Mitigating bias with Targeted Data Augmentations

Agnieszka Mikołajczyk

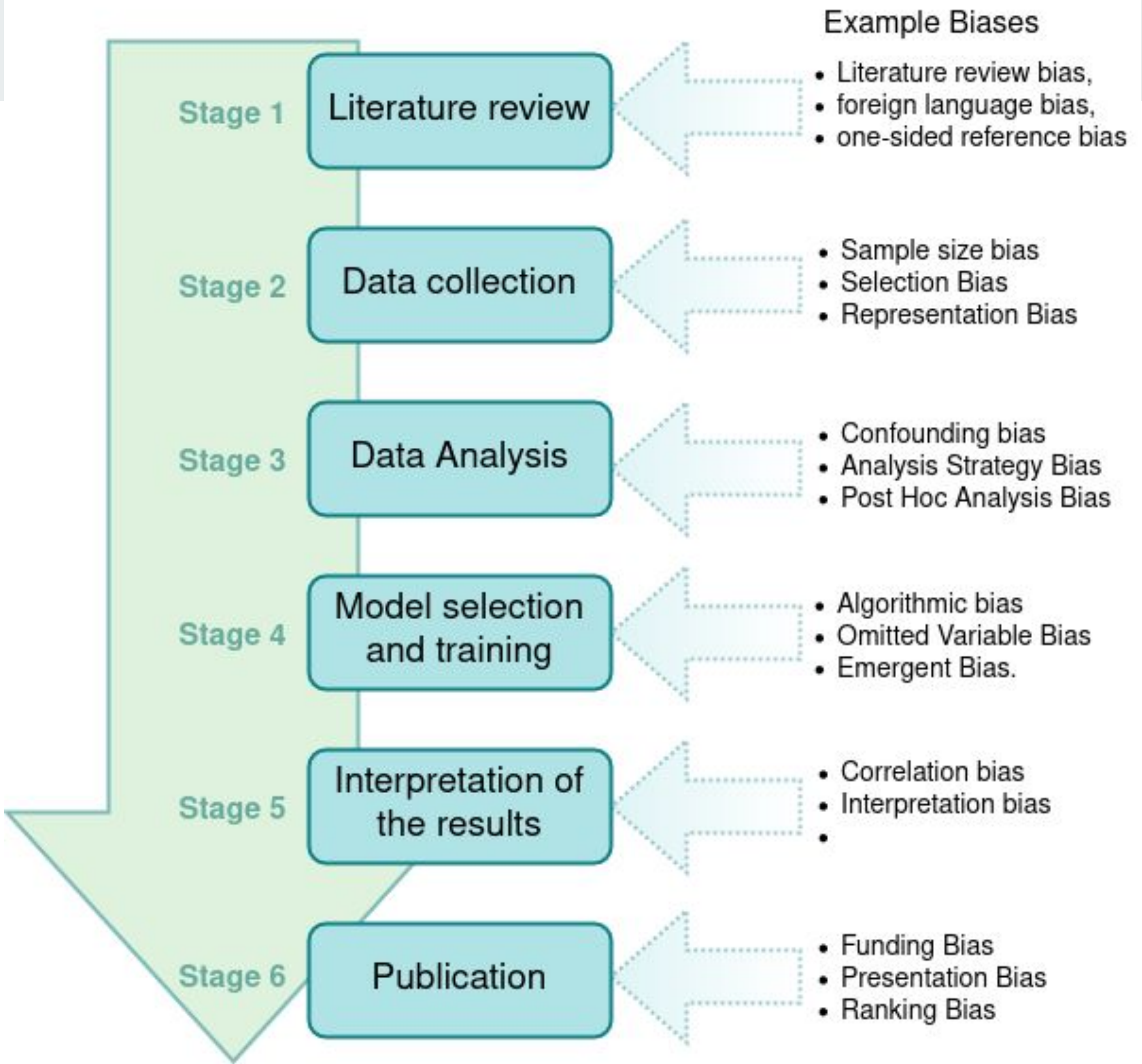
GDAŃSK UNIVERSITY OF TECHNOLOGY

VoiceLab

Plan

- Biases on different stages of ML research
- Skin lesion dataset - is it biased?
- Detecting bias
- Mitigating bias: Targeted data augmentation

Stages of ML project



Stage 1: Literature Review

Literature review bias - incomplete search due to poor keywords and search strategies or failure to include unpublished reports

Foreign language exclusion bias - when publications in foreign languages are ignored

One-sided reference bias: happens when researchers restrict their references to only those studies that support their position

Rhetoric bias - when authors try to convince the reader without any scientific fact

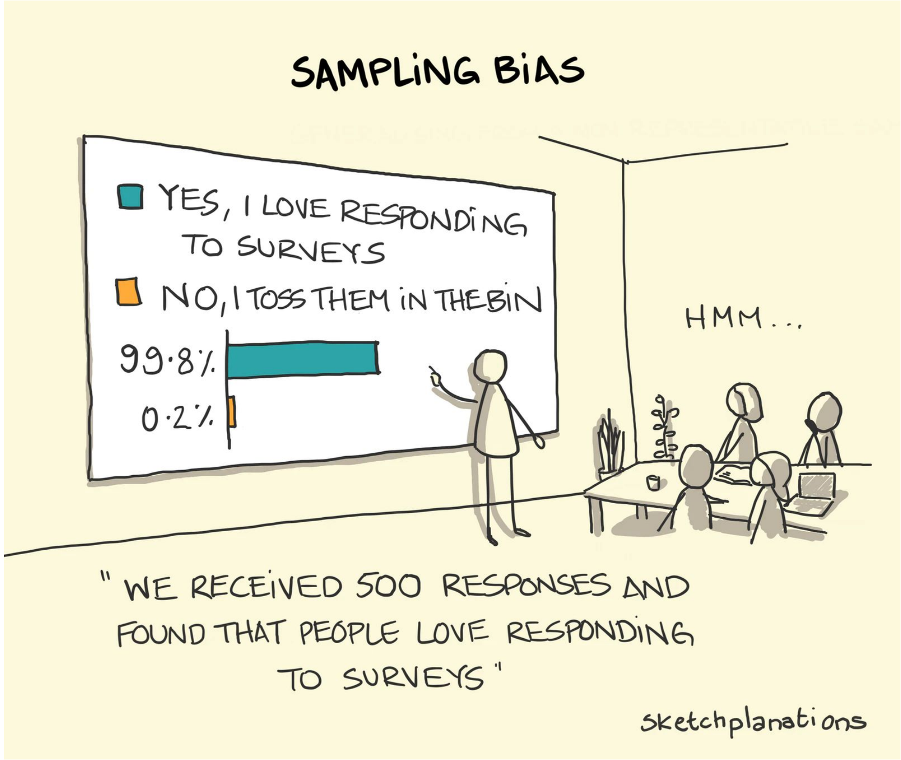
Stage 2: Data Collection

Selection Bias is defined as a deviation of data from the truth resulting from how samples were collected. It can arise when
a) the sampling frame is incomplete or inaccurate,
b) the sampling process was nonrandom, or
c) when some targets were excluded from data collection

