Int. J. Appl. Math. Comput. Sci., 2025, Vol. 35, No. 1, 143–155
DOI: 10.61822/amcs-2025-0011

TARGETED DATA AUGMENTATION FOR IMPROVING MODEL ROBUSTNESS

AGNIESZKA MIKOŁAJCZYK-BAREŁA [a], MARIA FERLIN [a], MICHAŁ GROCHOWSKI [a,*]

[a] Department of Intelligent Control Systems and Decision Support
Gdańsk University of Technology
Gabriela Narutowicza 11/12, 80-233 Gdańsk, Poland
e-mail: michal.grochowski@pg.edu.pl

This paper proposes a new and effective bias mitigation method called targeted data augmentation (TDA). Since removing biases is often tedious and challenging and may not always lead to effective bias mitigation, we propose an alternative approach: skillfully inserting biases during the training to improve model robustness. To validate the proposed method, we applied TDA to two representative and diverse datasets: a clinical skin lesion dataset and a dataset of male and female faces. We identified and manually annotated existing instrument and sampling biases in these datasets, explicitly focusing on black frames and ruler marks in the skin lesion dataset and glasses in the face dataset. Using the counterfactual bias insertion (CBI) method, we confirmed that these biases strongly affect the model performance. By randomly inserting identified biases into training samples, we demonstrated that TDA significantly reduced bias measures by two times to more than 50 times, with only a negligible increase in the error rate. We performed our research on three model families: EfficientNet, DenseNet and Vision Transformer.

Keywords: bias mitigation, classification, data augmentation, deep neural networks.

1. Introduction

It is well known that when designing a model, an appropriate selection of data used to identify their parameters is crucial, especially for data-driven models. The most widely used data-driven models are neural network models. Their parameters are identified in the training process using the available data. One of the issues we have to deal with during training is the presence of bias in the gathered data. Its presence always results in errors during the development and later model use. Detecting and mitigating biases involves a variety of approaches, including collecting diverse datasets, carefully selecting and designing algorithms, using models that are as transparent as possible, employing explainable AI, and continuously monitoring decision-making processes. This article focuses on reducing the impact of data bias on model performance and generalization ability by applying a unique approach to utilize the data augmentation approach.

Data augmentation is widely used in deep learning-based systems in cases where we have too little data to identify model parameters optimally. In the case of computer vision applications, images are usually augmented by simple linear transformations, color augmentations, and sometimes with more advanced methods such as cutout or neural transformations (Shorten and Khoshgoftaar, 2019). Data augmentation increases model efficiency when analyzed using standard evaluation metrics such as accuracy, precision, or recall. Moreover, it contributes to overcoming the impact of biases.

In this paper, we propose the targeted data augmentation (TDA) method to mitigate selected biases in data, resulting in more robust classification. Bias refers to systematic deviations in data that affect model performance, which can include repetitive artifacts. Artifacts are physical elements in images that might introduce bias by creating spurious correlations. The term “bias in data” mainly refers to four of the most common data biases in machine learning: *observer bias* (Mahtani *et al.*, 2018), which may appear when annotators are guided by their own opinions to label data; *sampling bias*, when data are acquired in such a way that not all samples have the same sampling probability (Mehrabi *et al.*, 2021); *data handling bias*, when the way of handling data distorts

[*] Corresponding author

the classifier's output; and *instrument bias*, which refers to imperfections in the instrument and/or method used to collect the data (He and van de Vijver, 2012).

The occurrence of bias in data is an ordinary, often unnoticed, and underestimated problem that degrades or distorts results (Gao *et al.*, 2020; Luengo-Oroz *et al.*, 2021; Surówka and Ogorzałek, 2022). In general, identifying and removing biases is a tedious and challenging task. In addition, the process of eliminating biases, such as removing hair from skin lesion images or camera reflections from an image, results in some remaining biases or the appearance of new artifacts, which in turn are challenging to retouch even with advanced methods such as image inpainting (Bissoto *et al.*, 2020; Bardou *et al.*, 2022). Therefore, an opposite approach seems attractive and reasonable. This paper proposes a concept that focuses on enriching the training sets with selected biases, forcing the model to ignore them.

This approach, called TDA, breaks the cycle of mistaking correlation with causation by disrupting spurious correlations. If one randomly adds biases to the input during training, the model will treat the bias-connected features as irrelevant. The methodology behind TDA consists of four steps: bias identification (Step 1), augmentation policy design (Step 2), training with data augmentation (Step 3), and model evaluation (Step 4).

Identifying bias in data includes a preliminary, supervised step that aims to detect possible unwanted biases in data. To achieve this, we used manual data exploration. Regarding the skin lesion dataset, we manually labeled 2,000 skin lesion images, while the face dataset was labeled automatically using a trained glasses detection model.

Then, according to the detected biases, we applied an augmentation policy to mimic them and inject them into the training data. In short, we propose to insert biases into the training data instead of removing them. To evaluate quantitatively the effect, after the training, we measured the bias with the counterfactual bias insertion (CBI) method introduced by Mikołajczyk *et al.* (2021). While previous studies (e.g., Mikołajczyk *et al.*, 2021) focused on detecting biases and assessing their impact on model performance, they did not explore methods for mitigating them through data augmentation. By contrast, our proposed TDA method actively addresses bias mitigation by skillfully inserting biases during training to enhance model robustness.

Our method has shown a significant drop in bias measures. In classification using a model trained with TDA, two to over fifty times fewer images switched classes compared with classification using a model trained classically. Moreover, training with TDA resulted in only a slight increase in the error rate.

The contributions of this paper can be defined as follows:

- proposing a bias mitigation method that can easily complement the machine learning pipeline;
- introducing a bias mitigation benchmark that includes two datasets, the publicly available code for TDA and CBI, detailed results, and prepared collections of masks and images to serve as a benchmark for bias testing;
- identifying and confirming the existence of bias in the gender classification dataset;
- demonstrating that some models are more prone to capturing biases in the data, and this tendency is not always well reflected in standard evaluation metrics;
- mitigating biases related to black frames and ruler marks in the skin lesion dataset and glasses in the face dataset.

The paper is organized as follows. We describe the related works referenced in Section 2, while in Section 3, we introduce the reader to the proposed TDA method and the analyzed datasets. The details of the conducted experiments are presented in Section 4, while the results are presented in Section 5; finally, we formulate the conclusions in Section 6.

2. Related works

2.1. Bias mitigation. It is well documented that models, in most cases, reflect the biases in data and even amplify them (Zhao *et al.*, 2017). Commonly, biases are often subtle, and their potential impact on the performance of the models is complex and not fully understood, especially in cases when the model demonstrates high-accuracy results. Bias mitigation methods from classical literature usually operate on simple, linear models (Wang *et al.*, 2020) that are not applicable to deep learning models.

