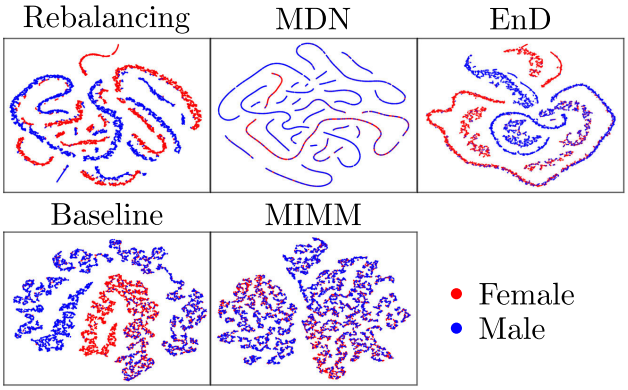


FIGURE 11. Brain MRI of UK Biobank and NAKO - age group and sex (#4): t-SNE visualization of the feature vector component F_Y, colored by the labels of Z, female (red) and male (blue). F_Y is the feature vector component that should only contain information about the anatomical features related to age. If female and male (i.e. the confounded variable) are not separable anymore, the independence of F_Y from Z is demonstrated. This is only given for the proposed MIMM model.

in Fig. 11, where a separation of the Z classes, female and male, from $\boldsymbol{F_Y}$ is still recognizable for the reference methods, whereas MIMM removed most of the information of Z from $\boldsymbol{F_Y}$. Correspondingly, the estimated MI (<0.5) of the MIMM model is smaller than the MI values of the baseline ($\approx$1) and rebalancing with ($\approx$3.1) as shown in Fig. 14d). This is an additional demonstration of the removal of MI in $\boldsymbol{F_Y}$ and $\boldsymbol{F_Z}$.

Given these results, the spurious correlation between the age group and sex is decreased by MIMM and this makes the prediction compared to the other methods more robust to distribution shifts and demonstrates its counterfactual invariance.

3) PREDICTION OF ALZHEIMER's DISEASE IN BRAIN MRI FROM ADNI

This experiment aimed to differentiate between healthy controls (HC) and Alzheimer's Disease (AD) using pseudo-3D brain MRI with six slices. Its results are displayed in Table 7. The training data contains a spurious correlation given by the scanner, which is either from the manufacturer Siemens or GE.

Changing the data distribution from validation to testing leads to decreased accuracy for all methods. Especially, the baseline, MDN, and EnD methods show an accuracy below 50%, clearly indicating a learned shortcut instead of the causal relationship between anatomical brain differences and Alzheimer's Disease. When evaluating the model on a balanced dataset, MIMM performs best with 67.7%. In general, we need to acknowledge that limited data and limited computational capacity, prevented our research from using deeper architectures and larger mini-batches to further improve the AD prediction. Nevertheless, the prediction of the scanner's manufacturer is not only the highest for MIMM but also with 97.4% on the balanced dataset close to 100 %. The reference methods achieved only about 93.4% on the same dataset.

Although rebalancing is trained on more samples due to resampling with replacement, it is not able to remove the influence of the spuriously correlated factor in the same way as MIMM. This is not only shown by the inverse estimation ($\boldsymbol{F_Y} \rightarrow Z$ and $\boldsymbol{F_Z} \rightarrow Y$, but also by the t-SNE plots in Fig. 12. For rebalancing, MDN, EnD, and baseline, clusters of the type of scanner are clearly distinguishable from the feature vectors $\boldsymbol{F_Y}$ in their respective plots. However, a distinction between the scanners is much more difficult for MIMM.

Fig. 13a) provides an exemplary illustration where samples from the major class during training, namely healthy controls by Siemens and Alzheimer's by GE, are correctly predicted by all models. However, examples from the underrepresented class during training (Fig. 13b)), namely healthy controls by GE and Alzheimer's acquired by Siemens, are only correctly predicted by our MIMM model.