I.e. Sampling bias

Stage 2: Data Collection

Sampling Bias (Representation Bias) is a bias in which data is acquired in such a way that not all samples have the same sampling probability, i.e., not all samples are equally likely to be selected in the study



Stage 2: Data Collection

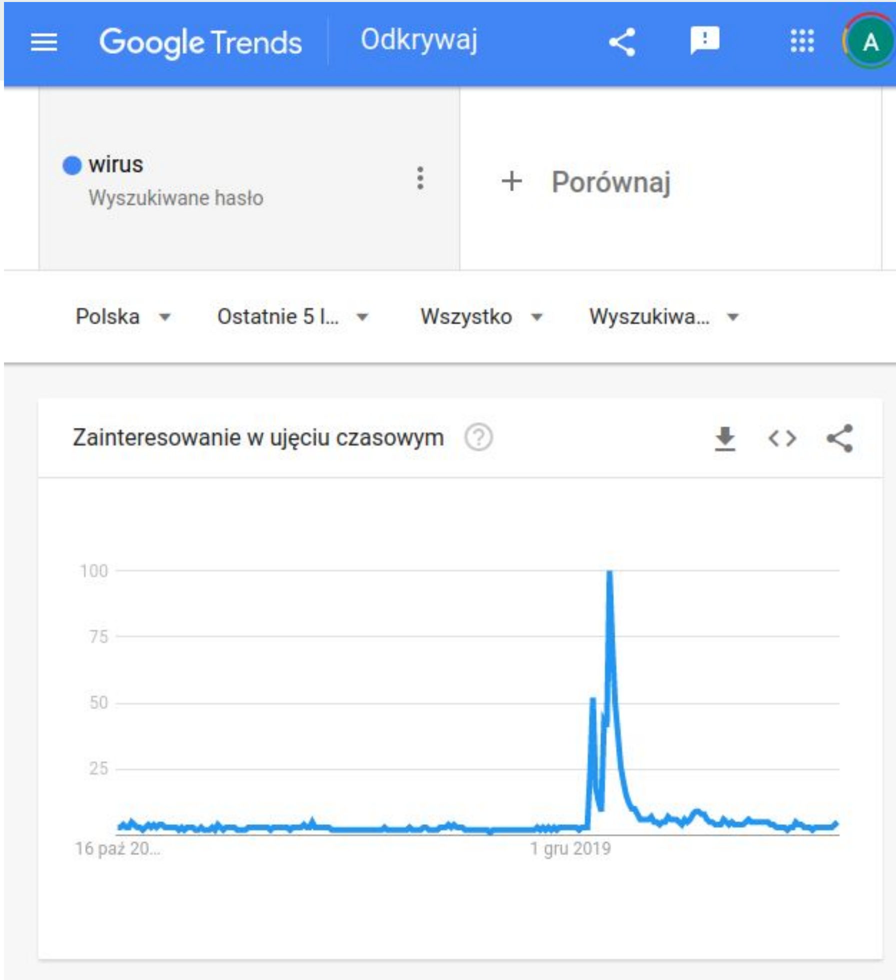
Nonrandom bias exists when the selection process is affected by the human choice, e.g., when sampling is nonrandom

Instrument Bias. This type of bias results from imperfections in the instrument or method used to collect the data

Stage 2: Data Collection

Temporal Bias systematic distortions across user populations or behaviors over time.

The **Popularity Bias** comes from increased public interest in a subject



Stage 2: Data Collection

Observer Bias It owes the name to its own definition: it tends to observe what the observer wants to see.
In ML, observer bias might appear when annotators use personal, subjective opinions to label data, resulting in incorrect annotations.

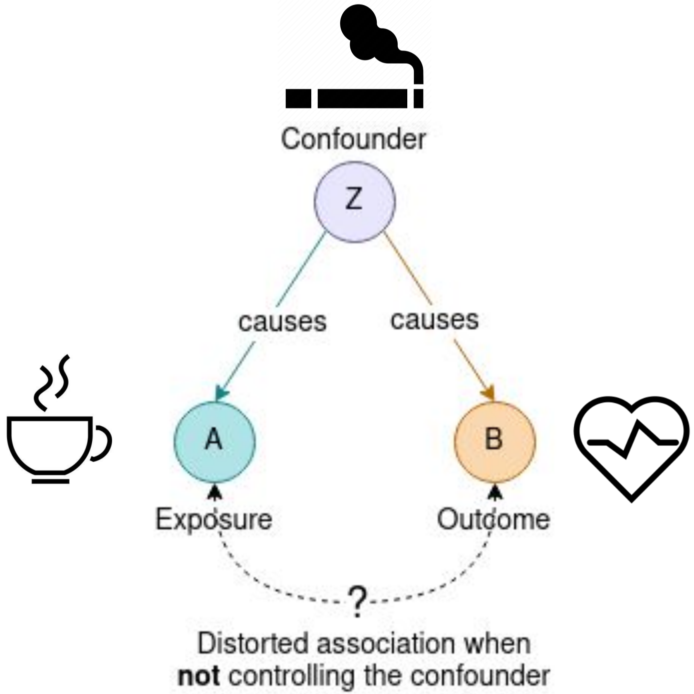
Cyril Burt

From Wikipedia, the free encyclopedia

Sir Cyril Lodowic Burt, FBA (3 March 1883 – 10 October 1971) was an English educational psychologist and geneticist who also made contributions to statistics. He is known for his studies on the heritability of IQ. Shortly after he died, his studies of inheritance of intelligence were discredited after evidence emerged indicating he had falsified research data, inventing correlations in separated twins which did not exist.