Data pre-processing approaches. The most obvious solution at first glance is to identify and remove all the biases (e.g., features or artifacts) at the pre-processing stage before training the models. This approach, fairness through blindness, was proposed by Wang *et al.* (2020). It is based on the idea that we can remove potentially biasing variables from the input data. For instance, we could remove information about a candidate's gender when evaluating a potential job candidate's résumé. However, in practice, this is not sufficient because some information about gender might be encoded in the résumé, e.g., feminine hobbies connected to gender or gender-specific adjectives. Moreover, removing all potential biases is often difficult, if not impossible, especially in computer vision. One of the applications in which the problem

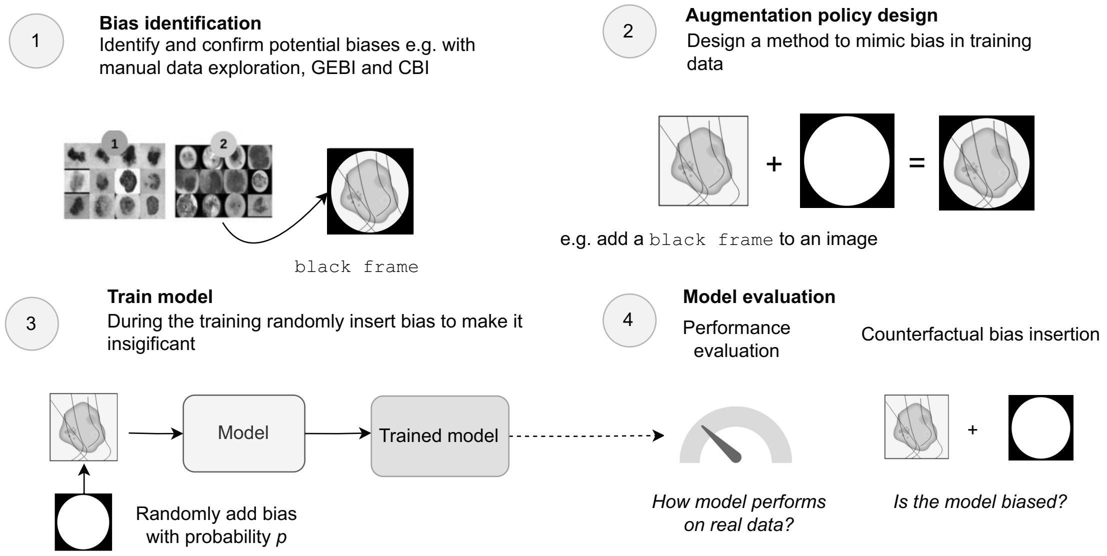
1
Bias identification
Identify and confirm potential biases e.g. with manual data exploration, GEBI and CBI
1
2
black frame
2
Augmentation policy design
Design a method to mimic bias in training data
+
=
e.g. add a black frame to an image
3
Train model
During the training randomly insert bias to make it insigificant
Model
Trained model
Randomly add bias with probability p
4
Model evaluation
Performance evaluation
Counterfactual bias insertion
+
How model performs on real data?
Is the model biased?

Fig. 1. Pipeline of the proposed targeted data augmentation (TDA) method for bias mitigation. The process involves four main steps: bias identification, where potential biases in the dataset are detected and measured using methods like counterfactual bias insertion for validation (1); augmentation policy design, where strategies are developed to introduce identified biases into the training data in a controlled manner (2); training with TDA, where the model is trained on data augmented with the biases according to the designed policy (3); and model evaluation, where the trained model is evaluated using standard metrics and CBI is employed again to assess the effectiveness of TDA in mitigating bias influence (4). This pipeline illustrates how TDA integrates bias identification and mitigation into the machine learning workflow, with CBI serving as a tool for both initial bias measurement and post-training evaluation. GEBI stands for global explanations for bias identification.

of artifact removal has been analyzed for many years is skin lesion analysis for possible cancer detection (Abbas *et al*., 2011; Oliveira *et al*., 2016).

Adversarial debiasing techniques. Another branch of approaches is adversarial bias mitigation, such as the supervised learning method proposed by Zhang *et al.* (2018). The task is to predict an output variable Y given an input variable X while remaining unbiased with respect to some variable Z. This approach uses the output layer of the predictor as an input to another model called the adversary network, which attempts to predict Z. This idea was further improved by Le Bras *et al.* (2020), who proposed the concept of adversarial debiasing filters. Their algorithm uses linear classifiers trained on different random data subsets at each filtering phase. Then, the linear classifier's predictions are collected, and a predictability score is calculated. High predictability scores are undesirable as their feature representation can be negatively exploited. Hence, Le Bras *et al.* (2020) proposed simply removing the top n instances with the highest scores. The above process can be repeated several times to reduce the bias influence.

Attention-based methods. Finally, there are a number of attempts to exploit the advantages of attention guidance. Early works on attention guidance in computer vision focused on improving segmentation tasks (Huang *et al*., 2019), enhancing classification using attention approaches from natural language processing (Barata *et al*., 2019), or even using attention maps to zoom closer to regions of interest (Li *et al*., 2019). The guidance provided by attention maps highlights relevant regions and suppresses unimportant ones, thus enabling better classification. A similar method is based on self-erasing networks that prohibit attention from spreading to unexpected background regions by erasing unwanted areas (Hou *et al*., 2018). Some researchers have proposed different ways to solve this problem, such as rule extraction, built-in knowledge, or built-in graphs (Chai and Li, 2020).

Other methods. Similarly, this problem is actively analyzed by many researchers in other fields and applications. Zhao *et al.* (2017) proposed an inference update scheme to match a target distribution to remove bias. Their method introduces corpus-level constraints so that selected features co-occur no more frequently than in the original training distribution. Dwork *et al.* (2018) proposed a scheme for decoupling classifiers that can be added to any black-box machine learning algorithm and then used to learn different classifiers for different groups.

Mikołajczyk *et al*. (2021) introduced the CBI method to detect and measure the influence of biases in datasets. However, their work did not propose any strategies for mitigating these biases. Our TDA method builds upon the understanding of biases revealed by CBI but innovates by introducing a data augmentation strategy specifically designed for bias mitigation.

Building on these approaches, we propose an alternative solution to increase the robustness of models by artificially introducing appropriately prepared, purpose-biased data into the analyzed dataset. We introduce our TDA method for bias mitigation (see Fig. 1). We have conducted experiments on two datasets to validate the approach: the International Skin Imaging Collaboration 2020 skin lesion benchmark (ISIC, 2020) and the gender classification dataset (Chauhan, 2019).

2.2. Comparison with existing bias mitigation methods. As described in the previous section, several methods have been proposed to mitigate biases in machine learning models. Adversarial debiasing methods, such as those introduced by Zhang *et al*. (2018), employ adversarial networks to remove sensitive information from learned representations. While effective in some cases, these methods often require complex training procedures and modifications to the model architecture, which can be computationally intensive and challenging to implement. Attention-based techniques, like the work by Barata *et al*. (2019), utilize attention mechanisms to focus the model on relevant, unbiased features. This can improve interpretability and reduce the impact of biases but may not fully eliminate the influence of spurious correlations, especially when biases are subtle or pervasive in the dataset.

