C. Reimers, P. Bodesheim, J. Runge, J. Denzler.: Conditional Adversarial Debiasing: Towards Learning Unbiased Classifiers from Biased Data. DAGM German Conference on Pattern Recognition (DAGM-GCPR) 2021, pp. 48–62. (c) Copyright by Springer.
The final publication is available at https://link.springer.com/chapter/10.1007/978-3-030-92659-5_4.

Conditional Adversarial Debiasing: Towards Learning Unbiased Classifiers from Biased Data

Christian Reimers[1,2][0000−0003−1127−136X], Paul Bodesheim[1][0000−0002−3564−6528], Jakob Runge[2,3][0000−0002−0629−1772], and Joachim Denzler[1,2][0000−0002−3193−3300]

[1] Computer Vision Group, Friedrich Schiller University Jena, 07743 Jena, Germany
[2] Institute of Data Science, German Aerospace Center (DLR), 07745 Jena, Germany
[3] Technische Universität Berlin, 10623 Berlin, Germany
creimers@bgc-jena.mpg.de

Abstract. Bias in classifiers is a severe issue of modern deep learning methods, especially for their application in safety- and security-critical areas. Often, the bias of a classifier is a direct consequence of a bias in the training set, frequently caused by the co-occurrence of relevant features and irrelevant ones. To mitigate this issue, we require learning algorithms that prevent the propagation of known bias from the dataset into the classifier. We present a novel adversarial debiasing method, which addresses a feature of which we know that it is spuriously connected to the labels of training images but statistically independent of the labels for test images. The debiasing stops the classifier from falsly identifying this irrelevant feature as important. Irrelevant features co-occur with important features in a wide range of bias-related problems for many computer vision tasks, such as automatic skin cancer detection or driver assistance. We argue by a mathematical proof that our approach is superior to existing techniques for the abovementioned bias. Our experiments show that our approach performs better than the state-of-the-art on a well-known benchmark dataset with real-world images of cats and dogs.

Keywords: Adversarial Debiasing · Causality · Conditional Dependence.

1 Introduction

Deep neural networks have demonstrated impressive performances in many computer vision and machine learning tasks, including safety- and security-critical applications such as skin cancer detection [15] or predicting recidivism [4]. However, many people and domain experts advise against employing deep learning in those applications, even if classifiers outperform human experts, as in skin lesion classification [20]. One reason for their concerns is bias in the classifiers. Indeed, almost all image datasets contain some kind of bias [22] and, consequently, the performance of classifiers varies significantly across subgroups [4,13].

One major reason for bias in classifiers is dataset bias. Every dataset is a unique slice through the visual world [19]. Therefore, an image dataset do not

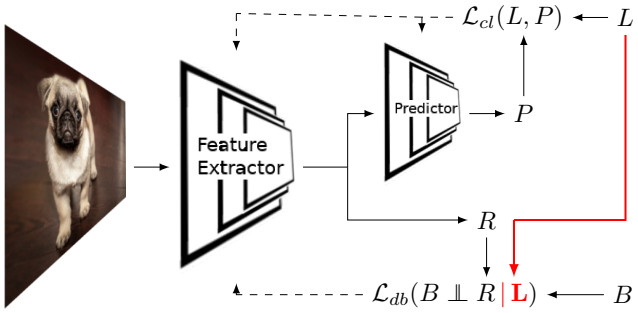

Fig. 1: In adversarial debiasing, a debiasing loss $\mathcal{L}_{db}$ is often used to enforce independence between the bias variable B and a representation R. In this work, we show that it is beneficial to condition this independence on the label L.

represent the real world perfectly but contains unwanted dependencies between meaningless features and the labels of its samples. This spurious connection can be caused by incautious data collection or by justified concerns, if, for example, the acquirement of particular examples is dangerous. A classifier trained on such a dataset might pick the spuriously dependent feature to predict the label and is, thus, biased. Mitigating such a bias is challenging even if the spurious connection is known. To this end, it is important to understand the nature of the spurious dependence. Therefore, we start our investigation at the data generation process. We provide a formal description of the data generation model for a common computer vision bias in Section 3.1. In contrast to other approaches that do not provide a model for the data generation process and, hence, rely solely on empirical evaluations, this allows us to investigate our proposed method theoretically.

The main contribution of our work is a novel adversarial debiasing strategy. The basic concept of adversarial debiasing and the idea of our improvement can be observed in Figure 1. For adversarial debiasing, a second loss $\mathcal{L}_{db}$ is used in addition to the regular training loss $\mathcal{L}_{cl}$ of a neural network classifier. This second loss penalizes the dependence between the bias variable B and an intermediate representation R from the neural network. The main difference we propose in this paper is replacing this dependence $B \not\perp\!\!\!\perp R$ by the conditional dependence $B \not\perp\!\!\!\perp R \,|\, L$ with L being the label. In fact, it turns out that this conditional dependence is better suited than the unconditional dependence for the considered kind of bias. The motivation for this replacement, and a mathematical proof for its suitability can be found in Section 3.2.

