

# Global explanations for discovering bias in data

Agnieszka Mikołajczyk\*, Michał Grochowski, Arkadiusz Kwasigroch

Department of Electrical Engineering, Control Systems and  
Informatics, Gdańsk University of Technology, Poland

## Abstract

*In the paper, we propose attention-based summarized post-hoc explanations for detection and identification of bias in data. We propose a global explanation and introduce a step-by-step framework on how to detect and test bias. Then, the bias is evaluated with proposed counterfactual approach to bias insertion. Because removing the unwanted bias is often a complicated and tremendous task, we automatically insert it, instead. We validate our results on the example of the skin lesion dataset. Using the method, we successfully identified and confirmed part of the possible bias-causing artifacts in dermoscopy images. We confirmed that the commonplace black frames in the training dataset images have a strong influence on the Convolutional Neural Network's prediction. After artificially adding a black frame to all images, around 22% of them changed the prediction from benign to malignant. We have shown that bias detection is an important step of making more robust models, and we discuss how to improve them.*

## 1. Introduction

In recent years, deep neural networks (DNNs) achieved state-of-the-art performance in various tasks. Currently, in contrast to shallow models exploited in the past, most of deep systems extract features automatically, and to do that, they tend to rely on a vast number of labeled data. Whereas the quality of dataset used to train neural networks has a significant impact on the model's performance, those datasets are often noisy, biased, and sometimes even contain incorrectly labeled samples. Moreover, DNNs usually have tens of layers with millions of parameters, and very complex latent space, which makes them very hard to interpret.

Nevertheless, those fragile black-box deep machine

learning models are used to solve sensitive and critical tasks, where the demand for clear reasoning and correct decision is high. Hence, there is raising awareness towards robust learning, formal verification, and extensive testing of models. However, without knowing that data is biased, training the model is a tricky and challenging task.

In the paper, we propose to detect bias in data with attention-based local-summarized global explanations coming from post-hoc Explainable Artificial Intelligence (XAI). We name our method GEBI – **G**lobal **E**xplanations for **B**ias **I**dentification. We focus on image classification and test it on the skin lesion recognition task, but GEBI can be applied to any other problem as well.

The proposed global explanation method is an improvement of the first global analyzer dedicated to summarizing attention-based explanations automatically (Spectral Relevance Analysis - SpRay [1]). We introduce and propose the solution to the previously unnoticed problem of biased XAI, which strongly focuses on the localization and shape of the model's attention but completely ignores an essential part of the explanation: why the attention focuses there. Our improved algorithm of global, relevance-based summarized post-hoc explanations for discovering biases in data takes inspiration in how humans analyze visual explanations: an attention map and input image altogether. In particular, the paper describes a novel GEBI method of global post-hoc explainability to help explain deep neural network decisions to justify them, to control their reasoning process, and to discover new knowledge. Moreover, we propose a simple framework on how to measure the impact of possible bias-causing artifacts. Because removing the unwanted bias is often a complicated and tremendous task, we automatically insert it, instead. Then, we measure how the prediction changed after such bias insertion.

Our major contribution includes:

- a proposition of GEBI method which improves SpRay by analyzing an explanation (attention map) along with the input,
- a proposition of a counterfactual approach for bias

---

\* Correspondence to agnieszka.mikolajczyk@pg.edu.pl

testing with our bias insertion algorithm,

In the Related works section, we bring closer the subject of explainable artificial intelligence and briefly review what approaches have been made in the past to uncover bias in data collections. Then, in the next section, we give a detailed methodology description. In the Experiments section, we present how our algorithm works on the example of a skin lesion dataset. We have manually examined detected clusters and analyzed them to find prediction patterns. Then, after detecting artifacts that might cause bias, we have measured how the presence of such artifact changes the predictions. Finally, we discuss our results and propose how to improve the biased model.

## 2. Related works

### 2.1. Explainable Artificial Intelligence

One of the ways of categorizing XAI methods is to divide them into local and global explanations. The local analysis aims to explain a single prediction, whereas a global one tries to explain how the whole model works in general [2].

The subcategory of local visual explanations covers methods like attention maps (heatmaps, saliency maps, relevance maps) [3], visualizing class-related patterns [4], or explaining by example [5]. An interesting branch of visual explanations is the category of methods based on decomposition [6]–[8] that in contrary to optimization-based methods [10] or techniques based on sensitivity analysis allow building self-consistent attention maps and consistent both in the space of models in the input-domain [9]. For instance, Layerwise Relevance propagation (LRP) [8], can be used to generate attention maps that show on which parts of an image a classifier focused the most. Local explanations are now an actively researched topic.