Data pre-processing approaches, such as fairness through blindness (Wang *et al*., 2020), attempt to remove selected attributes or biases from the dataset before training. However, removing all potential biases is often nearly impossible, and pre-processing methods can either inadvertently discard important information or introduce new artifacts. Other methods (e.g., Zhao *et al*., 2017) that addressed bias amplification in language models by introducing corpus-level constraints might be effective for textual data, but applying such constraints to deep learning models handling high-dimensional image data is non-trivial. The method would require careful crafting of constraints and may not scale well to large datasets common in computer vision tasks. Similarly, the proposed decoupled classifiers to achieve fairness by training separate classifiers for different groups (Dwork *et al*., 2018), reduce bias by ensuring that each group is modeled independently. In that case the method relies on having explicit group labels and results in increased model complexity and maintenance efforts. It may also become less practical in cases where there is a huge disproportion of the number of samples per class, like in the benign/malignant skin lesion classification.

In contrast, our proposed TDA method offers a straightforward and practical approach to bias mitigation. Hence, by intentionally introducing biases into the training data in a controlled manner, TDA reduces the model's reliance on spurious correlations without altering the model architecture or requiring extensive modifications to the training procedure. This makes TDA easy to integrate into existing pipelines and applicable to a wide range of models and datasets. Moreover, TDA could even be used along with those methods to further improve the training.

2.3. Bias in datasets. There is a limited number of communications on defining biases in datasets. Torralba and Efros (2011) examined cross-dataset generalization on popular benchmarks by evaluating the performance of the 'car' and 'person' classes when training on one dataset and testing on another. Regarding the skin lesion classification problem, the existence and influence of bias in skin lesion datasets have been previously analyzed by some researchers, yet the problem has not been thoroughly investigated and explained yet. Bissoto *et al.* (2019; 2020) conducted research on biases in skin lesion benchmarks and their impact on the quality of model performance. They showed that existing dermoscopy artifacts, such as frames, gel bubbles, or ruler marks, distort the results and are a common source of bias in data.

Van Molle *et al*. (2018) proved that the model, in addition to medically relevant features, was driven by artifacts such as specular reflections, gel application and rulers. Mikołajczyk *et al*. (2022) showed that there is a strong correlation between artifacts such as black frames and ruler marks and the skin lesion type (benign/malignant). They showed that models trained on biased data learned spurious correlations, resulting in more errors in images with such artifacts. Barata *et al.* (2019) proposed a hierarchical classification model for the diagnosis of skin lesions with attention maps that helped interpret the results. They showed that the system is able to identify clinically relevant regions in the lesions and biasing features. Mikołajczyk *et al*. (2021) proposed the global explanations for bias identification (GEBI) method, which can be used for the detection of bias in data or in model behavior. Based on a skin lesion case study, they showed that the method could detect dataset artifacts. In that paper, the authors also proposed injecting artificial artifacts such as black frames, ruler marks, and red circles to measure the model's robustness against those biases. Finally, Bissoto *et al*. (2020) conducted a comprehensive analysis of seven visual artifacts and their influence on deep learning models and employed debiasing methods to decrease their impact on model performance. Unfortunately, they concluded that the existing state-of-the-art methods for bias removal are not capable of handling these biases effectively.

Data used in the paper. The *International Skin Imaging Collaboration 2020* benchmark (ISIC, 2020) is the largest skin lesion dataset divided into two classes: benign and malignant. It contains 33,126 dermoscopic images from over 2,000 patients. The diagnoses were confirmed either by histopathology, expert agreement, or longitudinal follow-up. ISIC gathered the dataset from several medical facilities. The dataset was used in the SIIM-ISIC Melanoma Classification Challenge (Zawacki *et al*., 2020). The lesion is usually in the center and clearly visible in the images. Examples of artefacts in this dataset that may introduce bias into the model include hair, frames, rulers, pen marks, and gel drops. Past research showed that frames are correlated with the malignant class and ruler marks with the benign (Mikołajczyk *et al*., 2022).

The *gender classification* dataset (Chauhan, 2019) consists of cropped images of male and female faces. The data were collected from various Internet sources. It contains 47 009 images in the training set (23 766 male, 23 243 female) and 11 649 images in the testing set (5 808 male, 5 841 female), with a similar distribution between female and male subsets. The images are frontal portraits of individuals. We have discovered that glasses are a possible bias source, as subjects wore them more often than actresses.

3. Targeted data augmentation

Within the framework of the proposed TDA method, we specify the following stages: bias identification, augmentation policy design, training with TDA, and finally, model evaluation. The TDA pipeline is presented in Fig. 1.

Bias identification is a preliminary, supervised step in which we aim to detect unwanted biases within the data. In our case, we manually explored the data to detect potential biases such as black frames and rulers in skin lesion images or glasses in face images. To confirm and measure the influence of these biases, we employed our CBI method introduced in Mikołajczyk *et al.* (2021). It is important to note that, while our previous work focused on bias detection and measuring bias influence using methods like CBI, they did not propose any bias mitigation strategies involving data augmentation. By contrast, our work introduces TDA as a novel bias mitigation method that utilizes data augmentation to improve model robustness.

In the next step, *augmentation policy design*, we develop a strategy for how to augment the data to mitigate the identified biases. This involves specifying which features will be modified and determining the method of modification. For example, we might decide to add a black frame to skin lesion images or glasses to face images to disrupt any spurious correlations between these artifacts and the target classes.

Then, we proceed to train with *targeted data augmentation*. This process involves randomly adding the specified biases to the training data according to the designed augmentation policy. By artificially introducing these biases during training, we aim to make them less correlated with a given class and increase randomness, thereby encouraging the model to focus on the relevant features. This step is a key contribution of our work, as previous studies did not incorporate data augmentation for bias mitigation.

Finally, in the *model evaluation* stage, we assess the performance of the model trained with TDA. We compare its performance with that of a model trained without TDA using standard evaluation metrics. To measure the remaining bias influence after applying TDA, we use the CBI method as an evaluation tool. Here, CBI allows us to quantify the effectiveness of TDA in mitigating bias, but it is not part of the TDA method itself.

3.1. Bias identification.

The key to successful bias-targeted data augmentation is to thoroughly identify potential biases and their sources. This can be done through manual inspection of data, with the aid of global explanation methods (e.g., GEBI (Mikołajczyk *et al*., 2021)).

In this paper, we selected potential biases through manual data analysis. To make it more objective, the manual data inspection process was based on three basic metrics: *artifact cardinality*, *artifact ratio*, and *class ratio.*

Cardinality of artifacts. The cardinality of artifacts within a class is the total number of elements (images) in which a certain artifact is present. The cardinality cannot exceed the number of images annotated per class.

Artifact ratio. The artifact ratio Q^{artifact} is the number of images with certain artifacts divided by the total number of images investigated. The artifact ratio shows how many samples have a certain artifact out of all. The artifact ratios are calculated separately for each class.