To use our new conditional independence criterion for adversarial debiasing, we have to implement it as a differentiable loss. We provide three possible implementations in Section 3.3. We demonstrate that these new loss functions lead to an increase in accuracy on unbiased test sets. In Section 4, we provide results of experiments on a synthetic dataset, a dataset with real-world images of cats and dogs that is used by previous work to evaluate adversarial debiasing, and

an ablation study to show that the proposed change of the criterion causes the increase in accuracy.

2 Related Work

Traditionally, adversarial debiasing aims to learn a feature representation that is informative for a task but independent of the bias. Hence, a second neural network that should predict the bias from the feature representation is introduced to enforce this independence. The original network for classification and this second network are then trained in an adversarial fashion. To this end, different loss functions for the original network are suggested to decrease the performance of the second network for predicting the bias. Previous work aims at minimizing the cross-entropy between bias prediction and a uniform distribution [3] or maximizing the mean squared error between the reconstruction and the bias [23]. Further approaches enclose the joint maximization of the cross-entropy between the predicted and true distribution of the bias variable and the entropy of the distribution of the predicted bias [11] or the minimization of the correlation between the ground-truth bias and the prediction of the bias [1]. However, as shown in another study [17], independence is too restrictive as a criterion for determining whether a deep neural network uses a certain feature. This fact is also reflected in the experimental results of the abovementioned papers. The resulting classifiers are less biased, but this often leads to decreasing performance on unbiased test sets. As one example, significantly less bias in an age classifier trained on a dataset biased by gender has been reported [3], but the classification performance on an unbiased test set decreased from 0.789 to 0.781. Our work is fundamentally different. Instead of a different loss, we suggest a different criterion to determine whether a neural network uses a feature. We use the conditional independence criterion proposed by [17] rather than independence between the representation and the bias.

While the vast majority of adversarial debiasing methods acknowledge that bias has many forms, they rarely link the suggested solutions to the processes that generate the biased data. Instead, they rely exclusively on empirical evaluations. In contrast, we provide a specific model for a specific kind of bias as well as a theoretical proof that our approach is better suited for this case.

3 Proposed Conditional Adversarial Debiasing Approach

In this section, we motivate our novel approach for adversarial debiasing and introduce our novel adversarial debiasing criterion. We prove mathematically that the new criterion fits our specific bias model. Finally, we provide three possible implementations for loss functions that realize this criterion.

3.1 Bias Model

Different kinds of bias influence visual datasets in various ways [22]. We consider a specific kind of bias and the corresponding model covers many relevant tasks

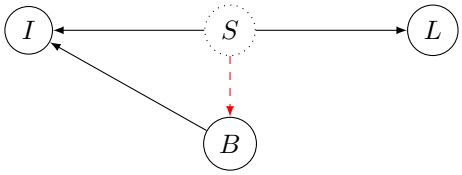

Fig. 2: A graphical representation of the specific bias. Circles represent variables, dotted circles represent unobserved variables. The label L is only dependent on a signal S, while the input I is also dependent on some variable B. In the training set, the signal S influences the variable B due to bias. This is indicated by the red dashed arrow. In contrast, in the test set, the two are independent.

in computer vision. To describe this bias model, we start with a graphical model of the underlying data generation process displayed in Figure 2.

For classification tasks, like separating cats from dogs, we assume that a signal S is contained in and can be extracted from the input I that contains the relevant information. Following the graphical model, a labeling process generates the label L ("cat" or "dog") from this signal S. However, the input I is a mixture of multiple signals. Besides S, another signal B influences I. In the cat/dog example, B might relate to the furs color. Since the furs color is not meaningful for distinguishing cats from dogs, B is independent of S and L during the application of the classifier in practice, i.e., on an unbiased test set, $B \perp\!\!\!\perp L$. In contrast, if taking images of dark-furred dogs would be bad luck, we might find an unwanted dependence between the signal S and the signal B in a training set, leading to $B \not\perp\!\!\!\perp L$, where we call B the bias variable. This dependence can be utilized by a classifier to reach a higher accuracy on the training set, resulting in a biased classifier and a lower accuracy on the test set.