In contrary, global analyzers are still a small part of XAI methods, but still, some existing methods can be used to find repeating errors in predictions. Global explanations are not only an essential tool to discover abnormalities in the whole model, but in fact, this is also a tool for comparing different models and even different datasets. One of the very first semi-automatic global explanation methods is Spectral Relevance Analysis [1]. We introduce a reader to LRP and SpRAy below.

**Layer-Wise Relevance Propagation** The general idea is to measure how pixels contribute to the positive and negative output by decomposing an input. Hence, the goal is to attribute a contribution, in other words, the relevance  $R_d$ , to each pixel  $x_d$  of an image  $x$  to a corresponding prediction. Bach et al. [3], propose to do that by decomposing the prediction  $f(x)$  to a sum of input dimensions (pixels). In the convolutional neural network, the first layer are the inputs

(pixels of image), and the last layer is the real-value prediction  $f(x)$  of the classifier. The idea is to find and assign each relevance score  $R_d(l)$  to each neuron at the  $l$  layer, starting from the last layer which is classifier output  $f(x)$  and moving backward to the first input layer  $x$ , while holding global conservation property  $\sum_i R_i = f(x)$ . Those relevance scores can be visualized in a form of so called attention maps.

**Spectral Relevance Analysis** uses local explanations in the form of attention maps for generating a summarized explanation of how the model works. Generated attention maps are later grouped with spectral clustering. Spectral clustering reveals some hidden patterns forming on the attention maps and allows a user to screen through a large dataset to find co-occurring patterns without manual, time-consuming analysis of individual explanations. The final step in this semi-supervised method is a visual inspection of interesting clusters by a user.

The steps of the method are as follows: **Step 0.** Select batch of samples for analysis. **Step 1.** Compute relevance scores with LRP and save them as attention maps. **Step 2.** Normalize and preprocess attention maps **Step 3.** Perform spectral clustering on normalized samples. **Step 4.** Find interesting clusters with eigengap analysis. **Step 5 (optional).** Visualize achieved clusters with t-SNE.

Results presented by Lapuschkin et al. [1] were very impressive, but the fact that the SpRAy method clusters the data based only on the attention maps makes the method itself biased. This biased XAI focuses only on the shape of detected objects on the attention maps, localization of those shapes, and sometimes textures, while not considering what is under the attention map. While localization, as well as the shape of the attention regions are essential, the information why the model focused on that area is even more critical. The algorithm should take into account the colors under the attention, the textures, and what exactly is there. This paper proposes an improvement of this method and delivers in-depth research regarding this newly-formulating branch of global explainability methods. We provide details in the Methodology section.

### 2.2. Bias in data

Bias in data is defined as any trend or deviation from the truth in data collection that can lead to false conclusions [11]. Bias in data might cause misinterpretation not only for highly data-dependable deep learning models but also for human experts, which makes identifying and avoiding bias in research a long-standing topic in general [12]. Most of the practical ML-related research problems start with a study on a whole population, e.g., a population of benign vs. malignant skin lesions. However, in practice, it is impossible to gather all possible cases from the whole

population. The population analysis uses only a small representative group of individuals. If the sample is not well represented, conclusions also will not be generalizable [11] – for instance, if all sensitive asthma patients were carefully hospitalized during their pneumonia and hence never got any complications, the model might conclude that asthma prevents from complications [13]. The influence of bias in data can be noticed in numerous applications. There is a known problem of gender and racial bias in sentiment analysis [14]. It appears that certain groups of people seem to be using specific words more often than others. When we want to analyze a slang, it could be a welcomed result, but in case of unpolarized text, we could get a wrong prediction, that was based only on the gender, race or age of the person speaking [15]. Similarly, in the case of creditworthiness prediction in the United States, predicted the credit risks were different depending on the race [16]. Even when it comes to widely accepted by the ML community benchmark datasets, a bias still can be found. For instance, the ImageNet [17], has many underrepresented classes. A car class is represented mostly by racing cars [18], and also, as reported, ImageNet seems to be undesirably biased towards texture [19].