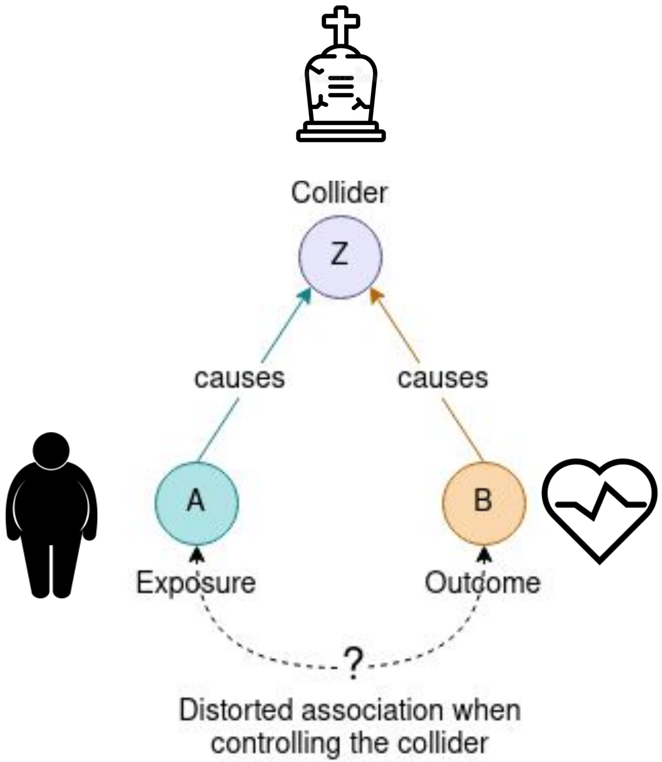
Stage 3: Data Analysis

Confounding bias - Confounder is a variable that influences both the dependent variable (i.e., disease) and independent variable (the factor being studied)



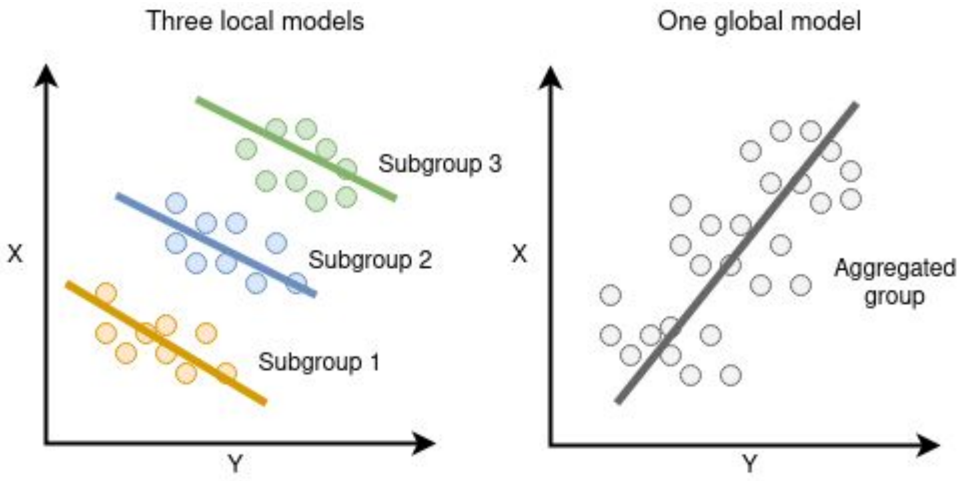
Stage 3: Data Analysis

Collider bias is defined as a causally influenced association created between two or more exposures when a shared outcome (collider) is included in the model as a covariate



Stage 3: Data Analysis

Reversal paradox happens when the association between two (or more) variables can be reversed when another variable is statistically controlled for**.** The most known subtype of reversal paradox is **Simpson's Paradox** which can be observed when the relationship between two variables differs within subgroups, and its aggregation



Stage 4: Model selection and training

Algorithmic Bias - when the model is the source of bias. Some sources also define an algorithmic bias as amplifying and adversely impacting existing inequities in, e.g., socioeconomic status, race, ethnic background, religion, gender, disability or sexual orientation by an algorithm

TABLE 2.1: ProPublica's table (2016) reporting model errors at the study cut point (Low vs. Not Low) for the General Recidivism Risk Scale [?]

COMPAS Risk Prediction	Reoffend	White	Black
High Risk	No	23.5%	44.9%
Low Risk	Yes	47.7%	28.0%

TABLE 2.2: (Low vs. Not Low) for the General Recidivism Risk Scale [3]

COMPAS Risk Prediction	Reoffend	White	Black
High Risk	No	41.0%	37.0%
Low Risk	Yes	29.0%	35.0%

Stage 4: Model selection and training

Emergent bias - this bias typically emerges a while after training is finished, as a result of changing societal knowledge, population, or even cultural values. Moreover, it can emerge when used by a population with different values than those assumed in the design

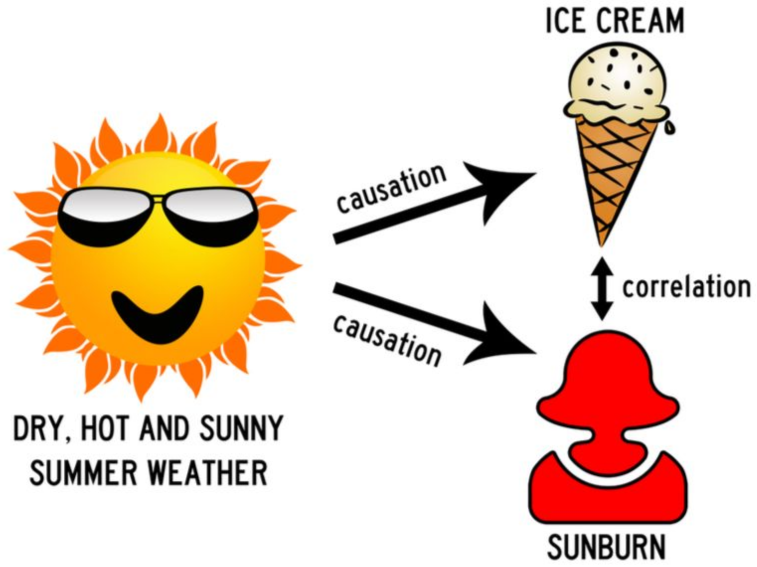
Stage 4: Model selection and training

Deployment Bias - when a system is used or interpreted in inappropriate ways, e.g., a model is used for a different purpose than the initially designed purpose.

Evaluation Bias - a bias that is introduced during the model's evaluation. This includes poorly selected evaluation data or inadequate metrics that do not measure the model's performance

Stage 5: Results interpretation

Correlation bias, also known as Cause-effect bias, which, as the name suggests, happens when the correlation is mistaken with causation.



Source:
https://towardsdatascience.com/correlation-is-not-causation-ae05d03c1f53

Stage 6: Publication

Funding Bias emerges when a party reporting results report them to satisfy the funding agency or financial supporter of the research study

Presentation bias which is defined as a result of how the research topic (information) is presented

"You are completely free to carry out whatever research you want, so long as you come to these conclusions."