To better understand the direction of the arrow from S to B, we want to emphasize, that data for a task is selected with a purpose. Images are included in the dataset because they show cats or dogs and one will, if necessary, deliberately accept imbalances in variables like fur-color. In contrast, if one find that our dataset misrepresents fur-color one would never accept a major misrepresentation of the ratio of cats and dogs to compensate for this problem. This demonstrates, that S influences B through the dataset creation while B does not influence S.

This bias model covers many, but not all, relevant situations in computer vision. To this end, we present one example for a situations where the bias model fits the data, and one example where it does not.

The first example is a driver assistance system that uses a camera to estimate aquaplaning risk [9]. To train such a system, example images of safe conditions and aquaplaning conditions are required. While the former can easily be collected in the wild, it is dangerous to drive a car under aquaplaning conditions. Thus, images for aquaplaning are collected in a specific facility. In this example, the signal S is the standing water, and the bias variable B is the location that

determines the background of the image. Because of the safety risk, they are dependent in the training set, but not at the time of application.

The second example is a system that predicts absenteeism in the workplace [2]. If an automated system predicts absenteeism, it should ignore pregnancy. Here, the bias variable B is the sex and the signal S is the chance of an employee to be absent from work. In this situation, similar to many others in algorithmic fairness, our bias model does not apply because the sex (B) does also influence the chance to be absent from work (L) during the application.

3.2 Conditional Independence for Debiasing

Deep neural networks unite a feature extractor and a predictor [16]. For adversarial debiasing, we separate the two at some intermediate layer. We denote the output of the feature extractor with R. Note that it is valid to use the whole network for feature extraction such that R contains the class predictions. Both networks are trained using a classification loss $\mathcal{L}_{cl}$, e.g., cross-entropy loss. Additionally, a debiasing loss $\mathcal{L}_{db}$ is used to prevent the extraction of the bias variable B. For a visualization, see Figure 1. Most approaches for adversarial debiasing [3,23,1,11] aim to find a representation R of I that is independent of the bias variable B while still being informative for the label L, i.e.,

$$R \perp\!\!\!\perp B \quad \wedge \quad R \not\perp\!\!\!\perp L. \tag{1}$$

In this work, we propose a novel strategy: Instead of independence, we aim for conditional independence of R and B, given the label L, i.e.,

$$R \perp\!\!\!\perp B \,|\, L \quad \wedge \quad R \not\perp\!\!\!\perp L. \tag{2}$$

First, we show that our strategy agrees with state-of-the-art results in explaining deep neural networks [17] and second, that an optimal classifier fulfills the conditional independence (2) but not the independence (1). We prove this statement for the case that all data generation processes are linear and L is scalar. Thus, loss functions that enforce the independence (1) will decrease the classifier's performance, while loss functions that ensure the conditional independence (2) will not.

The goal of debiasing is to prevent a deep neural network from using a biased feature. To reach this goal, we first need to determine whether a classifier uses a feature. So far, most approaches for adversarial debiasing use the dependence between a feature and the classifier's prediction to measure whether a classifier is using a feature. In contrast, we build on previous work for understanding deep neural networks [17]. While the independence criterion (1) obviously ensures that a bias variable B is not used for classification, the authors of [17] reveal that independence is too restrictive to determine whether a deep neural network uses a certain feature. They employ the framework of causal inference [14] to show that the ground-truth labels are a confounding variable for features of the input and the predictions of a deep neural network. In theoretical considerations and empirical experiments, they further demonstrate that the prediction of a neural

network and a feature of the input can be dependent even though the feature is not used by the network. The authors, therefore, suggest using the conditional independence (2) to determine whether a feature is used by a classifier.

Thus, using the independence criterion (1) is too strict. Even if the deep neural network ignores the bias, it might not satisfy (1) and, hence, not minimize a corresponding loss. Furthermore, minimizing such a loss based on the independence criterion will likely result in a less accurate classifier. Therefore, we use the conditional dependence criterion (2) for adversarial debiasing.

To corroborate this claim, we present a mathematical proof for the following statement. If the bias can be modeled as explained in Section 3.1, the optimal classifier, which recovers the signal and calculates the correct label for every input image, fulfills the conditional independence (2) but not the independence (1). In this work, we only include the proof for the linear, uni-variate case, i.e., all data generating processes (Φ, Ψ, Ξ) are linear and L is scalar. However, this proof can further be extended to the nonlinear case by using a kernel space in which the data generation processes are linear and replacing covariances with the inner product of that space.

Theorem 1. *If the bias can be modeled as described in Section 3.1, the optimal classifier fulfills the conditional independence (2) but not the independence (1).*

Proof. For this proof, we denote all variables with capital Latin letters. Capital Greek letters are used for processes, and lower-case Greek letters for their linear coefficients. The only exception is the optimal classifier denoted by F^*. First, we define all functions involved in the model. Afterward, since dependence results in correlation in the linear case, a simple calculation proves the claim.