When it comes to skin lesion datasets [20], [21], the possible bias was already discovered in 2019 [22], but the exact source of it was not identified. The common goal of skin lesion recognition is to classify skin lesions into benign or malignant type, or to specify its exact type, to find dangerous changes early. Dermatologists support their diagnosis by careful analysis of skin lesions with a broad set of dermoscopic methods along with their deep intuition. In contrast, deep models find relevant features during the training based on the provided dataset. Bissoto et al. [22] suspected that a widely used dataset of skin lesions might be biased, and hence they conducted a series of experiments regarding that matter. They used segmentation masks of each lesion and modified dataset by covering each lesion with a black segmentation mask. That modified dataset is used to train a convolutional neural network to differentiate benign and malignant skin lesions – but without any lesions in the dataset. Surprisingly, results showed that a model trained and tested on data without any lesions could classify them correctly with a performance (AUC) above 73%, which is only ten-percentage points less than performance on the original data. Because the shape of the skin lesion is a significant feature for dermatologists, the researchers changed segmentation masks to black boxes and repeated the experiments. The results were even more surprising because the performance was almost the same as in the previous tests. Those results raise an important question: should we blindly trust the machine learning systems, based only on the performance metrics? Those metrics are always generated based on the same, biased test set, which makes

internal validity doubtful. However, even if we know that the bias exists, we should ask ourselves another question: **what** exactly is the bias source and **how** to eliminate or at least mitigate it?

Barata et al. [23] tried to find the source of bias by manual analysis of skin lesions. They concluded that the model might be sensitive to the look of a skin lesion but also black frames, skin tone, and some artifacts such as white reflections. However, the manual inspection is time-consuming and might lead to overlooking some important large-scale patterns. Discovering the root of this problem is the first step to designing more robust and trustful systems. This paper attempts to answer those questions by providing a methodology that will help to find the origin of the bias in data.

### 3. Methodology

In this section, we propose an improvement of the spectral relevance analysis and show how it might be used for bias identification.

#### 3.1 Detecting bias with GEBI

On the example of the skin lesion dataset, we present that it allows detecting a few possible bias-causing artifacts.

The steps of the method are as follows:

**Step 0.** Select samples for analysis.

**Step 1.** Compute attention maps for the samples of one class.

**Step 2.** Normalize and preprocess **both input samples and accompanying attention maps** in the same manner.

**Step 3.** Reduce the dimension of each input sample and relevance map with dimension reduction algorithm, for instance, with Isomap

**Step 4.** Concatenate each reduced sample with a relevant reduced attention map.

**Step 5.** Perform spectral clustering on reduced vectors.

**Step 6.** Visualize and analyze obtained clusters.

**Step 7.** Formulate and test the hypothesis with bias insertion algorithm

Step 0 is an integral part of the analysis. Only one class should be analyzed at once to detect bias. Analyzing more than a single class at once should be performed only in some individual cases, e.g. when looking for possibly bias-causing artifacts that could exist in every class.

In the first step we apply LRP to selected input images, but any method of attention maps generation could be used.

In step 2, we normalize images with contrast enhancement to bring up some clinical attributes. An additional problem is white-balance hence every image was preprocessed with adaptive histogram equalization.

In step 3, instead of reducing dimensionality by image-downsizing, we used Isomap algorithm. In original SpRay method the simple downsizing of an image might

cause loose of important small-sized features in the image. The number of features should be selected for each kind of problem individually. In our case, we achieved the best results when the number of features of input images was around two times lower than the number of attention features. A number of features selected also depends on the chosen clustering method – many of the clustering methods have a problem with working on the high-dimensional data.

Step 4 is a simple concatenation of input features with attention features. This is a new step, because SpRAY method does not analyze input features along with attention features.

Step 5 covers grouping of a concatenated vector with features extracted from both images and attention maps. The only difference in this step is what is clustered: we are

clustering concatenated feature vectors instead of clustering downsized attention maps. Moreover, we suggest that a user can select any clustering method, not only the spectral clustering.

In step 6, we visualize clusters with Isomap algorithm in the 3d-space, and leave the analysis for the user.

Then, in the new last step, user formulates a hypothesis that defines the root of the bias, for instance, that the presence of an artifact in image cause bias. The influence of the bias in data can be tested with our bias insertion algorithm. We describe how to test the bias in the next subsection.

The workflow of the method is presented in figure 1. The visualization of the achieved clusters is shown in figure 4.