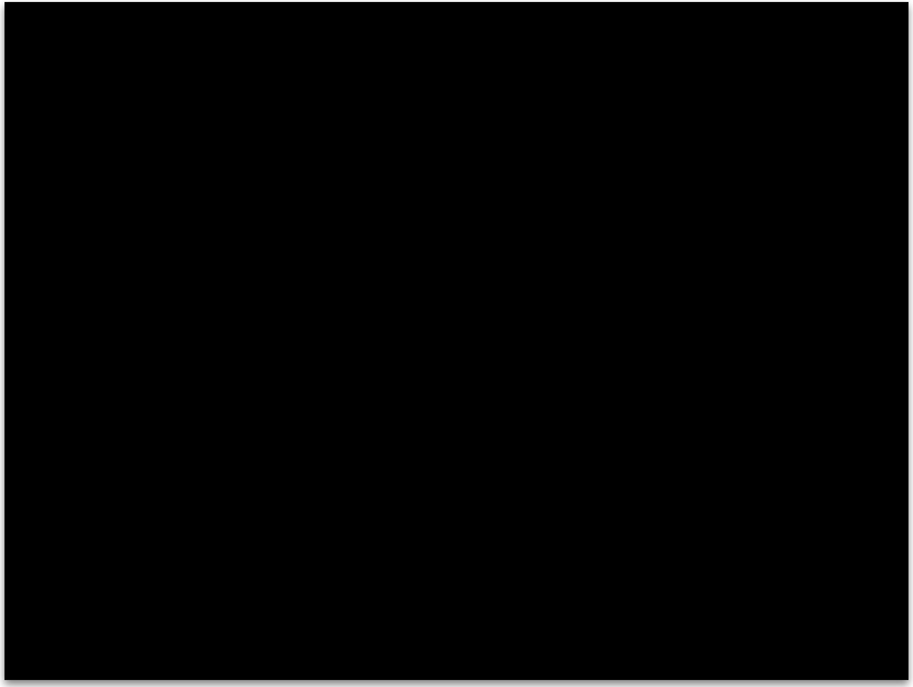
Source: https://twitter.com/clive_bates/status/649696820227108864

Mitigating biases in data and models

Skin lesion classification example

Skin lesion classification example

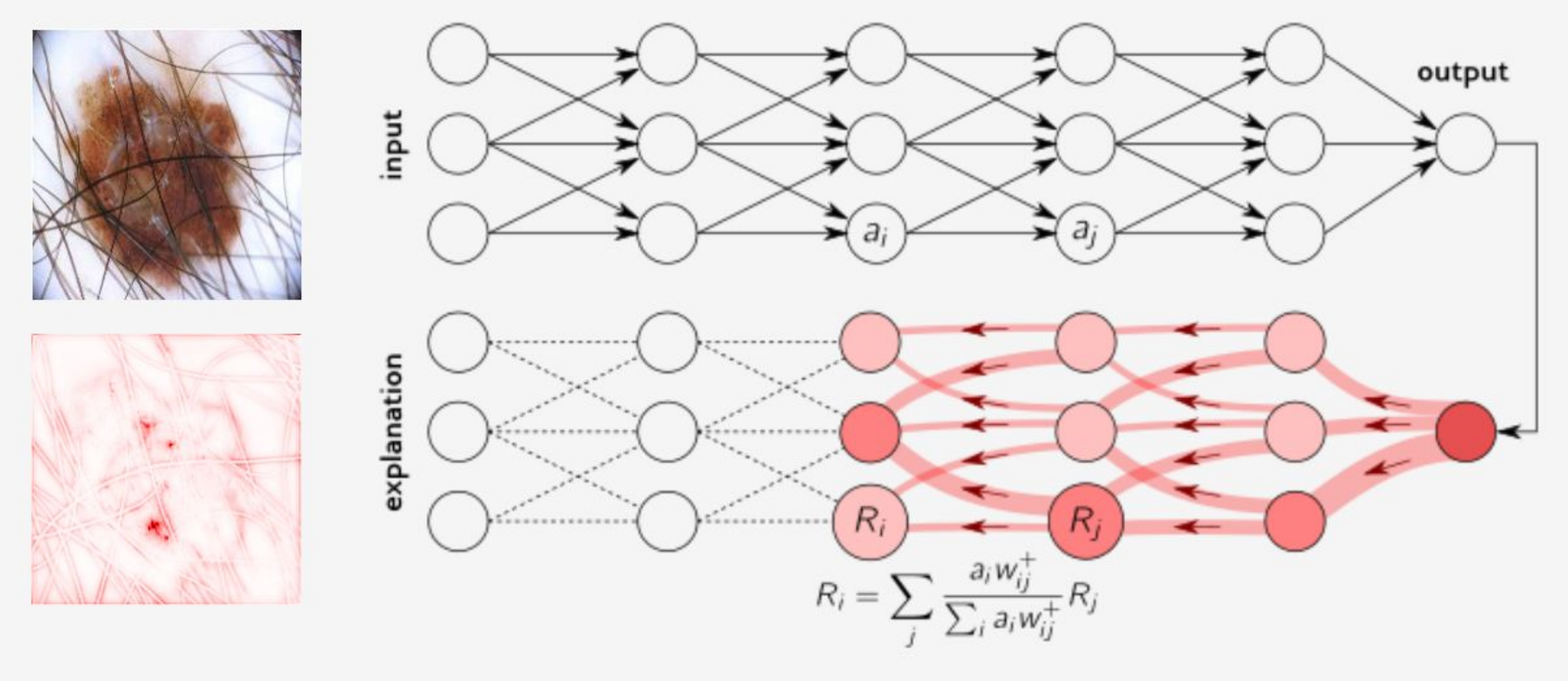
- Benign vs. malignant
- Imbalanced class distribution
- Various artifacts
- Sensitive problem