Let S denote the signal according to the bias model, as explained in Section 3.1. In the linear case, the bias variable B can be split into a part that is fully determined by S and a part that is independent of S. Let B^* be the part of the bias variable that is independent of S, e.g., noise. According to the bias model, the bias variable B, the image I, and the label L are given by:

$$B = \alpha_1 S + \alpha_2 B^* =: \Phi(S, B^*), \tag{3}$$

$$I =: \Psi(S, B) = \Psi(S, \Phi(S, B^*)), \tag{4}$$

$$L = \zeta_1 S =: \Xi(S). \tag{5}$$

Thus, the label L can be calculated from the signal S only. The optimal solution F^* of the machine learning problem will recover the signal and calculate the label. By the assumptions of the bias model, the signal can be recovered from the input. Thus, there exists a function $\Psi^\dagger$ such that $\Psi^\dagger(\Psi(S, B)) = S$ holds. Therefore, F^* is given by

$$F^* := \Xi \Psi^\dagger. \tag{6}$$

Now, we have defined all functions appearing in the model. The rest of the proof are two straightforward calculations. In the linear case, the independence of

variables is equivalent to variables being uncorrelated. We denote the covariance of two variables A, B with $\langle A, B\rangle$. To prove that (1) does not hold, we calculate

$$\langle F^*(I), B\rangle = \left\langle \Xi\Psi^\dagger\Psi\left(S, \Phi\left(S, B^*\right)\right), \Phi\left(S, B^*\right)\right\rangle = \zeta_1\alpha_1\left\langle S, S\right\rangle. \tag{7}$$

This is equal to zero if and only if either all inputs contain an identical signal ($\langle S, S\rangle = 0$), the dataset is unbiased ($\alpha_1 = 0$), or the label does not depend on the signal ($\zeta_1 = 0$). For conditional independence, we can use partial correlation. Using its definition we obtain

$$\langle F^*(I), B\rangle | L = \left\langle F^*(I) - \frac{\langle F^*(I), L\rangle}{\langle L, L\rangle}L, B - \frac{\langle B, L\rangle}{\langle L, L\rangle}L\right\rangle. \tag{8}$$

We substitute L by (5) and use the properties of the inner product to arrive at

$$\langle F^*(I), B\rangle - \frac{\langle\zeta_1 S, \zeta_1 S\rangle\langle\zeta_1 S, B\rangle}{\langle\zeta_1 S, \zeta_1 S\rangle} = \frac{\zeta_1\alpha_1\langle S, S\rangle - \alpha_1\zeta_1^3\langle S, S\rangle^2}{\zeta_1^2\langle S, S\rangle} = 0. \tag{9}$$

This completes the proof for the linear case. For more detailed calculations we refer to the supplementary material.

The optimal classifier does not minimize loss criteria based on the independence (1). Further, from (7), we see that the dependence contains ζ_1, which is the correlation between the signal S and the neural network's prediction. Loss functions based on that criterion aim to reduce this parameter and, hence, will negatively affect the classifier's performance. We demonstrate this effect using a synthetic dataset in Section 4. In contrast, loss terms based on our new criterion (2) are minimized by the optimal classifier. Thus, corresponding loss functions do not reduce the accuracy to minimize bias.

3.3 Implementation Details

In practice, we are faced with the problem of integrating our criterion into the end-to-end learning framework of deep neural networks. Hence, we provide three possibilities to realize (2) as a loss function. Turning an independence criterion into a loss function is not straightforward. First, the result of an independence test is binary and, hence, non-differentiable. Second, we need to consider distributions of variables to perform an independence test. However, we only see one mini-batch at a time during the training of a deep neural network. Nevertheless, multiple solutions exist for the unconditional case.

In this section, we describe three possible solutions, namely: mutual information (MI), the Hilbert-Schmidt independence criterion (HSIC) and the maximum correlation criterion (MCC). We adapt the corresponding solutions from the unconditional case and extend them to conditional independence criteria.

The first solution makes use of the MI of R and B as suggested in [11]. Here, the criterion for independence is $\mathrm{MI}(R; B) = 0$, and the MI is the differentiable

loss function. In contrast, we use conditional independence. Our criterion is $\mathrm{MI}(R; B|L) = 0$, and the loss is given by the conditional MI:

$$\mathrm{MI}(R; B|L) = \sum_{l \in \mathcal{L}, b \in \mathcal{B} r \in \mathcal{R}} p_{R,B,L}\,(r, b, l) \log \frac{p_L(l) p_{R,B,L}\,(r, b, l)}{p_{R,L}(r, l) p_{B,L}(b, l)}. \tag{10}$$

To incorporate this loss, we need to estimate the densities $p_{R,B}, p_R$ and p_B in every step. We use kernel density estimation on the mini-batches with a Gaussian kernel and a variance of one fourth of the mean pairwise distance within a batch. This setting proved best in preliminary experiments on reconstructing densities.