By training the MINE model with the features of the trained feature encoder to investigate the behavior of MI between $\boldsymbol{F_Y}$ and $\boldsymbol{F_Z}$, we found that the baseline and rebalancing are converging around 2.2 and 1.2, respectively, whereas the MI of MIMM is close to 0.1. This emphasizes that MIMM is the most successful at avoiding the spurious correlation. The MINE training curve of this experiment is shown in Fig. 14e) of the Appendix VII.

C. DISCUSSION

In this paper, we address the challenge of spurious correlation on non-medical benchmark and medical datasets. The complexity of confounding in databases is tremendous and threatens the robustness of deep learning-based prediction which is especially crucial for the future application of deep learning algorithms in clinical workflows. The introduction of our novel method MIMM demonstrates the feasibility of suppressing the occurrence of spurious correlations generated by a known confounder. We evaluated the method on non-medical benchmark and medical datasets and compared it with a baseline model without MI estimation, with MDN,

TABLE 7. Brain MRI of ADNI - Alzheimer's Disease and manufacturer (#5): Comparison of accuracy of the prediction of patient type Y (healthy controls vs. Alzheimer's Disease), and scanner Z (GE vs. Siemens). Accuracy is shown in percentage. The accuracy of $F_Y \rightarrow Z$ (estimating the manufacturer from Alzheimer's Disease class feature vector) and $F_Z \rightarrow Y$ (estimating Alzheimer's Disease from manufacturer class feature vector) are better if close to random guess (50%) since this tests the prediction of the opposed label.

	Y (patient type) Datasets			Z (scanner) Datasets		
Model	a)Val.	b)Test	c)Bal.	a)Val.	b)Test	c)Bal.
Rebalancing	67.9	51.4	65.0	92.1	88.6	93.3
MDN	75.9	45.8	61.2	-	-	-
EnD	90.2	32.7	61.5	-	-	-
Baseline	**87.5**	28.5	58.0	95.9	91.0	93.4
MIMM	82.8	**53.0**	**67.7**	**98.8**	**97.0**	**97.4**

Model	$F_Y \rightarrow Z$	$F_Z \rightarrow Y$
Rebalancing	65.2	54.3
MDN	69.1	-
EnD	86.2	-
Baseline	86.8	52.8
MIMM	**60.3**	**50.9**

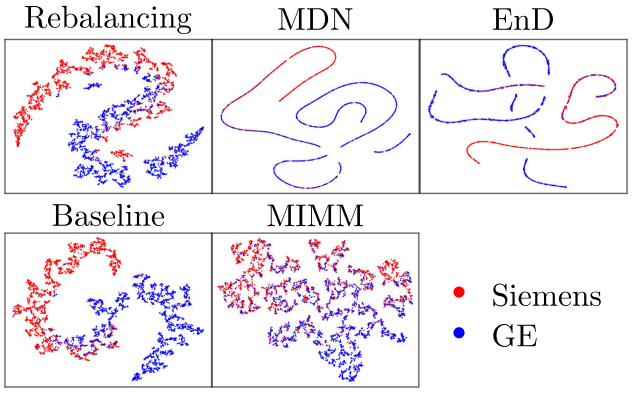


FIGURE 12. Brain MRI of ADNI - Alzheimer's Disease and manufacturer (#5): t-SNE visualization of the feature vector component F_Y, colored by the labels of Z, Siemens (red) and GE (blue). F_Y is the feature vector component that should only contain information about the anatomical features related to the manufacturer. If Siemens and GE (i.e. the confounded variable) are not separable anymore, the independence of F_Y from Z is demonstrated. This is only given for the proposed MIMM model.