Class ratio. The class ratio Q^{class} is equal to the fraction of artifact ratios from two classes C_1 and C_2:

$$Q^{\text{class, artifact}} = \frac{Q^{\text{artifact},C_1}}{Q^{\text{artifact},C_2}}. \tag{1}$$

A class ratio close to one means that both the classes have the same incidence of artifacts. A significantly lower or higher class ratio means that the examined artifact is more common in one class than in the other.

3.2. Augmentation policy design.

An augmentation policy should describe *what* feature is being modified and *how* to modify it. This can include modifying features, adding elements, and swapping categorical feature values.

For instance, if we recognize a potential bias based on the feature called *country-of-origin*, we can randomly modify the *country-of-origin* value during the training (i.e., randomly switch *Poland* to *China*). If we suppose that the bias in a *bird song classification* comes from city sounds when classifying pigeon songs (since pigeons are often recorded in cities unlike other bird species), then we might randomly add city noises to samples. If the algorithm is biased toward a certain *age* (i.e., works well on people aged 16–18 but poorly on 19-year-olds), we can modify the *age* values within a certain range.

In the reported research, we focused on examining two image classification problems: skin lesion classification and gender classification. After discovering the biasing factors, we decided to augment the first dataset with ruler marks and black frames and the second one with glasses.

Frame augmentation. Frames, also known as dark corners, are black or white round markings around the skin lesions, black rectangular edges, and vignettes. We focused on the black round and rectangular markings of different sizes and shapes. We performed the frame augmentation by randomly inserting different types of black frames during training. Additionally, each frame was randomly scaled and rotated. We used six different types of frames during training and a separate set of five frames for the evaluation procedure.

Ruler augmentation. The ruler marks are partially or fully visible ruler markings of different shapes and colors that can be found throughout the dermoscopic skin lesion datasets. We used pairs of images and segmentation masks of rulers from a designated subset of data to copy rulers from the source image to the target image. This enabled us to achieve good augmentation quality without a significant increase in computing time. Similarly to the previous case, the segmentation masks were randomly scaled and rotated.

Glasses augmentation. Glasses are objects that may be visible in the face image but do not belong to the face itself. In this research, we randomly inserted masks of different types of glasses, including sunglasses, into the image at eye level. In accordance with Wesker *et al.* (2015), the mask was placed at one-third of a human face from the top (i.e., at eye level). It was not randomly rotated nor scaled, as this could result in a strange position of the glasses relative to the face. We provided thirty different masks for training and eight other masks for evaluation.

Examples of black frames, rulers, and glasses augmentation are shown in Fig. 2.

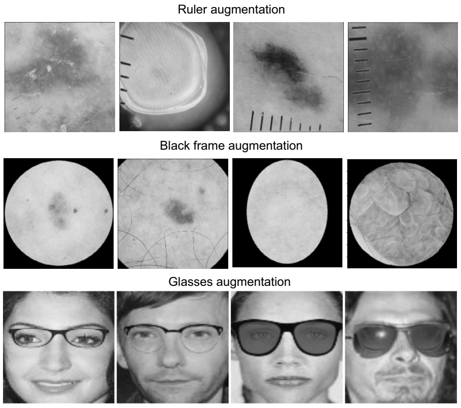

Fig. 2. Example augmentations.

3.3. Training with targeted data augmentation.

Training using TDA is almost identical to any other classical training. It requires a designed augmentation method that specifically targets bias. It can be used alongside other data augmentation methods and bias mitigation techniques. As the purpose of TDA is to mitigate spurious correlations between chosen features and outputs, it should be applied randomly, with a selected probability p.

3.4. Bias evaluation.

Bias evaluation is an important step in bias mitigation pipelines. Here, we measured the bias influence with CBI (Mikołajczyk *et al.*, 2021).

The model's prediction for the original input is compared with its prediction on the distorted input with the inserted bias. The steps of CBI are as follows. First, compute the predictions p for all samples from the dataset and store them. Next, insert the examined bias into every sample, and compute the predictions p^{biased} for all biased samples. Finally, compare the original predictions with the biased predictions.

Ideally, the model's predictions should remain the same after inserting minor artifacts or small data shifts. An example of inserting black frames into a skin lesion image is presented in Fig. 3.

The most basic measure proposed is the difference between the F_1 score values of the original samples and the F_1^{aug} values calculated for samples with the inserted bias.

Ideally, F_1^{aug} should be the same as F_1, which means that adding the bias to the data does not change the model's performance. Significant differences between F_1^{aug} and F_1 scores indicate greater susceptibility to bias. The F_1^{mean} index, which is the average of F_1 and F_1^{aug}, shows how well the model performs on both the original and modified data.

An additional measure proposed in the paper is the

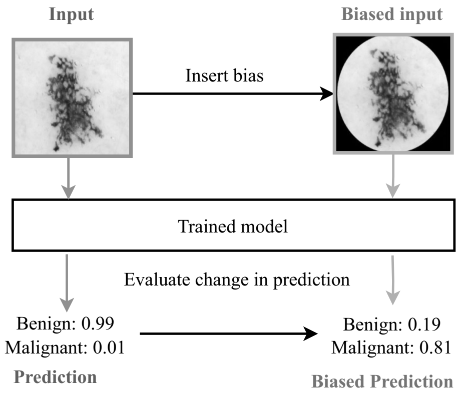

Fig. 3. Counterfactual bias insertion example.

number of switched classes. A prediction is considered *switched* when the predicted class has changed after inserting the bias. If the dataset is not biased with respect to the artifact under study, the number of switched classes should remain zero. The number of switched classes does not test the accuracy or correctness of the predicted category but only assesses the effect of bias on the prediction.

A switched class is defined as follows:

$$\text{switched}(k) = \begin{cases} 1 & \text{if } c^{\text{out}}_{p_k} \neq c^{\text{biased}}_{p_k}, \\ 0 & \text{otherwise,} \end{cases} \qquad (2)$$

where $c^{\text{out}}p_k$ is the predicted class (based on prediction p_k) for input k, and $c^{\text{biased}}p_k$ is the predicted class for the biased input k.

4. Experiments

We designed experiments to measure to what extent the TDA method helps mitigate biases, how the augmentation probability influences the results, and whether the obtained results differ between the considered models. Following the TDA pipeline, we first carefully identified the potential biases, i.e., frames and rulers in the skin lesion dataset and glasses in the gender dataset.

After bias identification, we measured the influence of the detected biases with the CBI. We added the artifacts to all images and measured the resulting change in the predictions. In the case of a perfect classifier, adding artifacts such as a black frame or a ruler mark to the image of a skin lesion, or eyeglasses to the image of a person, should not change the prediction.

After completing this step, we proceeded to the augmentation policy design. For this purpose, we used the augmentation methods described in the augmentation policy design (Section 3.2). We carefully prepared a separate subset of biases for training and testing for each dataset to avoid data leakage. For training, we applied rotations, zoom-in/out, and other modifications of the masks (in the case of skin lesions). We inserted each bias with a certain probability p, depending on the training. When testing two different biases on the skin lesion dataset, we augmented each bias separately.