As a second solution, we extend the Hilbert Schmidt independence criterion [8]. The variables are independent if and only if $\mathrm{HSIC}(R, B) = 0$ holds for a sufficiently large kernel space. The HSIC was extended to a conditional independence criterion by [7]. Multiple numerical approximations exist, we use

$$\operatorname{tr} G_R S_L G_B S_L = 0. \tag{11}$$

Here, S_L is given by $(\mathbb{I} + 1/m G_L)^{-1}$, where $\mathbb{I}$ is the identity matrix and $G_X = H K_X H$ with K_X the kernel matrix for $X \in \{B, R, L\}$ and $H_{ij} = \delta_{ij} - m^{-2}$ for δ_{ij} the Kronecker-Delta and m the number of examples. For the relation to HSIC and further explanations, we refer to [7]. We use the same kernel as above and estimate the loss on every mini-batch independently.

The third idea we extend is the predictability criterion from [1]

$$\max_f \quad \mathrm{Corr}(f(R), B) = 0. \tag{12}$$

To use this criterion within a loss function, they parametrize f by a neural network. However, this is not an independence criterion as it can be equal to zero, even if R and B are dependent. Therefore, it is unclear how to incorporate the conditioning on L. As a consequence, we decided to extend the proposed criterion in two ways. First, we use the maximum correlation coefficient (MCC)

$$\mathrm{MCC}(R, B) = \max_{f,g} \quad \mathrm{Corr}(f(R), g(B)) = 0, \tag{13}$$

which is equal to zero if and only if the two variables are independent [18]. Second, we use the partial correlation conditioned on the label L, which leads to

$$\max_{f,g} \quad \mathrm{PC}(f(R), g(B)\,|\,L) = 0. \tag{14}$$

To parametrize both functions f and g, we use neural networks. The individual effects of the two extensions can be observed through our ablation study in Section 4.2. Note that all three implementations can be used for vector-valued variables and, therefore, also for multiple bias variables in parallel.

3.4 Limitations

Debiasing with our method is only possible if the bias is known and a numerical value can be assigned to each image. This is, however, true for all adversarial

debiasing methods, e.g. the methods described in Section 2. Further, we assume the bias model from Section 3.1. While, this not the only bias possible, it covers many relevant situations, e.g. the one in Section 3.1.

One drawback of our method is that testing for conditional dependence is more complicated than testing for dependence. This is less a problem of calculation time and more of stability. Since the time complexity of both tests scales with the batch size, it can be ignored compared to the time complexity of backpropagation. However, the final data effects are stronger in the conditional dependence tests compared to their unconditional counterparts.

4 Experiments and Results

This section contains empirical results that confirm our theoretical claims. We first present experiments on a synthetic dataset that is designed to maximize the difference between the independence criterion (1) and the conditional independence criterion (2). Afterward, we report the results of an ablation study demonstrating that the gain in performance can be credited to the change of the independence criterion. Finally, we show that our findings also apply to a real-world dataset. For this purpose, we present experiments on different biased subsets of the cats and dogs dataset [12]. To evaluate our experiments, we measure the accuracy on an unbiased test set. Our debiasing approach is designed for applications in which the training set is biased, but where the classifier is used in an unbiased, real-world situation. Hence, the accuracy on an unbiased test set is our goal and therefore the most precise measure in this case. Further evaluations are included in the supplementary material.

4.1 Synthetic Data

If a feature is independent of the label for a given classification task, the independence criterion (1) and the conditional independence criterion (2) agree. Since we aim to maximize the difference between the two criteria, we use a dataset with a strong dependence between the label L and the variable B. We create a dataset of eight-by-eight pixels images that combine two signals. The first signal S, determines the shape of high-intensity pixels in the image, either a cross or a square, both consisting of the same number of pixels. The second signal B is the color of the image. To maximize the dependence between the label L and the bias variable B, every training image of a cross is green and every training image of a square is violet. In the test set, these two signals are independent. Example images from the training and test set can be seen in Figure 3.