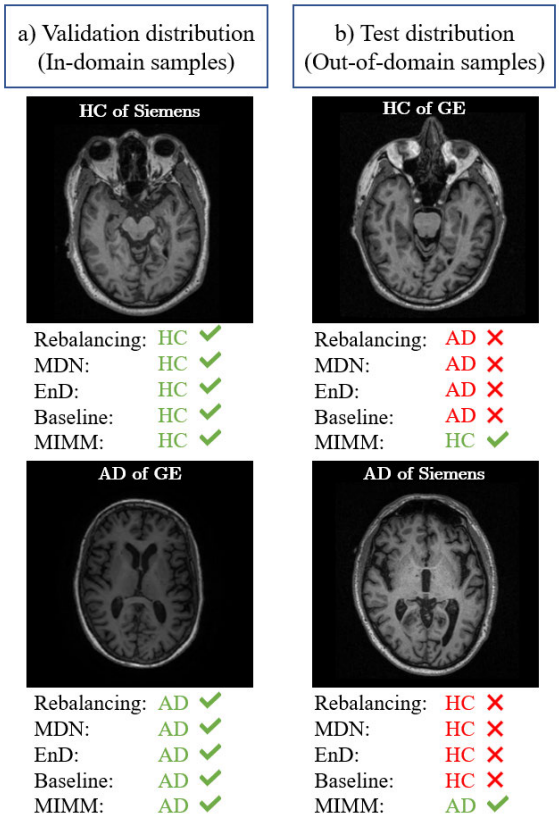


FIGURE 13. Prediction example Brain MRI of ADNI - Alzheimer's Disease and manufacturer (#5): (Left) Unseen samples of the major classes of the validation set(healthy control (HC) from Siemens scanner, Alzheimer's subject (AD) from GE scanner) are predicted correctly for all methods. (Right) unseen samples of the major class of the test set (HC from GE scanner, AD from Siemens scanner) are only correctly predicted by MIMM.

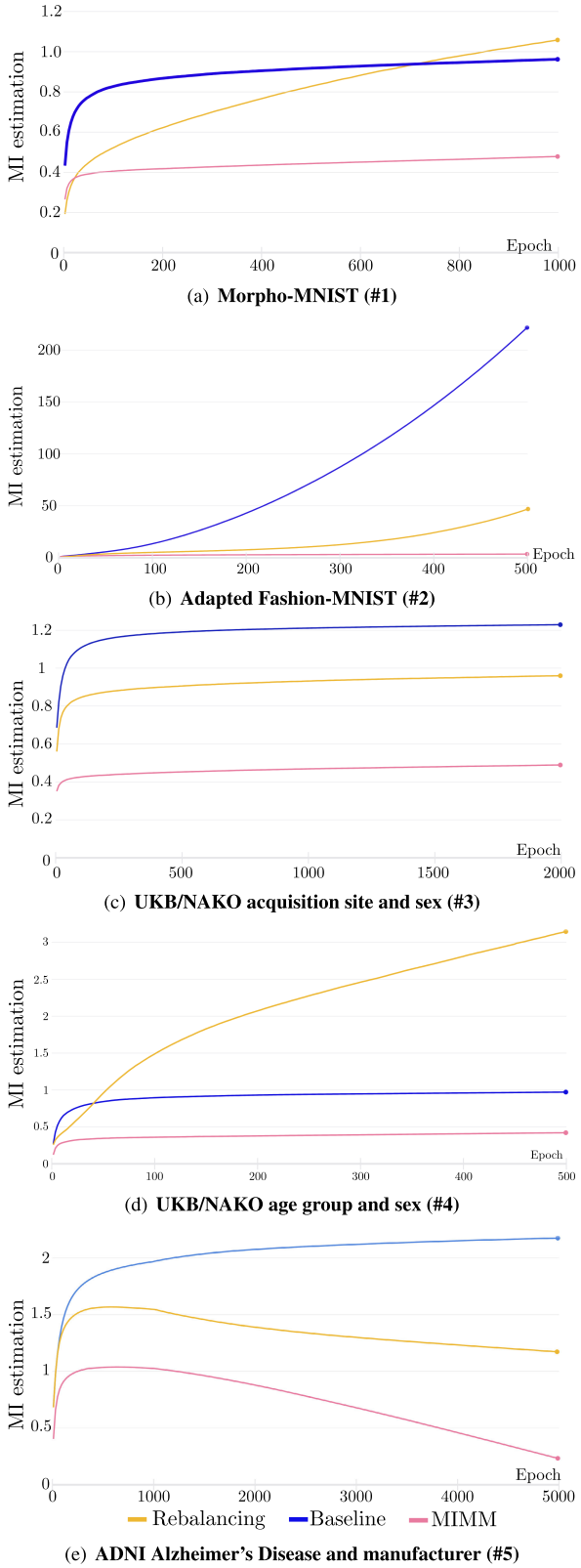
a method that removes the confounding after each DL layer, with EnD, which removes confounding by an additional penalty, and by rebalancing the minor represented class in the given training dataset to train the model on balanced data.