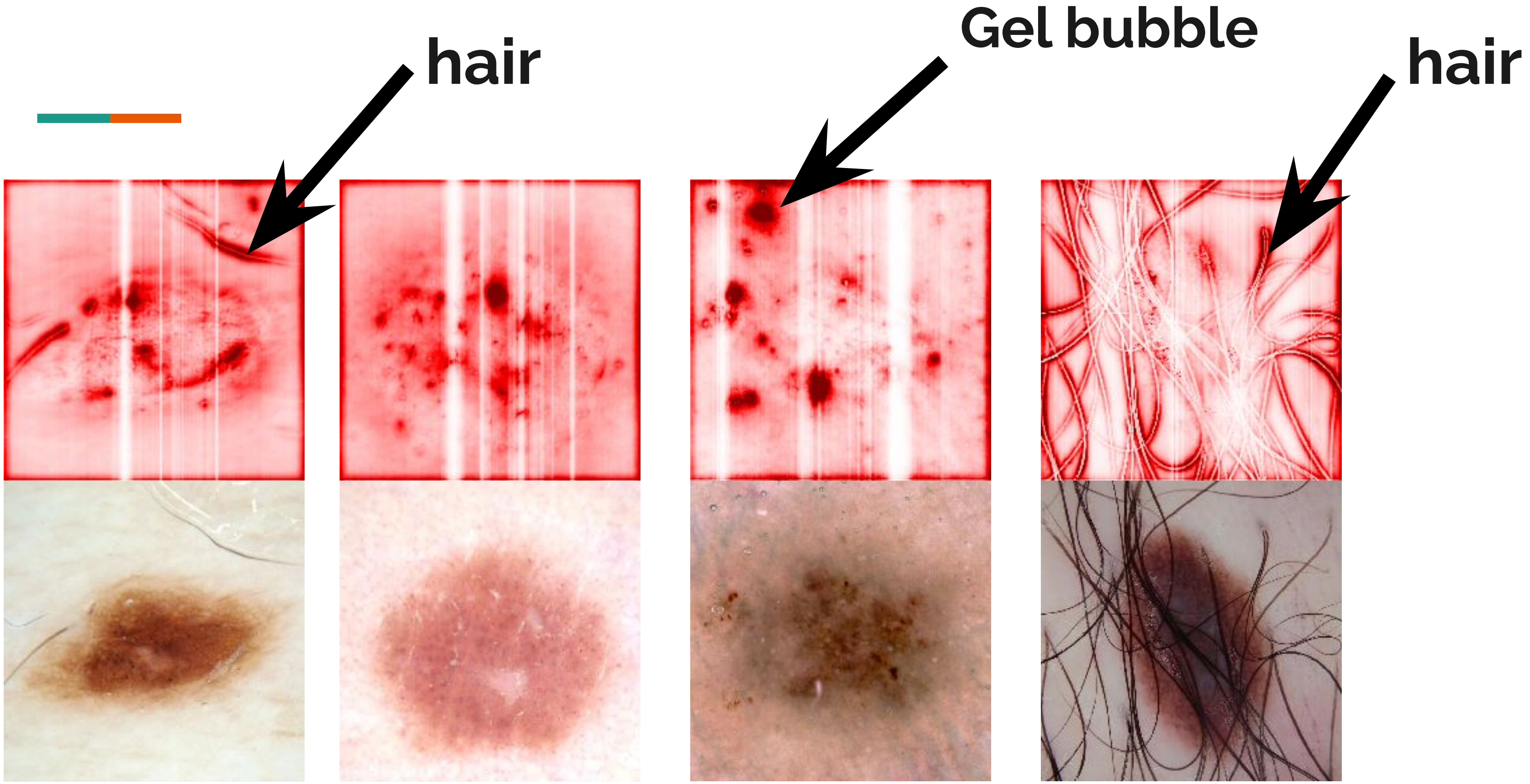


Are skin lesion datasets biased?

Preliminary results.

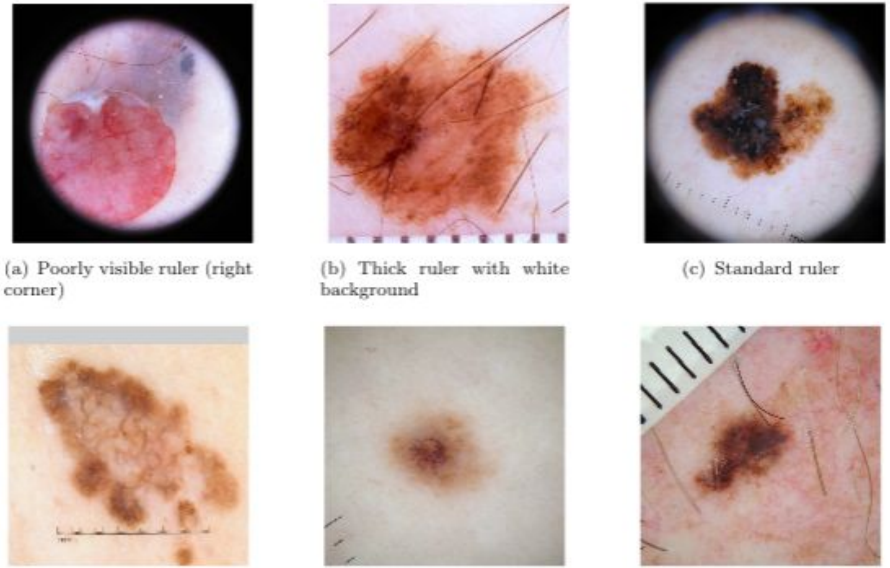
Layer-wise Relevance Propagation - LRP




hair
Gel bubble
hair

Skin lesion classification example

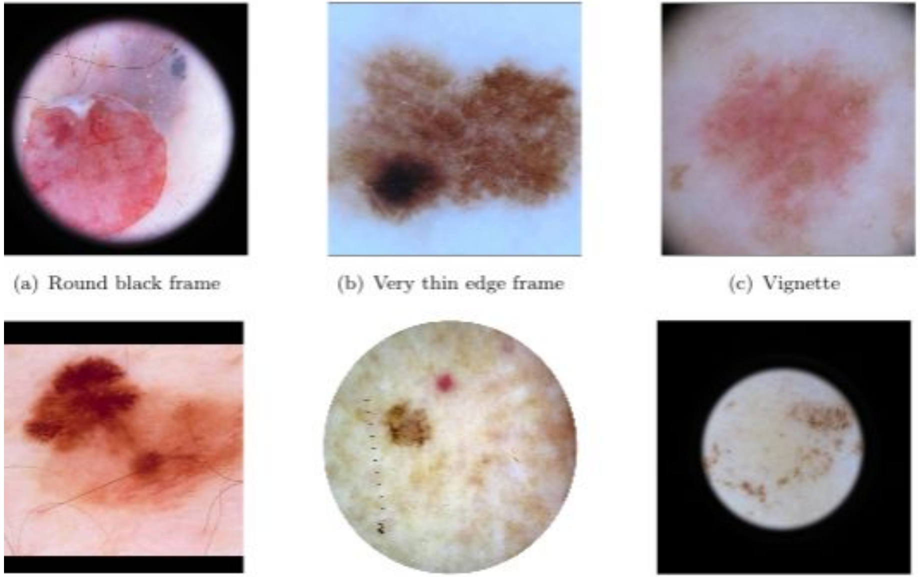
Images with ruler marks



(a) Poorly visible ruler (right corner)
(b) Thick ruler with white background
(c) Standard ruler
(d) Fully visible
(e) Partially visible
(f) Large with white back-groun

FIGURE 5.3: Example of images with ruler marks.

Images with frames



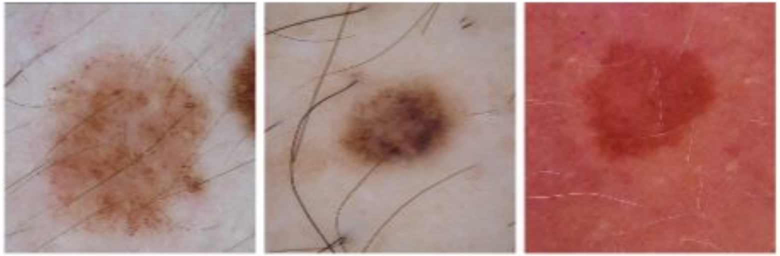
(a) Round black frame
(b) Very thin edge frame
(c) Vignette
) Rectangular black frame
(e) White round frame
(f) Large round frame

FIGURE 5.2: Example of images with frames.