For our first experiment (Setup I), we use the shape as the signal S to determine the label L and the mean color of the image as the bias variable B. To avoid any influence of shape- or color-preference of neural networks, in a second experiment, we use the inverse setting (Setup II). For this second experiment, we use the color as the signal S to determine the label L. Here, the bias variable B is calculated as the difference between the values of pixels in the square

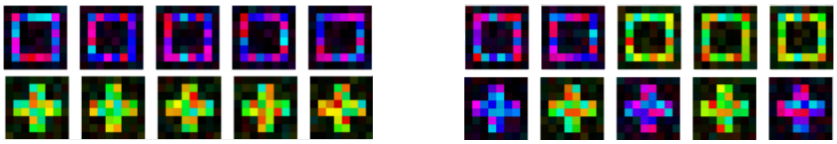

Fig. 3: Example images of the synthetic dataset, left: training set, right: test set.

Table 1: Results our method and all baselines on synthetic data. For both setups, we report mean accuracy ± standard error from 100 different random initializations on the same train/test-split. Best results in **bold**

	Baseline model	Adeli et al. [1]	Zhang et al. I [23]	Zhang et al. II [23]	Kim et al. [11]	Ours (MI)	Ours (HSIC)	Ours (MCC)
Setup I	0.819 ±0.016	0.747 ±0.015	0.736 ±0.018	0.747 ±0.016	0.771 ±0.012	0.840 ±0.014	0.846 ±0.021	**0.854** ±0.013
Setup II	0.791 ±0.016	0.776 ±0.014	0.837 ±0.017	0.750 ±0.013	0.767 ±0.016	**0.871** ±0.012	0.868 ±0.013	0.867 ±0.013

and those in the cross. We use a neural network with two convolutional layers (each with 16 filters of size 3×3) and two dense layers (128 neurons with ReLU activations and 2 neurons with softmax) as our backbone for classification. As a baseline, we use this network without any debiasing method. We have reimplemented four existing methods listed in Table 1. Two methods are proposed by Zhang et al. [23]. The first one called *Zhang et al. I* penalizes the predictability of B from R, the second one called *Zhang et al. II* penalizes predictability of B from R and L. Average accuracies of the competing approaches and from all three implementations of our proposed criterion are shown in Table 1.

Note that hyperparameter selection influences which feature (color or shape) the neural network uses for classifications, and we mitigate this by rigorous hyperparameter optimization. We use grid search for hyperparameters of our implementations as well as for general parameters like learning rates, and set hyperparameters for competing methods to values reported in the corresponding papers. We trained ten neural networks for each combination of different hyperparameters and evaluated them on an unbiased validation set. A list of the hyperparameters for every method is included in the supplementary material.

In Setup I, we see that differences between methods from the literature using the unconditional independence criterion (1) and the methods using our new conditional independence criterion (2) are much larger than the differences between methods within these groups. For the first experiment, the worst method using conditional independence performs 6.9 percentage points better than the best method using unconditional independence. The differences between the best and worst method within these groups is 1.4 percentage points and 3.5 percentage points, respectively. In the second setup, the results are similar. One competing

Table 2: The results of the ablation study. Every method is trained on a biased training set and evaluated on an unbiased test set. We report the accuracy averaged over 100 random initializations on the same train/test-split and the standard error. Best results are marked in **bold**

	Uncond. MI	Cond. MI	Uncond. HSIC	Cond. HSIC	Adeli et al. [1]	Uncond. MCC	Only PC	Ours (MCC)
Setup I	0.583 ±0.010	**0.840** ±0.014	0.744 ±0.011	**0.846** ±0.021	0.747 ±0.015	0.757 ±0.016	0.836 ±0.014	**0.854** ±0.013
Setup II	0.833 ±0.011	**0.871** ±0.012	0.590 ±0.011	**0.868** ±0.013	0.776 ±0.014	0.807 ±0.015	0.830 ±0.015	**0.867** ±0.013

method is surprisingly good in this experiment, but our models using conditional debiasing still perform much better with a gain of at least 3 percentage points.

We draw two conclusions from these experiments. First, none of the existing methods was able to improve the results of the baseline in Setup I, and only one approach did so for Setup II. This coincides with previous observations from the literature that adversarial debiasing methods are challenged in situations with strong bias, and reducing bias leads to decreased accuracy [3]. It also agrees with our findings discussed in Section 3.2. Second, we observe that all methods that use our new debiasing criterion with conditional independence reach higher accuracies than the baseline and, consequently, also a higher accuracies than existing methods. Note that we were able to reach the baseline performance for every method by allowing hyperparameters that deactivate the debiasing completely, e.g., by setting the weight of the debiasing loss to zero. To avoid this, we have limited the hyperparameter search to the range used in the respective publications.