We demonstrated that in all experiments our approach performs best. For most of our experiments, the reference method rebalancing and EnD could provide the closest results to our method.

Nevertheless, when estimating Y with the opposed feature vector $\boldsymbol{F_Z}$ and Z with the opposed feature vector $\boldsymbol{F_Y}$, MIMM provides the closest results to random, which shows us that it learns features without confounding influence. Moreover, by visualizing the learned feature vectors of the primary tasks, colored by the labels of the spuriously correlated factor, MIMM could show for all our experiments that a distinction of the spuriously correlated factor is impossible. Thus, MIMM is a counterfactual invariant predictor that is able to predict the target output based on the causal relationship with the input. In contrast, the presented comparative methods lack the ability to completely remove the influence of the spuriously correlated factor in all of our experiments. This leads to the conclusion that these methods fail to perform as counterfactual invariant predictors.

By inverting the dataset distribution between the validation and test datasets, we also showed that MIMM is robust to data distribution shifts and out-of-domain data as it showed the highest performance on the test dataset in all experiments. By studying different types of datasets with two-dimensional and three-dimensional acquisitions, we could show the simplicity of adapting MIMM for different tasks.

We acknowledge the following limitations of our study. These limitations will be addressed in our future work. While we have demonstrated that the MIMM approach performs well with non-binary tasks, our experiments have focused solely on the case of a priori known single confounding. Nevertheless, we note that the architecture of the MIMM model allows for a simple extension to accommodate multiple confounding factors. Adapting the mutual information (MI)



(a) **Morpho-MNIST (#1)**

(b) **Adapted Fashion-MNIST (#2)**

(c) **UKB/NAKO acquisition site and sex (#3)**

(d) **UKB/NAKO age group and sex (#4)**

(e) **ADNI Alzheimer's Disease and manufacturer (#5)**

FIGURE 14. **Training of MINE model to estimate the MI between F_Y and F_Z. The desired output of a small MI, which indicates less common information between the feature vector components, is best for the MIMM model.**

estimation process to multiple primary tasks and/or multiple confounding factors presents a more significant challenge, as the question of how to estimate the MI between the primary task and each spuriously correlated factor must be addressed. Extending our work to handle multiple and diverse (binary, multi-class, regression) confounding factors is an important direction for future work. In addition, in our future study, we will investigate the potential benefits of utilizing imbalanced feature subvector sizes, as opposed to our current approach of using equally split feature subvectors.

Moreover, the identification of unknown confounding factors will be scrutinized with further investigations on causal discovery algorithms. Currently, we assume to know the spuriously correlated factors and even their ground truth in the used datasets. The avoidance of shortcut learning will be more challenging if we know the spuriously correlated factors but don't have their ground truth (hidden or lost metadata like scanner type, acquisition site, and even sex and age) or if not all spuriously correlated factors have been identified.

Generally, in our study, we focused on basic deep learning architectures, especially due to limited computational capacity, in order to demonstrate enhanced performance and improved robustness. Consequently, we expect that our results can be further optimized using more advanced and sophisticated architectures, more data, and more computation power.