Images with hair

(a) Dense hair

(b) Short hair

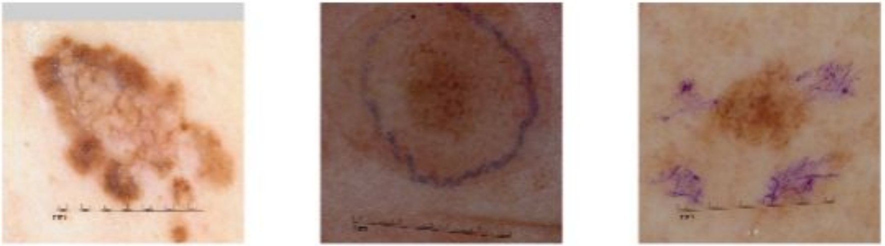


(c) Normal hair

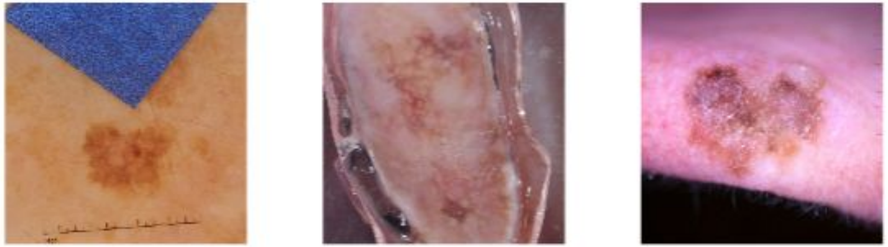
FIGURE 5.1: Example of images with hair.

Images with other artifacts

(a) Gel bubbles (b) Numbers or dates (c) Patches



(d) Dust (e) Ink circle (f) Ink stains



(g) Paper (h) Gel border (i) Visible background

FIGURE 5.4: Example of images with most common artifacts that were annotated as *other*.

Skin lesion classification example

FIGURE 5.5: Example clean images without any annotated artifacts.

Skin lesion classification - Statistics

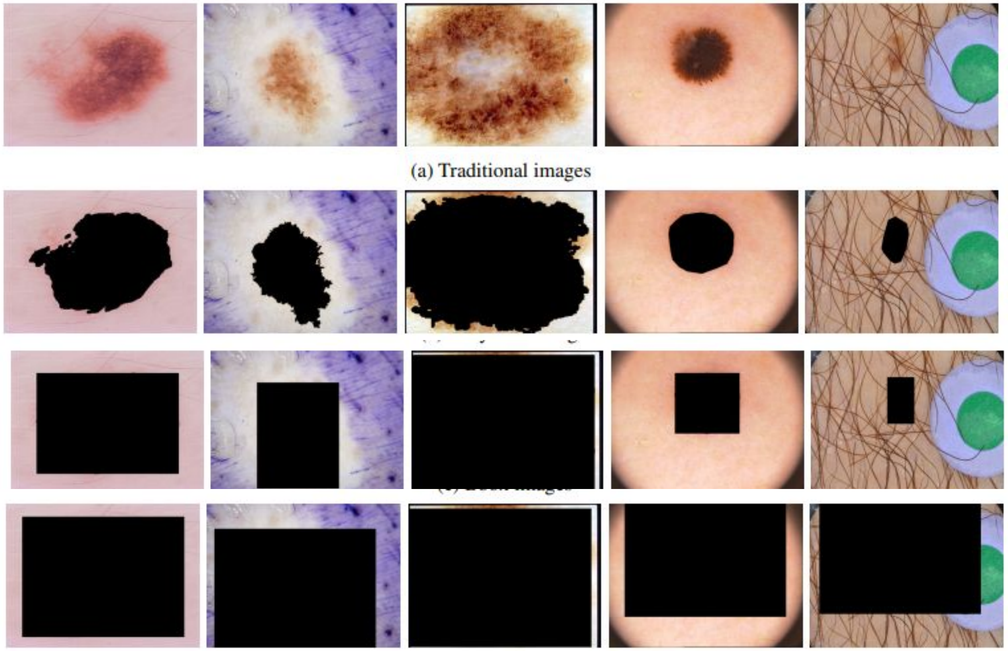
TABLE 5.1: Manually annotated artifacts in the skin lesion dataset ISIC 2019, [4–6] and ISIC 2020 [7].

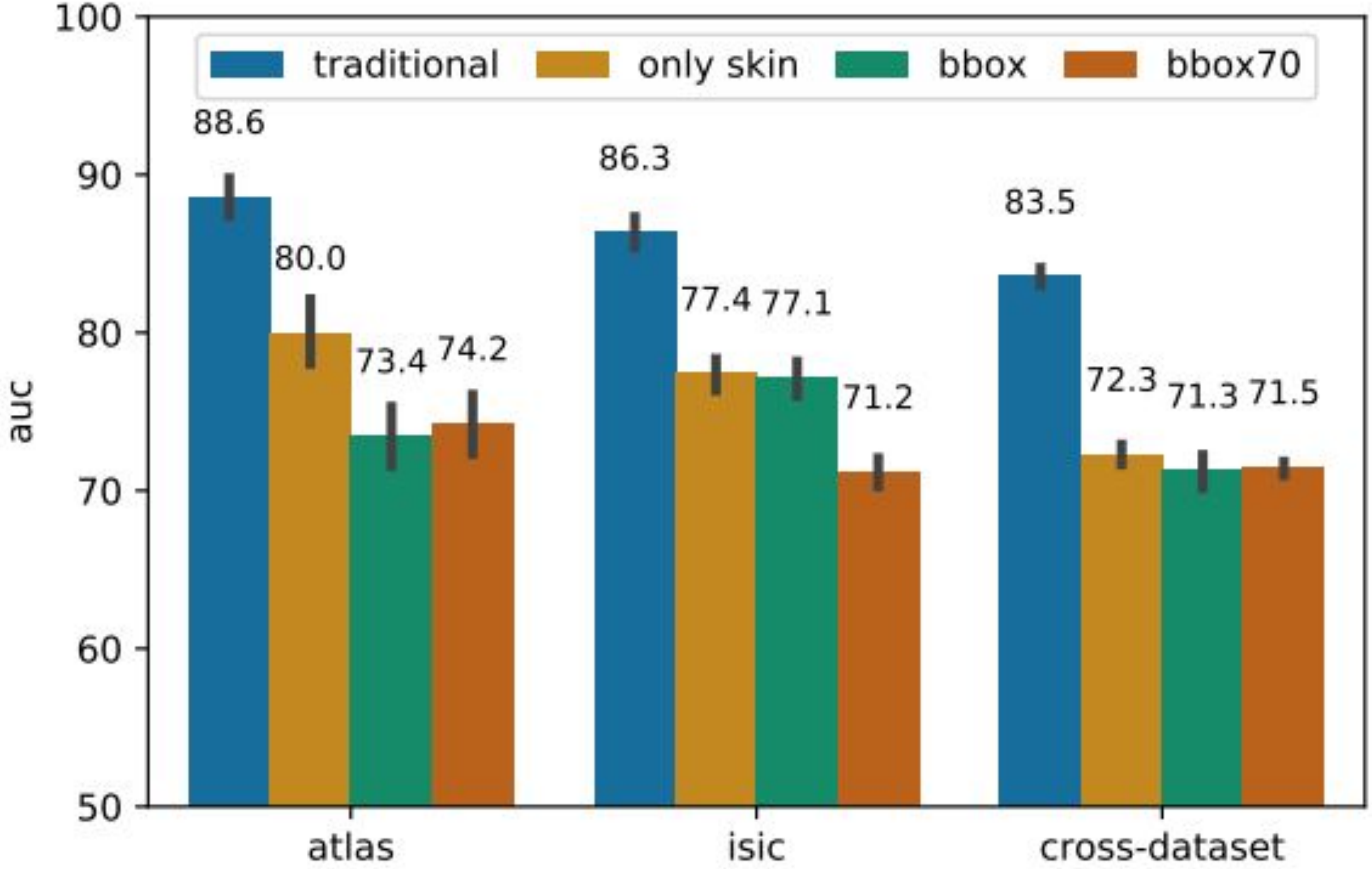
artifact	benign		malignant		$class_{ratio}$
	sum	ratio	sum	ratio	
frame	104	5.20%	521	26.05%	5.01
hair	958	47.88%	868	43.40%	0.91
dense	204	10.19%	99	4.95%	0.49
short	96	4.80%	103	5.15%	1.07
ruler	422	21.09%	586	29.30%	1.39
other	426	21.29%	818	40.90%	1.92
none	538	26.89%	268	13.40%	0.50
total	2001		2000		