4.2 Ablation Study

In the previous experiments, debiasing with our new criterion achieved higher accuracies than existing debiasing methods. We now conduct an ablation study for the three implementations from Section 3.3 to show that this increase can be attributed to our conditional independence criterion. More specifically, we report results for a method using unconditional mutual information and for a method using the unconditional HSIC as a loss. Furthermore, we present two methods that investigate the gap between the method of Adeli et al. [1] and our approach using the conditional maximal correlation coefficient. The first one uses the unconditional maximum correlation coefficient, and the second one incorporates the partial correlation (PC) instead of the correlation in (12). We use the settings and evaluation protocols of Setup I and Setup II from the previous section. The results are presented in Table 2.

The unconditional versions of MI and HSIC perform at least 3.8 percentage points worse than our conditional counterparts in both setups. For the third

method, we observe that the change from the predictability criterion to the maximum correlation coefficient ("Adeli et al. [1]" vs. "Uncond. MCC" and "Only PC" vs. "Ours(MCC)") increases the accuracy by at most 3.7 percentage points. In contrast, the change from correlation to partial correlation ("Adeli et al. [1]" vs. "Only PC" and "Uncond. MCC" vs. "Ours(MCC)") increases the accuracy by at least 5.4 percentage points. These observations indicate that the improvements found in Section 4.1 can be attributed to the difference between unconditional and conditional independence.

4.3 Real-World Data

Finally, we want to investigate whether increasing accuracies can also be observed for real-world image data. To evaluate the performance of our debiasing method, we require an unbiased test set.

Hence, we use a dataset with labels for multiple signals per image. This allows us to introduce a bias in the training set but not in the test set. We choose the cats and dogs dataset introduced by [12] for the same purpose. This dataset contains images of cats and dogs that are additionally labeled as dark-furred or light-furred. We first remove 20% of each class/fur combination as an unbiased test set. Then, we create eleven training sets with different levels of bias. We start with a training set that contains only light-furred dogs and only dark-furred cats. For each of the other sets, we increase the fraction of dark-furred dogs and light-furred cats by ten percent. Therefore, the last dataset contains only dark-furred dogs and light-furred cats. All training sets are created to have the same size for a fair evaluation. Hence, the number of training images is restricted by the rarest class/fur combination, leading to only less than 2500 training images.

We use a ResNet-18 [10] as a classifier. Details about this network and the selected hyperparameters for this experiment can be found in the supplementary material. To solely focus on the bias in our training sets, we refrain from pretraining on ImageNet [6] because ImageNet already contains thousands of dog images, and train from scratch instead. We only report accuracies for our approach with HSIC because it was most robust for different hyperparameters.

The results shown in Table 3 are averaged across three runs. Since the labels L and the bias variable B are binary, the two signals are indistinguishable for 0% and 100% dark-furred dogs, respectively. Furthermore, we obtain an unbiased training set for 50% dark-furred dogs. Our method reaches the highest accuracies in seven out of the remaining eight biased scenarios and the highest overall accuracy of 0.875 for 40% dark-furred dogs in the training set. For six out of these seven scenarios, the baseline was outside of our method's 95% confidence interval. We observe that competing methods only outperform the baseline in situations with little bias. This result supports our finding that existing methods are not suited for the bias model described in Section 3.1.

To further investigate the effectiveness of our approach, we compare the conditional and unconditional HSIC in Table 3 as well. We see that the conditional HSIC outperforms the unconditional HSIC in all biased scenarios. The stronger the bias, the bigger is the difference between the two methods. The correlation

Table 3: Experimental results on the cats and dogs dataset. All methods were trained on training sets in which p% of all dogs are dark-furred dogs and p% of all cats are light-furred. The first column indicates the fraction p, the others contain the accuracies on an unbiased test set. Best results in **bold**

Frac. p	Baseline model	Adeli et al.[1]	Zhang et al. I [23]	Zhang et al. II [23]	Uncond. HSIC	Ours (HSIC)
0%	**0.627** ±0.004	0.597 ±0.004	0.590 ±0.002	0.617 ±0.001	0.611 ±0.003	0.615 ±0.005
10%	0.800 ±0.001	0.774 ±0.002	0.779 ±0.005	0.785 ±0.007	0.759 ±0.012	**0.801** ±0.001
20%	0.845 ±0.003	0.829 ±0.000	0.812 ±0.002	0.809 ±0.005	0.816 ±0.002	**0.855** ±0.004
30%	0.852 ±0.007	0.842 ±0.003	0.837 ±0.004	0.834 ±0.003	0.834 ±0.002	**0.863** ±0.002
40%	0.859 ±0.007	0.855 ±0.004	0.870 ±0.002	0.850 ±0.001	0.861 ±0.003	**0.875** ±0.003
50%	0.859 ±0.006	**0.866** ±0.003	0.856 ±0.001	0.853 ±0.001	0.863 ±0.004	0.860 ±0.002
60%	**0.866** ±0.006	0.837 ±0.001	0.850 ±0.003	0.860 ±0.004	0.844 ±0.001	0.856 ±0.005
70%	0.844 ±0.003	0.854 ±0.003	0.835 ±0.005	0.841 ±0.005	0.835 ±0.003	**0.859** ±0.000
80%	0.829 ±0.002	0.822 ±0.005	0.820 ±0.005	0.826 ±0.003	0.820 ±0.007	**0.836** ±0.002
90%	0.773 ±0.010	0.743 ±0.001	0.758 ±0.001	0.731 ±0.002	0.757 ±0.003	**0.791** ±0.004
100%	0.612 ±0.001	0.612 ±0.004	0.604 ±0.001	0.609 ±0.001	0.606 ±0.002	**0.616** ±0.002