VII. CONCLUSION

In conclusion, we proposed and evaluated a framework, called Mutual Information Minimization Model (MIMM), which allows to minimize spurious correlation learned by DL models from non-medical benchmark and medical databases. These correlations can lead to incorrect predictions on counterfactuals or in new environments, such as when the model is applied to out-of-distribution data. By separating the features used for the primary task prediction from those of the spuriously correlated factor, MIMM learns causal relationships, which results in more precise and trustworthy predictions. Our extensive experiments demonstrate the effectiveness of this approach, making it a promising approach to prevent shortcut learning in real-world medical imaging analysis.

APPENDIX

The supplemental graphs in Fig. 14 show the behavior of the estimated MI when training a randomly initialized MINE model with the feature vectors from the trained feature encoders for all experiments.

ACKNOWLEDGMENT

This project was conducted with data from federal states and the Helmholtz Association, with additional financial support from the participating universities and institutes of the Leibniz Association. The authors thank all participants who

took part in the GNC study and the staff in this research program.

This work was carried out under U.K. Biobank Application 40040. They also thank all participants who took part in the UKB study and the staff in this research program.

ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd., and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research and Development, LLC.; Johnson & Johnson Pharmaceutical Research and Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This publication was funded by the German Research Foundation (DFG) grant "Open Access Publication Funding/2023-2024/University of Stuttgart" (512689491).

REFERENCES

[1] G. Hinton, "Deep learning—A technology with the potential to transform health care," *J. Amer. Med. Assoc.*, vol. 320, no. 11, pp. 1101–1102, 2018.

[2] G. Marcus, "Deep learning: A critical appraisal," 2018, *arXiv:1801.00631*.

[3] T. Kustner, J. Pan, C. Gilliam, H. Qi, G. Cruz, K. Hammernik, T. Blu, D. Rueckert, R. Botnar, C. Prieto, and S. Gatidis, "Self-supervised motion-corrected image reconstruction network for 4D magnetic resonance imaging of the body trunk," *APSIPA Trans. Signal Inf. Process.*, vol. 11, no. 1, p. e12, 2022.

[4] T. Kustner, N. Fuin, K. Hammernik, A. Bustin, H. Qi, R. Hajhosseiny, P. G. Masci, R. Neji, D. Rueckert, R. M. Botnar, and C. Prieto, "CINENet: Deep learning-based 3D cardiac CINE MRI reconstruction with multi-coil complex-valued 4D spatio-temporal convolutions," *Sci. Rep.*, vol. 10, no. 1, p. 13710, Aug. 2020.

[5] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated MRI data," *Magn. Reson. Med.*, vol. 79, no. 6, pp. 3055–3071, Jun. 2018.

[6] S. Wang, Z. Su, L. Ying, X. Peng, S. Zhu, F. Liang, D. Feng, and D. Liang, "Accelerating magnetic resonance imaging via deep learning," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 514–517.

[7] D. K. Sharma, M. Chatterjee, G. Kaur, and S. Vavilala, "Deep learning applications for disease diagnosis," in *Deep Learning for Medical Applications With Unique Data*. Amsterdam, The Netherlands: Elsevier, 2022, pp. 31–51.

[8] M. Khojaste-Sarakhsi, S. S. Haghighi, S. F. Ghomi, and E. Marchiori, "Deep learning for Alzheimer's disease diagnosis: A survey," *Artif. Intell. Med.*, vol. 130, Aug. 2022, Art. no. 102332.

[9] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J. R. Ledsam, M. K. Schmid, K. Balaskas, E. J. Topol, L. M. Bachmann, P. A. Keane, and A. K. Denniston, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis," *Lancet Digit. Health*, vol. 1, no. 6, pp. 271–297, Oct. 2019.

[10] W. Ren, A. H. Bashkandi, J. A. Jahanshahi, A. Q. M. AlHamad, D. Javaheri, and M. Mohammadi, "Brain tumor diagnosis using a step-by-step methodology based on courtship learning-based water strider algorithm," *Biomed. Signal Process. Control*, vol. 83, May 2023, Art. no. 104614.