We selected three model families: EfficientNet (Tan and Le, 2019), DenseNet (Huang *et al*., 2016), and Vision Transformer (ViT) (Dosovitskiy *et al*., 2021), and trained on the same dataset, either using classical training or with bias augmentation, with identical hyperparameters. During the evaluation, we compared the results on clean images without the bias ($p = 0.0$) with those on images with inserted biases ($p = 1.0$), using unmodified masks to avoid unnecessary randomness. We averaged the obtained results for each bias. We used the *switched* metric introduced in Section 3.4 to measure how many predictions changed after adding the bias to the data.

4.1. Skin lesion classification. In the case of the skin lesion dataset, the relevant literature shows that black frames are correlated with malignant lesions and ruler marks with the benign class (Mikołajczyk *et al*., 2022).

This finding was further supported by our research, where we hand-labeled an additional 2,000 samples. Table 1 presents the results of these 2,000 images with annotations aggregated with public labels from (Mikołajczyk *et al*., 2022).

The metrics used for comparison were the *cardinality of artifacts*, *artifact ratio*, and *class ratio* introduced in Section 3.1. Statistics show that frames are not only a very common artifact but are also strongly correlated with the malignancy class. The total number of images with frames is about 5% for the benign class and 26% for malignant, for malignant, meaning that frames are five times more common in malignant cases ($Q^{\text{class}} \approx 5.01$). Ruler marks are even more common but less correlated with the skin lesion type than frames, with $Q^{\text{class}} \approx 1.39$. This confirms the presence of potentially biasing features that may affect the models.

Then, we conducted CBI experiments to measure how bias affects the models. For each test image, we inserted five different types of frames or ruler masks into the data and measured the effect of inserting the given bias. We present the averaged results (with augmentation probability $p = 0.0$) in Table 2.

The obtained results clearly show that all three architectures were strongly affected by inserting the frames. Each neural model had a significantly lower F_1 score after bias insertion: the smallest difference was $F_1^{\text{diff}} = 12.7\%$ for EfficientNet, then a larger difference of 36.3% for DenseNet, and finally 56.62% for the Vision Transformer (ViT). A higher difference means a greater influence of the inserted bias on the prediction. In all measured cases, predictions were strongly affected by the

Table 1. Aggregated annotations from the work of Mikołajczyk *et al.* (2022) and manually annotated artifacts in the skin lesion dataset *ISIC 2019* (Combalia *et al.*, 2019; Codella *et al.*, 2018; Tschandl *et al.*, 2018) and *ISIC 2020* (ISIC, 2020). *Ben* and *mal* stand for benign and malignant, respectively.

Type	$\|\text{ben}\|$	Q^{artifact}	$\|\text{mal}\|$	Q^{artifact}	Q^{class}
Frame	104	5.20%	521	26.05%	5.01
Hair	958	47.88%	868	43.40%	0.91
Ruler	422	21.09%	586	29.30%	1.39
Others	426	21.29%	818	40.90%	1.92
None	538	26.89%	268	13.40%	0.50
Total (artifacts)	2448		3061		
Total (images)	2001		2000		

biases. Almost all samples switched from the benign to malignant class after inserting the frame bias.

Similarly, when testing the ruler insertion, we observed a drop in accuracy, although not as large as in the case of the frames. Depending on the neural model's architecture, the mean difference between F_1 and F_1^{aug} was approximately 5% for DenseNet, around 13% for EfficientNet, and only 1.4% for ViT. Most of the cases with inserted ruler marks switched from malignant to benign. These observations point out the need for bias mitigation. For this purpose, we used segmentation masks with ruler marks to mitigate ruler bias and six masks of frames. The training details are described in Section 4.3 and the results of training are gathered in Table 2.

In all examined cases, training with TDA resulted in a significant decrease in switched predictions, as well as an increase in F_1^{aug}, and sometimes even an increase in F_1. The effect of frame bias, as measured by the *switched* metric, decreased by a factor of 38 (from almost 2,000 cases to only 50) in DenseNet, 2.7 times for EfficientNet, and an impressive 57 times for ViT. In the case of ruler marks, we observed a notable decrease in *switched* values: by a factor of 1.8 for DenseNet and EfficientNet, and by a factor of 5 for ViT.

4.2. Gender classification. To measure the bias towards glasses in the gender classification dataset, we additionally annotated a small subsample of this dataset for the presence of eyeglasses. We used this along with the *glasses or no glasses* dataset,[1] which was generated by a generative adversarial network (GAN), to train an EfficientNet-B2 model to detect images with glasses. The final glasses classifier achieved a high performance score of about $F_1 \approx 0.96$. We automatically annotated the gender dataset using this classifier and compared the metrics between samples in both gender categories. This confirmed our hypothesis about the bias in the dataset. The results are presented in Table 3.

Glasses are not very common in this dataset, the total number of images with glasses is less than 12% for men and a little over 1% for women. However, glasses are over seven times more common ($Q^{\text{class}} \approx 7.79$) in images of men than in images of women. This makes it the strongest single trait disparity between classes in the present comparison.

Since the total number of images with glasses in the dataset is quite small, we additionally conducted CBI experiments to confirm this hypothesis. For each test image, we inserted nine different types of glasses and measured the effect of bias. The averaged results (with augmentation probability $p = 0.0$) are gathered in Table 4.

We confirmed that, similarly to the skin lesion dataset, all models were affected by glasses insertion – a feature that theoretically should not affect the result. Each model had a lower F_1 score after the insertion of the bias: F_1^{diff} was equal to 4.93 for DenseNet, 4.6 for EfficientNet, and 1.21 for ViT. Consequently, almost all model predictions for selected samples, switched from the female to the male class, once again confirming the correlation between glasses and gender.

We used 30 masks with both corrective glasses and sunglasses to test the bias mitigation algorithm on this dataset. The training details are described in Section 4.3, while the results of the training are reported in Table 4.

In all the tested cases, training with TDA resulted in a significant decrease in switched predictions, as well as an increase in F_1^{aug}, and even sometimes in F_1. The glasses bias, represented by the *switched* metric, decreased 3.8 times for DenseNet, 3.3 times for EfficientNet, and 1.8 times for ViT. This shows that even in task where bias does not occur very often, TDA can have a positive effect.

4.3. Training details. We performed the experiments using the *ISIC 2020* and *gender classification* datasets described in Section 2.3. In the case of skin lesion classification, we applied 411 ruler segmentation masks from (Ramella, 2021) which were published in open repositories.

We tested and trained three different architectures:

[1] Glasses or no glasses dataset: `https://www.kaggle.com/datasets/jeffheaton/glasses-or-no-glasses` (from the course T81-855: Applications of Deep Learning at Washington University in St. Louis).

Table 2. Counterfactual bias insertion results on frame and ruler bias testing with and without targeted data augmentation: *'ben'* and *'mal'* stand for benign and malignant, respectively.