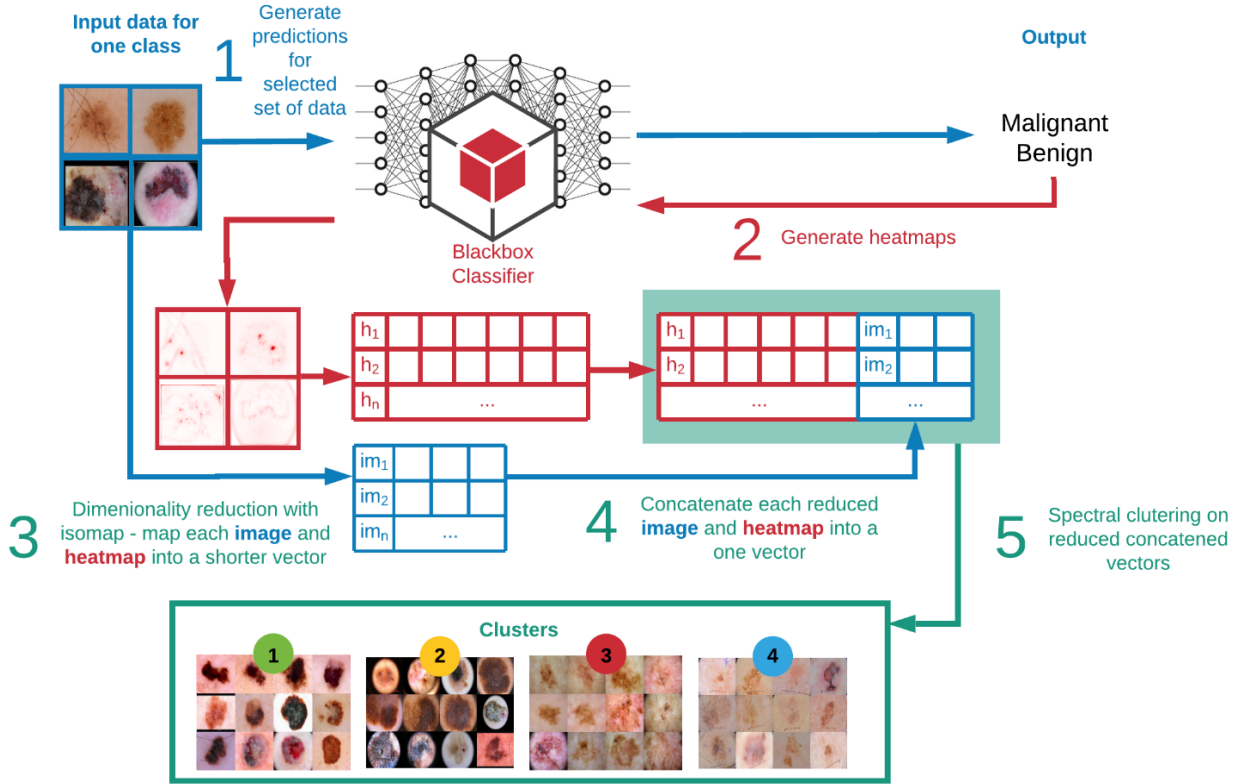


Figure 1: Illustration of the idea behind improved spectral attention analysis

### 3.2 Bias testing – counterfactual approach

We also propose to test the influence of possible bias by bias-insertion experiments. At first, similarly as in [24], the user has to find an answer to the question: what might cause bias? The answer can be formulated as the hypothesis and then, once the cause is identified, it should be carefully confirmed. For example, in the computer vision task, if in

the task of dog vs. cat classification, there is one cluster with dogs behind bars and no clusters of cats behind bars, one can think that bars might be a significant feature while classifying dogs. Then, we could add bars to every image in the dataset and change how the prediction score changes. If the average change of prediction is high, it means that the hypothesis was correct. Otherwise- possibly not.