[11] S. Foersch, C. Glasner, A. C. Woerl, M. Eckstein, D. C. Wagner, S. Schulz, F. Kellers, A. Fernandez, K. Tserea, M. Kloth, and A. Hartmann, "Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer," *Nature Med.*, vol. 29, pp. 430–439, Jan. 2023.

[12] M. A. Myszczynska, P. N. Ojamies, A. M. B. Lacoste, D. Neil, A. Saffari, R. Mead, G. M. Hautbergue, J. D. Holbrook, and L. Ferraiuolo, "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," *Nature Rev. Neurol.*, vol. 16, no. 8, pp. 440–456, Jul. 2020.

[13] J. Fan, J. Wang, Z. Chen, C. Hu, Z. Zhang, and W. Hu, "Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique," *Med. Phys.*, vol. 46, no. 1, pp. 370–381, Nov. 2018.

[14] J. D. Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, and G. Van Den Driessche, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Med.*, vol. 24, no. 9, pp. 1342–1350, Aug. 2018.

[15] P. Jyothi and A. R. Singh, "Deep learning models and traditional automated techniques for brain tumor segmentation in MRI: A review," *Artif. Intell. Rev.*, vol. 56, no. 4, pp. 2923–2969, Apr. 2023.

[16] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, no. 3, p. 1224, Jan. 2021.

[17] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *J. Digit. Imag.*, vol. 32, no. 4, pp. 582–596, Aug. 2019.

[18] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1562–1573, Jul. 2018.

[19] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[20] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, Apr. 2023.

[21] S. Choi, S. Jung, H. Yun, J. T. Kim, S. Kim, and J. Choo, "RobustNet: Improving domain generalization in urban-scene segmentation via instance selective whitening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11575–11585.

[22] N. Díaz-Rodríguez, R. Binkyte, W. Bakkali, S. Bookseller, P. Tubaro, A. Bacevicius, S. Zhioua, and R. Chatila, "Gender and sex bias in COVID-19 epidemiological data through the lens of causality," *Inf. Process. Manag.*, vol. 60, no. 3, May 2023, Art. no. 103276.

[23] A. Lynch, G. J. S. Dovonon, J. Kaddour, and R. Silva, "Spawrious: A benchmark for fine control of spurious correlation biases," 2023, *arXiv:2303.05470*.

[24] T. M. Mitchell, "The need for biases in learning generalizations," in *Readings in Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1980, pp. 184–191.

[25] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. CVPR*, Jun. 2011, pp. 1521–1528.

[26] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[27] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Mach. Intell.*, vol. 3, no. 7, pp. 610–619, May 2021.

[28] V. Veitch, A. D'Amour, S. Yadlowsky, and J. Eisenstein, "Counterfactual invariance to spurious correlations: Why and how to pass stress tests," 2021, *arXiv:2106.00545*.

[29] S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 456–473.

[30] A. Rao, J. M. Monteiro, and J. Mourao-Miranda, "Predictive modelling using neuroimaging data in the presence of confounds," *NeuroImage*, vol. 150, pp. 23–49, Apr. 2017.

[31] R. Pomponio, G. Erus, M. Habes, J. Doshi, D. Srinivasan, E. Mamourian, V. Bashyam, I. M. Nasrallah, T. D. Satterthwaite, and Y. Fan, "Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan," *NeuroImage*, vol. 208, Mar. 2020, Art. no. 116450.

[32] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 23, pp. 12592–12594, Jun. 2020.

[33] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature Med.*, vol. 27, no. 12, pp. 2176–2182, Dec. 2021.

[34] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 531–540.

[35] B. Schölkopf, "Causality for machine learning," in *Probabilistic and Causal Inference: The Works of Judea Pearl*, 1st ed. New York, NY, USA: Association for Computing Machinery, 2022, pp. 765–804, doi: 10.1145/3501714.3501755.

[36] F. Locatello, B. Poole, G. Ratsch, B. Schölkopf, O. Bachem, and M. Tschannen, "Weakly-supervised disentanglement without compromises," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6348–6359.