Model	Type	p	switched		mal to ben	ben to mal	F_1	F_1^{aug}	F_1^{diff}
DenseNet121	Frame	0	1928.8	25.61%	0.79%	99.21%	88.18%	51.88%	36.30%
		0.25	68.0	0.90%	66.76%	33.24%	87.79%	85.50%	2.29%
		0.5	**49.8**	**0.66%**	59.84%	40.16%	88.22%	87.11%	1.11%
		0.75	82.0	1.09%	11.95%	88.05%	**88.54%**	**88.29%**	**0.24%**
		1	73.6	0.98%	70.65%	29.35%	87.51%	86.66%	0.85%
	Ruler	0	142.4	1.89%	78.51%	21.49%	89.62%	84.82%	4.80%
		0.25	**77.2**	**1.03%**	86.01%	13.99%	89.84%	88.31%	1.52%
		0.5	92	1.22%	90.87%	9.13%	88.92%	88.58%	0.34%
		0.75	77.8	1.03%	83.03%	16.97%	**90.40%**	**89.01%**	1.39%
		1	86.6	1.15%	90.30%	9.70%	88.60%	88.60%	**-0.01%**
EfficientNet-B2	Frame	0	504.2	6.70%	4.72%	95.28%	**86.56%**	73.81%	12.76%
		0.25	**187.0**	**2.48%**	8.34%	91.66%	83.23%	**78.61%**	4.62%
		0.5	310.0	4.12%	12.97%	87.03%	79.06%	73.22%	5.84%
		0.75	411.8	5.47%	41.04%	58.96%	78.12%	78.00%	**0.11%**
		1	204.8	2.72%	9.57%	90.43%	73.02%	72.90%	0.12%
	Ruler	0	173.4	2.30%	97.00%	3.00%	87.89%	76.67%	11.22%
		0.25	167.2	2.22%	95.81%	4.19%	87.92%	79.39%	8.52%
		0.5	157.4	2.09%	92.25%	7.75%	88.97%	81.64%	7.33%
		0.75	171.2	2.27%	93.57%	6.43%	85.48%	75.67%	9.81%
		1	**93**	**1.24%**	86.67%	13.33%	88.56%	**86.55%**	**2.01%**
ViT	Frame	0	5014.2	66.59%	0.00%	100.00%	88.85%	32.22%	56.62%
		0.25	187.0	2.48%	0.08%	91.66%	83.23%	78.61%	4.62%
		0.5	161.2	2.14%	2.61%	97.39%	90.18%	86.88%	3.30%
		0.75	115.8	1.54%	63.90%	36.10%	88.33%	87.77%	**0.55%**
		1	**87.2**	**1.16%**	60.55%	39.45%	**90.30%**	**88.49%**	1.81%
	Ruler	0	189.4	2.52%	93.24%	6.76%	88.85%	**87.47%**	1.38%
		0.25	37.4	0.50%	63.64%	36.36%	79.87%	78.99%	0.88%
		0.5	**37**	**0.49%**	69.19%	30.81%	80.30%	80.22%	**0.08%**
		0.75	290	3.85%	96.97%	3.03%	78.94%	76.82%	2.12%
		1	33.6	0.45%	25.00%	75.00%	78.76%	78.54%	0.22%

EfficientNet-B2, DenseNet121, and Vision Transformer (base version, 16 patch, 224).

We trained all models for five epochs. Regarding the hyperparameters, the learning rate was set to $lr = 510^{-4}$ for EfficientNet-B2 and DenseNet121 and to $lr = 510^{-5}$ for ViT. The step scheduler reduced the learning rate by multiplying it by 0.9 every epoch. A batch size of 64 was used for skin lesion classification and 2 for gender classification.

Depending on the experiment scenario, we additionally randomly inserted the biases—a ruler, a frame, and glasses—with different probabilities $p \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$.

4.4. Code and data availability. The code for the targeted data augmentation and counterfactual bias insertion methods, along with the scripts used for training and evaluation, are publicly available in our GitHub repository[2].

Full datasets with all skin lesion artifacts annotations (4k samples), used segmentation masks and eyeglasses, and bias annotations are available to download in an open-access repository[3].

The datasets used in our experiments are publicly accessible:

- ISIC 2020 skin lesion dataset,[4]
- gender classification dataset,[5]

[2] Repository of target data augmentation code: https://github.com/AgaMiko/targeted-data-augmentations

[3] Bias mitigation data repository: https://mostwiedzy.pl/pl/open-research-data/bias-mitigation-benchmark-that-includes-two-datasets,227084836236401-0?_share=322e9564d0341d8a

[4] ISIC 2020 skin lesion dataset: https://www.kaggle.com/c/siim-isic-melanoma-classification/data.

[5] Gender classification dataset: https://www.kaggle.com/datasets/cashutosh/gender-classification-dataset.

Table 3. Artifact statistics for a semi-automatically annotated gender classification dataset.

Type	$\lvert$male$\rvert$	Q^{artifact}	$\lvert$female$\rvert$	Q^{artifact}	Q^{class}
Glasses	2659	11.19%	334	1.44%	7.79
None	21107	88.81%	22909	98.56%	0.90
Total	23766		23243		

Table 4. Counterfactual bias insertion results on glasses bias testing with and without targeted data augmentation.

Model	Type	p	switched		mal to fem	fem to mal	F_1	F_1^{aug}	F_1^{diff}
DenseNet121	Glasses	0	908.4	7.8%	9.98%	90.02%	96.90%	91.98%	4.93%
		0.25	**235.6**	**2.02%**	55.99%	44.01%	96.95%	**95.88%**	**1.07%**
		0.5	284.8	2.44%	44.56%	55.44%	**98.78%**	95.59%	1.18%
		0.75	282.9	2.43%	53.02%	46.98%	96.19%	95.27%	0.92%
		1	337.4	2.90%	26.24%	73.76%	93.42%	94.14%	0.71%
EfficientNet-B2	Glasses	0	800.9	6.88%	21.24%	78.76%	96.41%	91.81%	4.60%
		0.25	300.7	2.58%	54.36%	45.64%	**96.90%**	**95.76%**	1.15%
		0.5	273.6	2.35%	53.57%	46.43%	96.62%	95.49%	1.12%
		0.75	**237.4**	**2.04%**	39.07%	60.93%	96.66%	95.69%	0.97%
		1	333.7	2.86%	63.97%	36.03%	95.59%	95.45%	**0.13%**
ViT	Glasses	0	262.1	2.25%	11.45%	88.55%	**97.69%**	96.48%	1.21%
		0.25	187.1	1.61%	14.31%	85.69%	97.62%	**96.90%**	0.72%
		0.5	160.3	1.38%	27.10%	72.90%	97.31%	96.80%	0.51%
		0.75	178.9	1.54%	37.33%	62.67%	97.17%	96.66%	0.51%
		1	**144.6**	**1.24%**	24.52%	75.48%	97.34%	96.88%	**0.46%**

- segmentation masks of hair and rulers,[6]
- skin lesion artifacts annotations (2k samples) (Mikołajczyk *et al.*, 2022).[7]

Detailed instructions for reproducing our experiments are provided in the README file of the code repository.