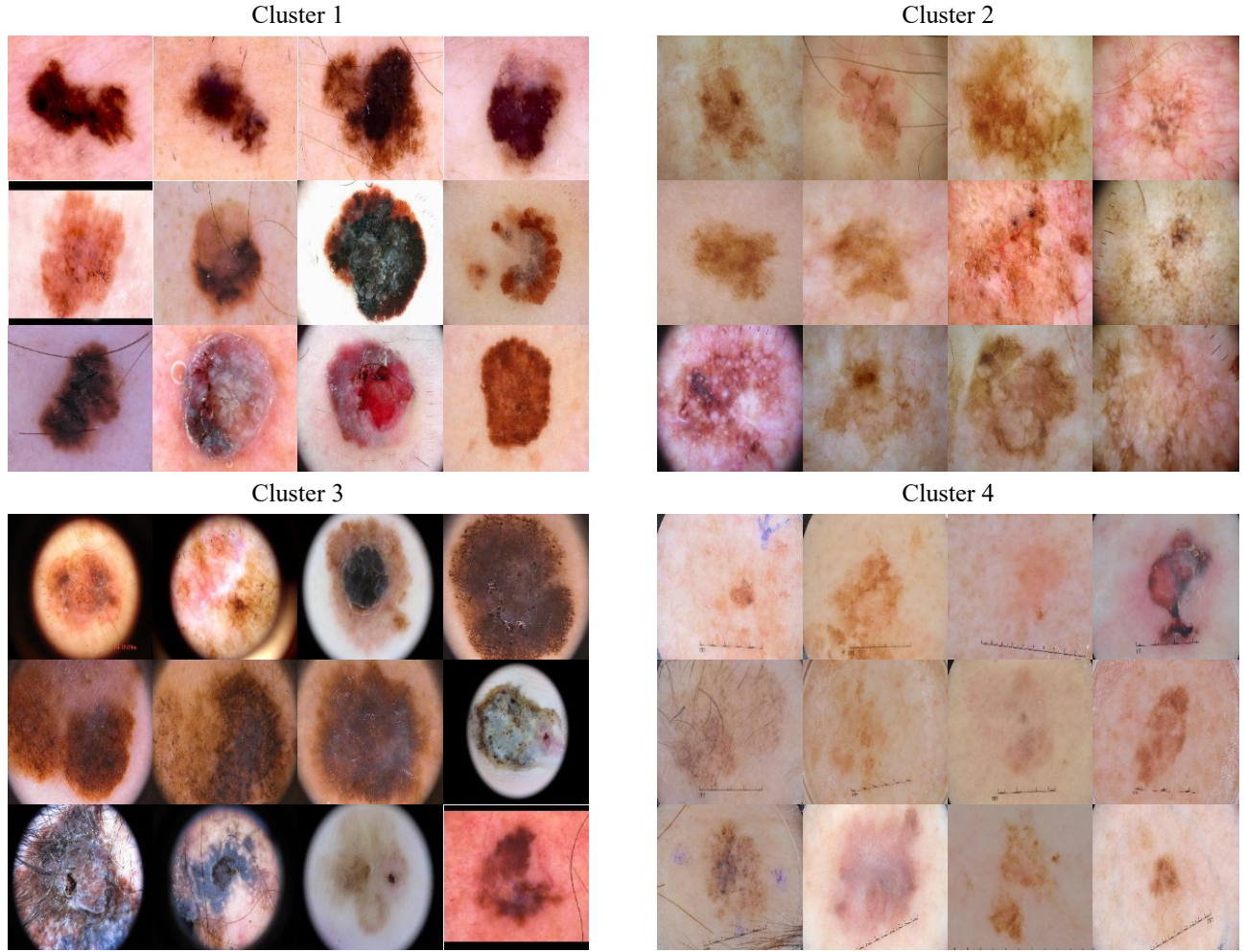


Figure 2: Example images from four different clusters discovered with modified spectral clustering on concatenated reduced attention maps and input images. Cluster 1 shows mostly dark skin lesions with clear border; Cluster 2 shows very textured skin lesions with numerous visible structures; Cluster 3 contains images with black frames; Cluster 4 contains mostly light-colored skin lesions with metrics, a single hair, blue markings (the first column was marked with red rings to show possibly misleading artifacts)

The process of bias insertion is also similar in different types of models and data. In the case of tabular data, a bias, for instance, in the assessment of a client's creditworthiness, one could change the sex of a client and check if the model's output changed. That operation could be tested on many records, and the differences in prediction be calculated and averaged afterward. Such a test would also be a crucial procedure for measuring possible unfairness.

In Natural Language Processing, in the case of sentiment analysis, we could similarly insert bias. We could switch selected word that, in our opinion, does not change the polarity of the text, to other, of the same meaning, and check the change in prediction. For instance, many papers show that sentiment analysis seems to be biased by gender or race.

## 4. Experiments

### 4.1 Implementation details

We followed a training procedure presented by Mikołajczyk et al. [25]. In the experiments, we have fine-tuned widely used DenseNet121 [26] architecture with traditional data augmentation (rotation, zoom, shear, reflection) and early stopping. Final network had an AUC score of 0.869 on a test set.

We have tested several types of attention maps generation such as LRP, LRP flat A, LRP flat B, Deep Taylor Decomposition (DTD). The results were similar for every type of attention visualization. Attention maps presented in this paper are generated with DTD [27]. Each image is preprocessed with histogram equalization and



contrast-enhancing. Then, to reduce dimensionality, we have used Isomap algorithm [28]. We have reduced each image to the 10-dimensional vector and each attention map to 20-dimensional vector.

Then, we concatenated reduced vectors together and clustered all vectors. We have tested the DBSCAN, k-means, spectral clustering, affinity propagation, mean shift, OPTICS, and birch methods [29]. In case of skin lesion dataset we approached, with huge intra-class variation and small interclass variation, where images seem to be very similar, the best results were achieved with spectral clustering and traditional k-means. Presented in the paper results were achieved by using a spectral clustering algorithm [29]. We used elbow method [30] to estimate the optimal number of clusters – four clusters seemed to be the most suitable solution. Finally, we examined clusters and additionally visualized the results by presenting 3d animated plots.