[37] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, "CausalVAE: Disentangled representation learning via neural structural causal models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9588–9597.

[38] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021.

[39] M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, and J. Bian, "Causal inference and counterfactual prediction in machine learning for actionable healthcare," *Nature Mach. Intell.*, vol. 2, no. 7, pp. 369–375, Jul. 2020.

[40] A. Subbaswamy and S. Saria, "Counterfactual normalization: Proactively addressing dataset shift and improving reliability using causal mechanisms," 2018, *arXiv:1808.03253*.

[41] J. Peters, P. Buhlmann, and N. Meinshausen, "Causal inference by using invariant prediction: Identification and confidence intervals," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 78, no. 5, pp. 947–1012, Nov. 2016.

[42] X. Wang, M. Saxon, J. Li, H. Zhang, K. Zhang, and W. Y. Wang, "Causal balancing for domain generalization," 2022, *arXiv:2206.05263*.

[43] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," 2019, *arXiv:1907.02893*.

[44] M. Ilse, J. M. Tomczak, and P. Forre, "Selecting data augmentation for simulating interventions," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4555–4562.

[45] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–50, Jun. 2017.

[46] A. Gupta, A. Murali, D. P. Gandhi, and L. Pinto, "Robot learning in homes: Improving generalization and reducing dataset bias," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[47] M. Lu, Q. Zhao, J. Zhang, K. M. Pohl, L. Fei-Fei, J. C. Niebles, and E. Adeli, "Metadata normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10912–10922.

[48] A. Vento, Q. Zhao, R. Paul, K. M. Pohl, and E. Adeli, "A penalty approach for normalizing feature distributions to build confounder-free models," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2022, pp. 387–397.

[49] E. Tartaglione, C. A. Barbano, and M. Grangetto, "EnD: Entangling and disentangling deep representations for bias correction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13503–13512.

[50] H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh, "Learning de-biased representations with biased representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 528–539.

[51] N. Quadrianto, V. Sharmanska, and O. Thomas, "Discovering fair representations in the data domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8219–8228.

[52] W. Zhu, H. Zheng, H. Liao, W. Li, and J. Luo, "Learning bias-invariant representation by cross-sample mutual information minimization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14982–14992.

[53] Y. Hong and E. Yang, "Unbiased classification through bias-contrastive and bias-balanced learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 26449–26461.

[54] S. Jung, D. Lee, T. Park, and T. Moon, "Fair feature distillation for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12110–12119.

[55] Q. Zhao, E. Adeli, and K. M. Pohl, "Training confounder-free deep learning models for medical applications," *Nature Commun.*, vol. 11, no. 1, p. 6010, Nov. 2020.

[56] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, and D. Sculley, "No classification without representation: Assessing geodiversity issues in open data sets for the developing world," 2017, *arXiv:1711.08536*.

[57] S. Roychowdhury, M. Diligenti, and M. Gori, "Regularizing deep networks with prior knowledge: A constraint-based approach," *Knowl.-Based Syst.*, vol. 222, Jun. 2021, Art. no. 106989.

[58] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, no. 6, pp. 1191–1253, Jun. 2003.

[59] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 34–42.

[60] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1931–1939.

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[62] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. Conf. Fairness, Accountability Transparency*, 2018, pp. 77–91.

[63] C. Wachinger, A. Rieckmann, and S. Pölsterl, "Detect and correct bias in multi-site neuroimaging datasets," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101879.

[64] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[65] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, Apr. 2017.

[66] S. Yucer, F. Tektas, N. Al Moubayed, and T. P. Breckon, "Racial bias within face recognition: A survey," 2023, *arXiv:2305.00817*.

[67] T. Quan, F. Zhu, Q. Liu, and F. Li, "Learning fair representations for accuracy parity," *Eng. Appl. Artif. Intell.*, vol. 119, Mar. 2023, Art. no. 105819.