How to detect bias without manually screening thousands of images?

Training the model to classify skin lesions but without any skin lesions

“(De)Constructing Bias on Skin Lesion Datasets”, 2019, CVPR





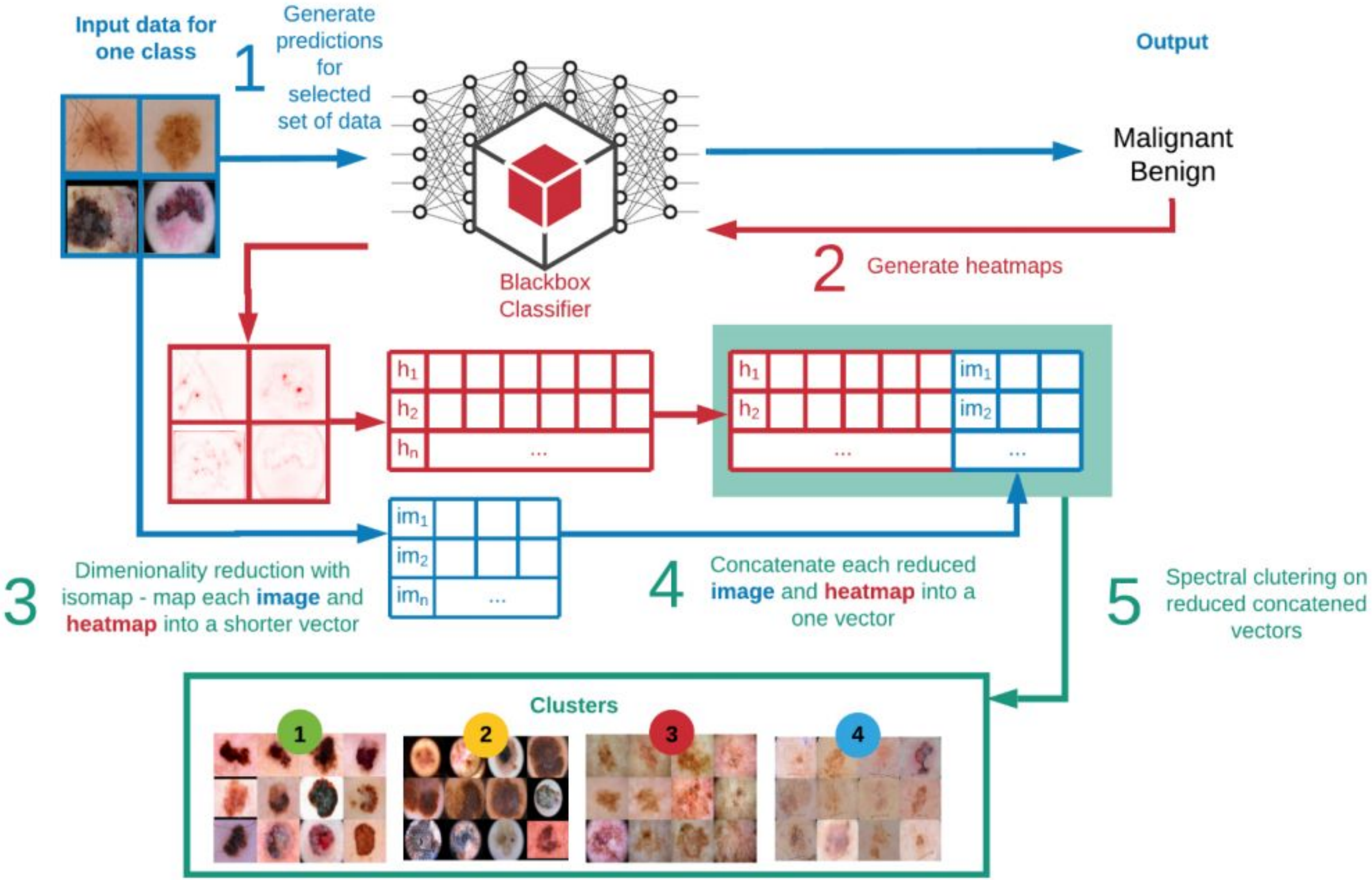
100
90
80
70
60
50
auc
traditional
only skin
bbox
bbox70
88.6
80.0
73.4
74.2
86.3
77.4
77.1
71.2
83.5
72.3
71.3
71.5
atlas
isic
cross-dataset

What is causing bias? How strong is it?

- **Detect bias:** GEBI - Global explanations for bias identification
- **Evaluate bias:** Counterfactual bias insertion

Global Explanations for Bias Identification

- **The idea:** generate local explanations for every instance, and cluster them to find patterns in prediction


Input data for one class
1
Generate predictions for selected set of data
Blackbox Classifier
Output
Malignant
Benign
2
Generate heatmaps
h1
h2
hn
...
im1
im2
imn
...
3
Dimenionality reduction with isomap - map each image and heatmap into a shorter vector
4
Concatenate each reduced image and heatmap into a one vector
5
Spectral clutering on reduced concatened vectors
Clusters
1
2
3
4

➔ Why do we need concatenate attribution maps with input instances?