#### 4.2 Identification of prediction strategies

With our proposed method, we have identified four different clusters. Each cluster reveals unique characteristics in the look of analyzed data set, related with the skin tone, skin lesions, but also with the presence of the unwanted artifacts. The first and the second cluster seem to group images based on the skin lesions similarity, which is a welcomed result in this case.

On the contrary, the third cluster mostly gathers images with round or rectangular black frames. The last, fourth cluster contains mostly light skin lesions, very often with a visible ruler. The proposed method is semi-automated, so a field expert should analyze the clusters. In this case, we focused our attention on clusters 3 and 4, where some additional artifacts grouped those images. Hence, we could formulate a hypothesis that black frames and ruler marks might cause possible bias in models. To check if those features have a significant influence on the prediction, we conducted another experiment: inserting possible bias and testing its influence.

#### 4.3 Inserting possible bias

To test how the prediction will change if feature on the image is present, we propose to compare models outputs for the same image with and without the given feature. Because removing artifacts from the images is very complicated task, we have decided to insert it, instead. We wanted to mimic the real artifacts that we found in the dataset, as well as to add a new one for comparison.

**Black frames** were added in the same way to all images, without any variations in size and position. Frames can be commonly found in numerous images, and are often recognized as unwanted artifacts [31]. Their visibility usually depends on the type of used dermatoscope.

**Rule marks** were prepared beforehand and placed on the image in slightly different sizes, angles, and positions. Rulers are usually used by a doctor to show the size of a skin lesion on the dermoscopic image.

**Red circles** cannot be naturally found in the ISIC archive, SD-198, and Derm7pt datasets. For a clear comparison, we have also placed those markings. The single red circle was placed randomly in the image, both within the skin and lesion areas.

We present the examples of such modifications in figure 3.

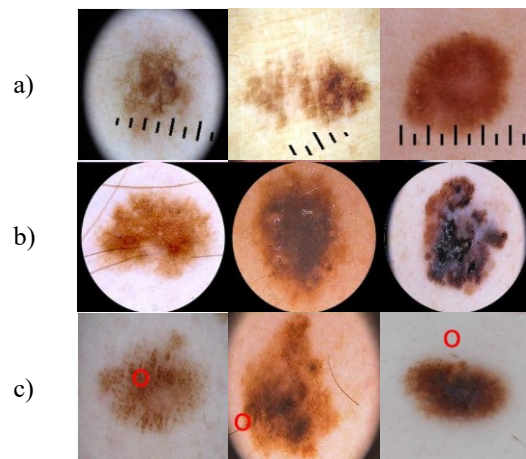


Figure 3: Modified examples by insertion of artificial bias: a) ruler markings, b) black frames, c) red circles

#### 4.4 Testing the bias influence

After we modified the dataset by placing in images selected artifacts, we began testing our hypothesis by asking a question: *Are those artifacts causing bias in model's performance?* To answer this question we measured how predictions will change depending on the presence of the artifacts. In our case, we checked what will happen if we add to all images: black frames, black ruler, and red circles.

The idea behind testing the bias influence is presented in the figure 4.

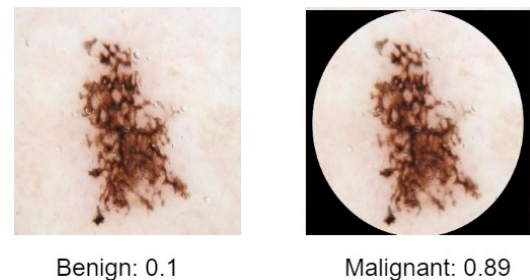


Figure 4: Idea behind the counterfactual bias insertion. Inserting the artifact changes the prediction score by 0.79 points

We have calculated the differences in predictions for 884 randomly selected malignant and benign skin lesions, separately for each type of transformation. The difference in prediction score is simply a difference between the prediction on unmodified image and the prediction after artifact insertion. Hence, the higher difference, the higher impact of the tested artifact on the final prediction. We averaged the results and gathered them in table 1.