5. Results

The results from the experiments suggest that various neural network architectures exhibit different degrees of susceptibility to biases, and understanding the architectural differences can provide insights into these variations. Specifically, the vision transformer (ViT) appears to be more affected by certain biases, such as black frames, compared to convolutional architectures like DenseNet121 and EfficientNet-B2. We suspect that this discrepancy might arise due to the following architectural distinctions:

- **Convolutional networks (DenseNet121, EfficientNet-B2).** They rely on convolutional layers that use local spatial hierarchies by applying convolutional filters to usually small regions of the input image. This localized filtering process allows CNNs to be more robust to small perturbations or artifacts in the image, as the convolutional filters are designed to focus on *local patterns* (e.g., textures, edges, like gel blobs), rather than global features.
- **Vision transformer (ViT).** Unlike convolutional networks, the ViT architecture processes an image by splitting it into a few larger patches and applying self-attention mechanisms across these patches. Since the ViT does not inherently encode spatial locality, it may be more prone to biases that are *globally distributed* across the image, such as frames or rulers. This lack of inductive bias towards local spatial patterns makes ViT more vulnerable to spurious correlations introduced by global artifacts like frames. The self-attention mechanism may focus disproportionately on these artifacts, interpreting them as relevant features during classification.

The susceptibility of ViT to frame biases, in particular, may be due to its global attention mechanism that treats every patch of the image with equal importance, potentially mistaking the repetitive nature of frames as significant features. This explains why ViT showed the largest drop in the F_1 score after the insertion of the frame bias, compared with the relatively smaller drops in convolutional architectures (DenseNet121 and EfficientNet-B2).

We observed an interesting trend that highlights the importance of selecting an optimal probability for applying the TDA. For most models, lower augmentation

[6] Segmentation masks of hair and rulers: `https://github.com/gramella/HR`.

[7] Manual annotations and public labels: `https://github.com/AgaMiko/debiasing-effect-of-gans`.

probability ($p = 0.25$ to 0.5) models were exposed to the bias often enough to learn to ignore it, but not so frequently that the model overfitted to the augmented data. This balance helped improve robustness while keeping the error rates low. For example, DenseNet121 and EfficientNet-B2 showed the most consistent reduction in the number of switched classes and minimal differences between F_1 and F_1^{aug} at $p = 0.5$. At higher probabilities ($p = 0.75$ to 1.0), where the model was exposed to biases in almost every training sample, the effectiveness of the augmentation diminished slightly. Although the models became robust to the biases they were trained on, their performance on clean, unmodified images sometimes decreased suggesting that overexposure to the biases can lead to model overfitting to the augmented data, limiting its ability to generalize to clean data. This shows that a probability in the mid-range ($p = 0.5$) generally offered the best trade-off between robustness and accuracy across the tested models.

6. Conclusions

The robustness of a neural network model refers to its ability to remain effective even when it is subjected to data with distributions different from those on which it was trained, containing artefacts (visible or hidden), noise or clear imbalances. In general, biases present in the training data weaken the robustness, particularly hampering the model's ability to generalise since it learns patterns specific to the biases, which can lead to overfitting. In our study, we focused on showing the impact of instrument and sampling bias on model robustness in a classification task. We confirmed our hypotheses regarding bias influence by training the models on datasets and testing them using the CBI method. The results showed that the models are strongly affected by the biases selected for the experiments, with the largest *switched* metrics observed for the frame bias—a commonly observed artifact that is strongly correlated with the malignant class. To mitigate the discovered biases, we trained the models by randomly inserting biases during the training.

Our experiments demonstrate that TDA effectively reduces the influence of biases on model predictions, as evidenced by significant decreases in the number of switched classes and improved robustness measures. Unlike adversarial debiasing, TDA does not require additional adversarial networks or complex loss functions, simplifying implementation and reducing computational overhead. The TDA method is model-agnostic, allowing it to be used with any neural network without modifying the architecture. This contrasts with attention-based methods, which may depend on specific model designs or require additional components like attention layers. Moreover, TDA does not rely on removing biases, which can be impractical or introduce new issues. Instead, it embraces the presence of biases by incorporating them into the training process, teaching the model to ignore these features during prediction. However, the method also shares the limitation of any bias mitigation method, that is, it requires manual effort for bias identification.

We examined three different deep neural network architectures: DenseNet121, EfficientNet-B2, and Vision Transformer (ViT). All of them exhibited similar behaviour when subjected to TDA. All models showed significant improvement in terms of robustness to bias after training with TDA. Notably, the vision transformer was strongly influenced by frame and ruler biases, yet was the least affected by the glasses bias. This was most likely due to architectural differences between these networks, as ViT does not include convolution layers in its design. The results have shown that the F_1 score is not always the best indicator of robustness. For instance, the ViT model with the highest F_1 values was the most affected model by the frame bias. This indicates the need for careful and comprehensive model evaluation, beyond standard performance metrics. Future work could include automating the bias identification and annotation process, testing how TDA handle multiple biases simultaneously, and exploring the applicability of TDA in other domains or with other types of data (e.g., textual data).

An interesting point relates to the augmentation probability coefficient p. In each case, experiments with $p > 0$ resulted in a reduction of the *switched* metric, and in most cases, p within the range 0.25 to 0.75 gave the best results. Similarly, using TDA reduced the difference between F_1 and F_1^{aug}, which is a clear evidence supporting the postulated immunization of the model to the bias in the analyzed data.

In this paper, we propose a new and effective method for mitigating biases called TDA. Although this method requires partial manual user involvement, it can successfully debias models without removing biases (e.g., artifacts) from inputs. Furthermore, we propose confirming the manual bias identification using CBI to avoid improperly chosen bias factors.

Acknowledgment

The research on bias reported in this publication was supported by the Polish National Science Centre (Preludium project no. UMO-2019/35/N/ST6/04052). The authors thank the Polish National Science Centre for its support.

References

Abbas, Q., Celebi, M.E. and García, I.F. (2011). Hair removal methods: A comparative study for dermoscopy images, *Biomedical Signal Processing and Control* **6**(4): 395–404.

Barata, C., Marques, J.S. and Celebi, M.E. (2019). Deep attention model for the hierarchical diagnosis of skin lesions, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, USA*, pp. 2757–2765.

Bardou, D., Bouaziz, H., Lv, L. and Zhang, T. (2022). Hair removal in dermoscopy images using variational autoencoders, *Skin Research and Technology* **28**(3): 445–454.

Bissoto, A., Fornaciali, M., Valle, E. and Avila, S. (2019). (DE)Constructing bias on skin lesion datasets, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, USA*, pp. 1–9.

Bissoto, A., Valle, E. and Avila, S. (2020). Debiasing skin lesion datasets and models? Not so fast, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, USA*, pp. 3192–3201.

Chai, C. and Li, G. (2020). Human-in-the-loop techniques in machine learning, *IEEE Data Engineering Bulletin* **43**(3): 37–52.