[68] E. Adeli, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, L. Fei-Fei, J. C. Niebles, and K. M. Pohl, "Representation learning with statistical independence to mitigate bias," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2512–2522.

[69] S. Park, J. Lee, P. Lee, S. Hwang, D. Kim, and H. Byun, "Fair contrastive learning for facial attribute classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10379–10388.

[70] T. Wang, J. Zhao, M. Yatskar, K. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5309–5318.

[71] M. Alvi, A. Zisserman, and C. Nellåker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 556–572.

[72] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," 2015, *arXiv:1511.00830*.

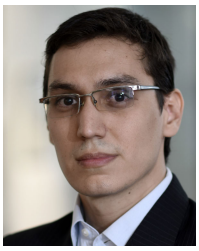
[73] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[74] H. Reichenbach, *The Direction of Time*. New York, NY, USA: Dover, 1956.

[75] D. C. Castro, J. Tan, B. Kainz, E. Konukoglu, and B. Glocker, "MorphoMNIST: Quantitative assessment and diagnostics for representation learning," *J. Mach. Learn. Res.*, vol. 20, no. 178, pp. 1–29, 2019.

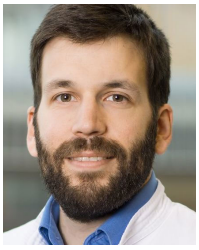
[76] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.

LOUISA FAY received the M.Sc. degree in electrical engineering and information technology with a focus on deep learning from the University of Stuttgart, Germany, in 2020. She is currently pursuing the Ph.D. degree with the Group of Medical Imaging and Data Analysis (MIDAS.lab), University Hospital of Tübingen, Germany, and the Institute for Signal Processing and System Theory, University of Stuttgart. Her research interest includes causality in machine learning with a focus on medical imaging.



ERICK COBOS received the bachelor's and master's degrees from Tec de Monterrey, Mexico, and the M.Sc. degree in intelligent systems, in December 2016. He is currently pursuing the Ph.D. degree with the Max Planck Institute for Intelligent Systems, Tübingen, Germany, under the supervision of Bernhard Schölkopf and Sergios Gatidis. He worked for three years as a Research Associate with the Neuroscience Department, Baylor College of Medicine, Houston, TX, USA. His current research interests include explainability in medical imaging and compositional image representations.

BIN YANG (Senior Member, IEEE) received the Dipl.-Ing. and Ph.D. degrees in electrical engineering in Germany. He is a Full Professor and the Head with the Institute of Signal Processing and System Theory, University of Stuttgart, Germany. His research interests include methods and algorithms of statistical signal processing, machine learning, and in particular deep learning. The research spectrum covers discriminative models, generative models, self-supervised learning, domain adaptation, anomaly detection, and causal reasoning, in connection with a variety of applications like medical imaging, autonomous driving, radar, speech, localization, and semiconductor test.



SERGIOS GATIDIS received the M.D. degree from the University of Tübingen, in 2011, and the M.Sc. degree in mathematics from the University of Hagen, in 2014. In 2017, he was appointed as an Assistant Professor, and in 2020, as an Associate Professor in radiology with the Department of Radiology, University Hospital Tübingen. His research interest includes automated analysis of multiparametric medical image data.

THOMAS KÜSTNER (Member, IEEE) received the Ph.D. degree from the University of Stuttgart, Germany, in 2017. From 2018 to 2020, he was with the School of Biomedical Engineering and Imaging Sciences, King's College London, U.K. Since 2020, he has been co-leading the Medical Imaging and Data Analysis (MIDAS.lab) and got appointed a professorship at the University Hospital of Tübingen, Germany, in 2023. He is the Chair with the MIDAS.lab, University Hospital of Tübingen. His research interests include artificial intelligence-enabled multiparametric and multimodality medical imaging methods in acquisition and reconstruction, and the automated analysis of clinical and epidemiological studies. He is particularly focused on MR-based motion imaging, correction and reconstruction, and the advents of artificial intelligence in MRI. He is a Junior Fellow of ISMRM.