Table 1: Results in percentage points

Added Feature	Type	Average Change in prediction*	Maximum Change in prediction
Ruler	Mal	2.21	22.01
	Ben	1.23	19.91
Frame	Mal	<b>30.77</b>	<b>62.43</b>
	Ben	<b>32.04</b>	<b>63.66</b>
Red circle	Mal	2.27	15.51
	Ben	1.50	12.78

The highest differences in model predictions were caused by adding a black frame to an image, whereas the ruler and the red circle did not change prediction scores much on average. The interesting part of this experiment is that the black frame did not change how skin lesion looks like in any way, but the changes in predictions were very high for both malignant and benign skin lesions. On average, every output changed by 33%. Moreover, adding this type of artifact seems to be biasing the model toward classifying a skin lesion as a malignant. A number of images classified as malignant raised from 31 to 228 when tested on a benign dataset. Hence, 197 out of 884 skin lesions switched prediction to malignant, considering the classification threshold equal to 0.5. This means that around 22.29% of the checked skin lesion changed their classes after introducing such slight modification. Black frames usually do not cover any part of skin lesion, hence such a significant change in prediction score should wake up some doubts in the models' behavior. It is a very interesting notice, which should be taken into consideration while training new models in the future.

Ruler marks caused, on average, only a slight difference in model predictions of around 1.23 and 2.21 pp., but still, it might be a dangerous reaction in some cases when the change in prediction is high. What is interesting, in the case of those markings, there were a few cases that changed the model's decision from malignant to benign, in both subsets.

Adding a red circle did not make a huge difference in the output, but surprisingly, it was quite similar to the average change for ruler placement. A small number of approximately 1.5% of images switched prediction from benign to malignant. We hypothesize that the reason is that

part of the malignant skin lesions tends to have atypical structures like blobs, dots, and streaks [32]. The red circle might be similar in some way to dermatological attributes. Those structures are defined, for example in 7-check point list or in ABCD rule [32].

#### 4.5 Codes and data availability

Our source code, user-friendly tutorials, and generated attention maps for quick experiments are available at [github.com/agamiko/gebi](https://github.com/agamiko/gebi). Source code for adding bias such as black frames and rule marks are available at [github.com/agamiko/bias-insertion](https://github.com/agamiko/bias-insertion). The source code for LRP is available at [github.com/albermax/innvestigate](https://github.com/albermax/innvestigate). Source code for clustering and Isomap reduction is available at [scikitlearn](https://scikitlearn.org). Dataset of skin lesions is available at [isic-archive.com](https://isic-archive.com).

#### 5. Discussion

Currently the subject of interpretable and explainable artificial intelligence is constantly raising. More and more people are aware that machine learning and deep learning models requires extensive testing and that its inner work should be known. Unfortunately, bias in data is still not widely discussed. Authors of real-data applications usually test their models only in terms of accuracy performance or computation efficiency. That approach to production ML should be changed, especially when tackling safety-critical systems.

Our paper presents a new method that might be used for detection of bias in data collection, as well as in model's behavior. We introduced the problem of biased XAI which could result with the wrong interpretation of models decision-making process. Then, we presented our results on the example of skin lesion classification task. With a few, simple but effective modifications of the SpRAy, our GEBI method gained a significant improvement. The GEBI for example, allowed detecting that black frames, commonly existing in skin lesion dataset images, have a significant impact on models predictions. We have tested our hypothesis regarding bias in skin lesion dataset with our bias insertion algorithm. In fact, each image was predicted with around 32 percentage points more towards malignant skin lesions when added a black frame, which confirmed a suspicion of many researchers from the past [33]–[35].

However, bias detection and confirmation is just a first step of making more reliable and robust models. The next step should be a further development of this approach. Improvement can be done e.g. by deleting the bias from datasets. Many researchers have tried artifact removal approach [33]–[35], as a first preprocessing step before classification. Although removing all of the biases is nearly impossible, and does not solve a problem. Another

approach is making model to focus on a right features. This could be done for example with data augmentation, with specially prepared data: in case of speech recognition it can be done by randomly removing parts of recordings with low energy [36]. In our case, it could be done by randomly inserting the bias into images during the training, similarly to online data augmentation.

And finally, we could force a model to look at important parts of data, for example, by attention-training [23]. Such approaches modify loss functions to check not only the models classification performance but also if it focuses on the right regions.

We present our results in open-science manner and attach relevant codes for our method, as well for the bias insertion.