Chauhan, A. (2019). Gender classification dataset, `https://www.kaggle.com/datasets/cashutosh/gender-classification-dataset`.

Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H. and Halpern, A. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC), *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, USA*, pp. 168–172.

Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C. Puig, S. and Malvehy, J. (2019). BCN20000: Dermoscopic lesions in the wild, *arXiv*: 1908.02288.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale, *International Conference on Learning Representations, Vienna, Austria*.

Dwork, C., Immorlica, N., Kalai, A.T. and Leiserson, M. (2018). Decoupled classifiers for group-fair and efficient machine learning, *1st Conference on Fairness, Accountability and Transparency, New York, NY*, pp. 119–133.

Gao, D., Wu, R., Liu, J., Fan, X. and Tang, X. (2020). Finding robust transfer features for unsupervised domain adaptation, *International Journal of Applied Mathematics and Computer Science* **30**(1): 99–112, DOI: 10.34768/amcs-2020-0008.

He, J. and van de Vijver, F. (2012). Bias and equivalence in cross-cultural research, *Online Readings in Psychology and Culture* **2**(2): 2307–0919.

Hou, Q., Jiang, P., Wei, Y. and Cheng, M.-M. (2018). Self-erasing network for integral object attention, *32nd Conference on Advances in Neural Information Processing Systems, NeurIPS 2018*.

Huang, G., Liu, Z. and Weinberger, K.Q. (2016). Densely connected convolutional networks, *CoRR:* abs/1608.06993.

Huang, Q., Chen, X., Metaxas, D. and Nadar, M.S. (2019). Brain segmentation from k-space with end-to-end recurrent attention network, *in* D. Shen *et al.* (Eds), *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2019*, Springer, Cham, pp. 275–283.

ISIC (2020). SIIM-ISIC 2020 challenge dataset, International Skin Imaging Collaboration, `https://challenge2020.isic-archive.com/`.

Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M., Sabharwal, A. and Choi, Y. (2020). Adversarial filters of dataset biases, *Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria*, pp. 1078–1088.

Li, H., Liu, Y., Ouyang, W. and Wang, X. (2019). Zoom out-and-in network with map attention decision for region proposal and object detection, *International Journal of Computer Vision* **127**(3): 225–238.

Luengo-Oroz, M., Bullock, J., Pham, K.H., Lam, C.S.N. and Luccioni, A. (2021). From artificial intelligence bias to inequality in the time of COVID-19, *IEEE Technology and Society Magazine* **40**(1): 71–79.

Mahtani, K., Spencer, E.A., Brassey, J. and Heneghan, C. (2018). Catalogue of bias: Observer bias, *BMJ Evidence-Based Medicine* **23**(1): 23–24.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021). A survey on bias and fairness in machine learning, *ACM Computing Surveys* **54**(6): 1–35, DOI: 10.1145/3457607.

Mikołajczyk, A., Grochowski, M. and Kwasigroch, A. (2021). Towards explainable classifiers using the counterfactual approach—Global explanations for discovering bias in data, *Journal of Artificial Intelligence and Soft Computing Research* **11**(1): 51–67.

Mikołajczyk, A., Majchrowska, S. and Limeros, S.C. (2022). The (de)biasing effect of GAN-based augmentation methods on skin lesion images, *arXiv:* 2206.15182.

Oliveira, R.B., Mercedes Filho, E., Ma, Z., Papa, J.P., Pereira, A.S. and Tavares, J.M.R. (2016). Computational methods for the image segmentation of pigmented skin lesions: A review, *Computer Methods and Programs in Biomedicine* **131**: 127–141.

Ramella, G. (2021). Hair removal combining saliency, shape and color, *Applied Sciences* **11**(1): 447.

Shorten, C. and Khoshgoftaar, T.M. (2019). A survey on image data augmentation for deep learning, *Journal of Big Data* **6**(1): 1–48.

Surówka, G. and Ogorzałek, M. (2022). Segmentation of the melanoma lesion and its border, *International Journal of Applied Mathematics and Computer Science* **32**(4): 683–699, DOI: 10.34768/amcs-2022-0047.

Tan, M. and Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks, *Proceedings of the 36th International Conference on Machine Learning, Long Beach, USA*, pp. 6105–6114.

Torralba, A. and Efros, A.A. (2011). Unbiased look at dataset bias, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, USA*, pp. 1521–1528.

Tschandl, P., Rosendahl, C. and Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, *Scientific Data* **5**: 180161.

Van Molle, P., De Strooper, M., Verbelen, T., Vankeirsbilck, B., Simoens, P. and Dhoedt, B. (2018). Visualizing convolutional neural networks to improve decision support for skin lesion classification, *in* D. Stoyanov *et al.* (Eds), *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, Springer, Cham, pp. 115–123.

Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K. and Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA*, pp. 8919–8928.

Wesker, K.H., Radlanski, R.J. and Kaczmarzyk, T. (2015). *Face: Atlas of Clinical Anatomy*, Kwintesencja, Warsaw, (in Polish).

Zawacki, A., Helba, B., Shih, G., Weber, J., Elliott, J., Combalia, M., Kurtansky, N., Codella, N., Culliton, P. and Rotemberg, V. (2020). SIIM-ISIC melanoma classification, `https://kaggle.com/competitions/siim-isic-melanoma-classification`.

Zhang, B. H., Lemoine, B. and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning, *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, USA*, pp. 335–340.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V. and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints, *arXiv:* 1707.09457.

Agnieszka Mikołajczyk-Bareła, PhD, is an author of datasets, scientific papers, and other publications, holding numerous scholarships and awards. She works on large language models like the first Polish GPT model called TRURL. She is a co-organizer of the PolEval2021 and PolEval 2022 tasks with punctuation prediction and restoration. She organizes and actively contributes to the scientific community in her free time: she has managed and led the team during the HearAI project focused on modeling sign language (`hearai.pl`). She is also a former organizer and a team leader at the open-source project detectwaste.ml.

Maria Ferlin received her BSc and MSc degrees in 2018 and 2019, respectively, in the field of control engineering and robotics at the Warsaw University of Technology. Currently, she is enrolled in PhD studies at the Gdańsk University of Technology in the field of medical applications utilizing machine-learning approaches. Her research is focused on the applicability of machine-learning systems in terms of trustworthiness and interpretability. Moreover, she actively participates in non-profit projects aiming at ethical use of AI.

Michał Grochowski graduated from the Faculty of Automatic Control and Robotics at the Gdansk University of Technology (GUT). In 2004 he received a PhD degree in automatic control and robotics there. In 2020 he received his DSc in the field of automation, electronics, and electrical engineering. Currently, he is a professor and the head of the Department of Intelligent Control and Decision Support Systems at the same university. His present research is focused on computational intelligence and machine learning methods and their utilization in decision support, data analysis, fault detection, and diagnosis systems.

Received: 22 May 2024
Revised: 3 October 2024
Accepted: 12 November 2024