between the bias, measured as the absolute value between the difference of fractions of dark- and light-furred dogs, and the difference in accuracy between the conditional and unconditional HSIC method is 0.858.

5 Conclusion

In this work, we investigated a specific kind of dataset bias with a graphical model for data generation. Our exact model formulation allowed us to provide a mathematical proof to confirm our proposed conditional adversarial debiasing approach. Hence, our work differs from related work on adversarial debiasing, which solely relies on empirical evaluations. Our experimental results also support our theoretical claims. If a bias can be modeled with the investigated bias model, our conditional independence criterion is a better choice compared to an unconditional one. This is confirmed by our experiments. On synthetic data, the difference between conditional and unconditional debiasing criteria has been maximized. We further demonstrated in an ablation study that the conditional independence criterion is the reason for an increase in accuracy on unbiased test data, and improved accuracies have also been observed for real-world data.

In the future we aim to extend these empirical evaluations to get a better practical understanding of the method. This includes more detailed investigations in synthetic datasets, experiments on biased real-world datasets, e.g. HAM10000[21] or CAMELYON17[5], and experiments on unbiased datasets.

References

1. Adeli, E., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Fei-Fei, L., Niebles, J.C., Pohl, K.M.: Representation learning with statistical independence to mitigate bias. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2513–2523 (2021)
2. Ali Shah, S.A., Uddin, I., Aziz, F., Ahmad, S., Al-Khasawneh, M.A., Sharaf, M.: An enhanced deep neural network for predicting workplace absenteeism. Complexity **2020** (2020)
3. Alvi, M., Zisserman, A., Nellåker, C.: Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
4. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. propublica. See https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016)
5. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al.: From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE transactions on medical imaging **38**(2), 550–560 (2018)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: Kernel measures of conditional dependence. In: Advances in neural information processing systems. pp. 489–496 (2008)
8. Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Schölkopf, B., Smola, A.J.: A kernel statistical test of independence. In: Advances in neural information processing systems. pp. 585–592 (2008)
9. Hartmann, B., Raste, T., Kretschmann, M., Amthor, M., Schneider, F., Denzler, J.: Aquaplaning - a potential hazard also for automated driving. In: ITS automotive nord e.V. (Hrsg.), Braunschweig (2018)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning not to learn: Training deep neural networks with biased data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9012–9020 (2019)
12. Lakkaraju, H., Kamar, E., Caruana, R., Horvitz, E.: Discovering blind spots of predictive models: Representations and policies for guided exploration. arXiv preprint arXiv **1610** (2016)
13. Muckatira, S.: Properties of winning tickets on skin lesion classification. arXiv preprint arXiv:2008.12141 (2020)
14. Pearl, J.: Causality. Cambridge university press (2009)
15. Perez, F., Vasconcelos, C., Avila, S., Valle, E.: Data augmentation for skin lesion analysis. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, pp. 303–311. Springer (2018)
16. Reimers, C., Requena-Mesa, C.: Deep learning–an opportunity and a challenge for geo-and astrophysics. In: Knowledge Discovery in Big Data from Astronomy and Earth Observation, pp. 251–265. Elsevier (2020)

17. Reimers, C., Runge, J., Denzler, J.: Determining the relevance of features for deep neural networks. In: European Conference on Computer Vision (ECCV) (2020)
18. Sarmanov, O.V.: The maximum correlation coefficient (symmetrical case). In: Doklady Akademii Nauk. vol. 120, pp. 715–718. Russian Academy of Sciences (1958)
19. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528. IEEE (2011)
20. Tschandl, P., Codella, N., Akay, B.N., Argenziano, G., Braun, R.P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., et al.: Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. The lancet oncology **20**(7), 938–947 (2019)
21. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**, 180161 (2018)
22. Wang, A., Narayanan, A., Russakovsky, O.: Revise: A tool for measuring and mitigating bias in visual datasets. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
23. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)