## References

- [1] S. Lapuschkin, S. Wäldchen, ... A. B.-N., and undefined 2019, "Unmasking Clever Hans predictors and assessing what machines really learn," *nature.com*.
- [2] A. B. Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Oct. 2019.
- [3] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," 2015.
- [4] R. Ramprasaath, D. Abhishek, V. R.-C. 2016, and undefined 2016, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization."
- [5] S. Wachter, B. Mittelstadt, and C. Russell, "COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR," *Harv. J. Law Technol.*, vol. 31, no. 2, 2018.
- [6] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," Aug. 2017.
- [7] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-Down Neural Attention by Excitation Backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, Oct. 2018.
- [8] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K. R. Müller, "Layer-Wise Relevance Propagation: An Overview," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11700 LNCS, Springer Verlag, 2019, pp. 193–209.
- [9] W. Samek, A. Binder, ... G. M.-I. transactions on, and undefined 2016, "Evaluating the visualization of what a deep neural network has learned," *ieeexplore.ieee.org*.
- [10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier."
- [11] A. M. Šimundić, "Bias in research," *Biochem. Medica*, vol. 23, no. 1, pp. 12–15, Feb. 2013.
- [12] C. J. Pannucci and E. G. Wilkins, "Identifying and avoiding bias in research," *Plast. Reconstr. Surg.*, vol. 126, no. 2, pp. 619–625, Aug. 2010.
- [13] R. Ambrosino, B. G. Buchanan, G. F. Cooper, and M. J. Fine, "The use of misclassification costs to learn rule-based decision support models for cost-effective hospital admission strategies.," *Proc. Annu. Symp. Comput. Appl. Med. Care*, pp. 304–308, 1995.
- [14] M. Thelwall, "Gender bias in sentiment analysis," *Online Inf. Rev.*, vol. 42, no. 1, pp. 45–57, 2018.
- [15] P.-S. Huang *et al.*, "Reducing Sentiment Bias in Language Models via Counterfactual Evaluation," Nov. 2019.
- [16] M. Hardt Google, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," 2016.
- [17] J. Deng, W. Dong, R. Socher, L. Li, ... K. L.-2009 I. conference, and undefined 2009, "Imagenet: A large-scale hierarchical image database," *ieeexplore.ieee.org*.
- [18] A. Torralba, A. E.-C. 2011, and undefined 2011, "Unbiased look at dataset bias," *ieeexplore.ieee.org*.
- [19] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," *7th Int. Conf. Learn. Represent. ICLR 2019*, Nov. 2018.
- [20] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, Mar. 2018.
- [21] X. Sun, J. Yang, M. Sun, and K. Wang, "A Benchmark for Automatic Visual Classification of Clinical Skin Disease Images."
- [22] A. Bissoto, M. Fornaciali, E. Valle, and S. Avila, "(De)Constructing Bias on Skin Lesion Datasets," Apr. 2019.
- [23] C. Barata, J. S. Marques, and M. E. Celebi, "Deep Attention Model for the Hierarchical Diagnosis of Skin Lesions."
- [24] C. J. Anders, T. Marinč, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin, "Analyzing ImageNet with Spectral Relevance Analysis: Towards ImageNet un-Hans'ed," Dec. 2019.
- [25] A. Mikołajczyk, M. G.-2019 24th International, and undefined 2019, "Style transfer-based image synthesis as an efficient regularization technique in deep learning," *ieeexplore.ieee.org*.
- [26] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Aug. 2016.
- [27] G. Montavon, S. Lapuschkin, A. Binder, W. S.-P. Recognition, and undefined 2017, "Explaining nonlinear classification decisions with deep taylor decomposition," *Elsevier*.
- [28] M. Balasubramanian, "The Isomap Algorithm and Topological Stability," *Science (80-. )*, vol. 295, no. 5552, pp. 7a – 7, Jan. 2002.
- [29] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, Jun. 2015.
- [30] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, 2013.
- [31] J. Jaworek-Korjakowska, "A Deep Learning Approach to Vascular Structure Segmentation in Dermoscopy Colour Images," *hindawi.com*, 2018.
- [32] R. H. Johr, "Dermoscopy: Alternative melanocytic algorithms - The ABCD rule of dermatoscopy, menzies scoring method, and 7-point checklist," *Clin. Dermatol.*, vol. 20, no. 3, pp. 240–247, May 2002.
- [33] T. Majtner, K. Lidayová, S. Yildirim-Yayilgan, and J. Y. Hardeberg, "Improving skin lesion segmentation in dermoscopic images by thin artefacts removal methods," in *Proceedings of the 2016 6th European Workshop on Visual Information Processing, EUVIP 2016*, 2016.
- [34] A. Sultana, I. Dumitrache, M. Vocurek, and M. Ciuc, "Removal of artifacts from dermoscopic images," in *IEEE International Conference on Communications*, 2014.
- [35] M. E. Celebi, H. Iyatomi, G. Schaefer, and W. V. Stoecker, "Lesion border detection in dermoscopy images," *Comput. Med. Imaging Graph.*, vol. 33, no. 2, pp. 148–153, Mar. 2009.
- [36] C. Kim, K. Kim, and S. R. Indurthi, "Small energy masking for



improved neural network training for end-to-end speech recognition,” Feb. 2020.