Example visualization of occlusion-based attribution maps

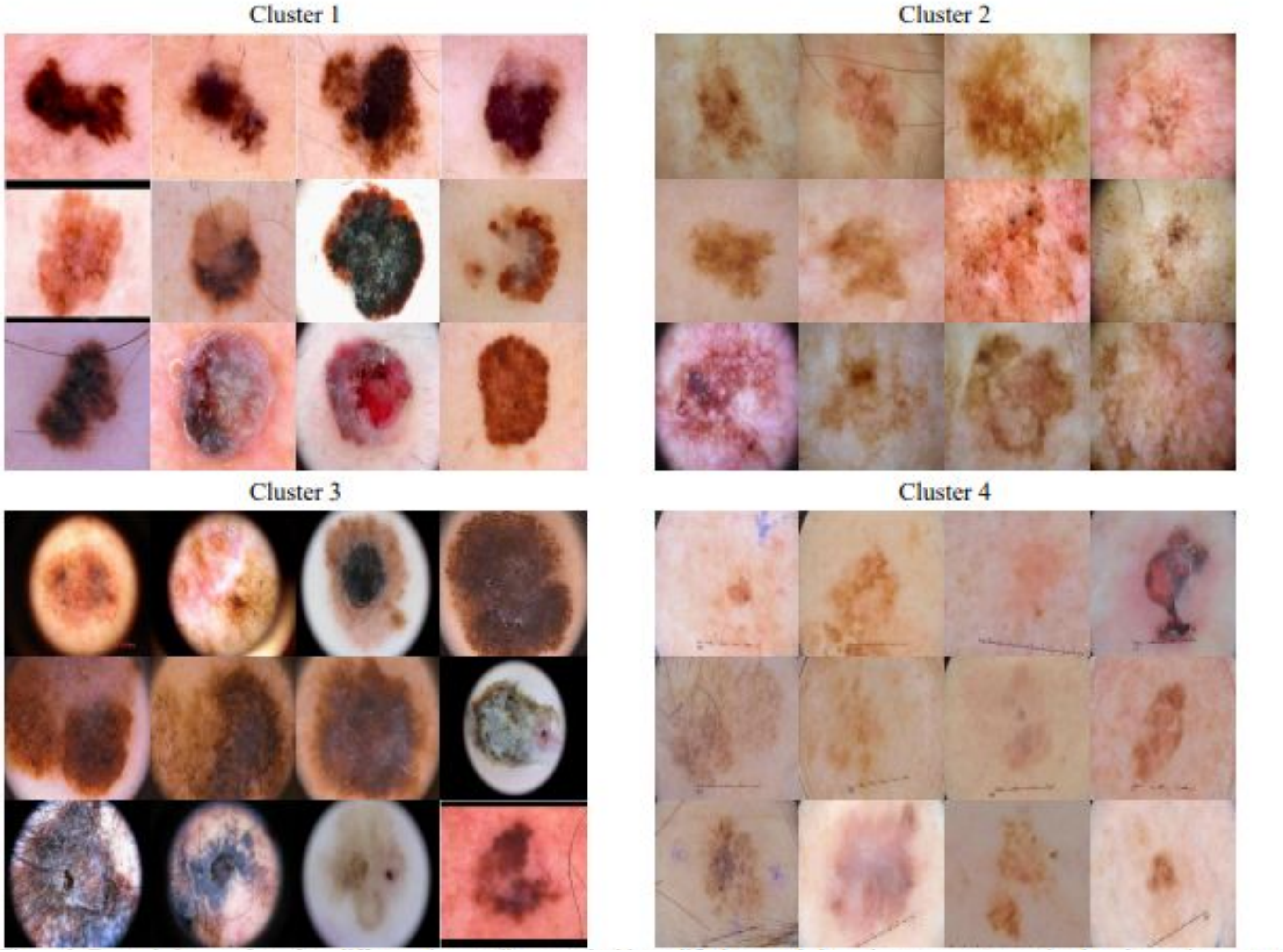
(a) Small skin lesion with smooth borders on the center of the image with strongly textured skin

(b) Large protruding skin lesion with well-defined borders

(c) Medium round skin lesion with irregular border with streaks and atypical dots

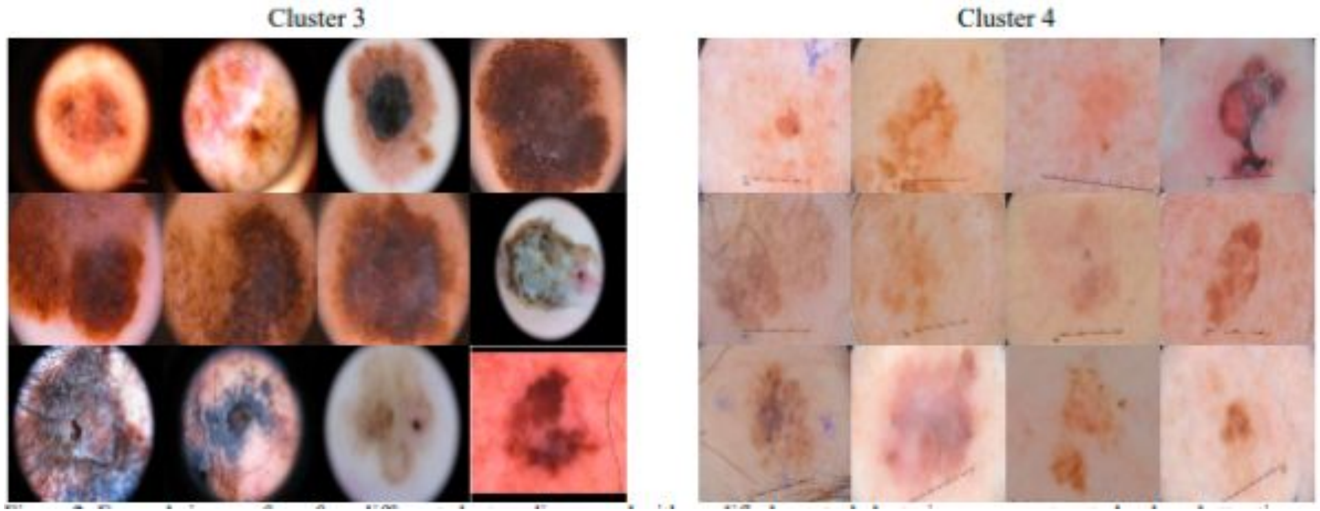
FIGURE 4.1: Example visualization of occlusion-based explanations. In the heatmap, a darker green color means stronger attribution. Visualized with captum [1].

Results



GEBI

➔ Black frames and ruler marks are possibly biasing factors



What’s next?

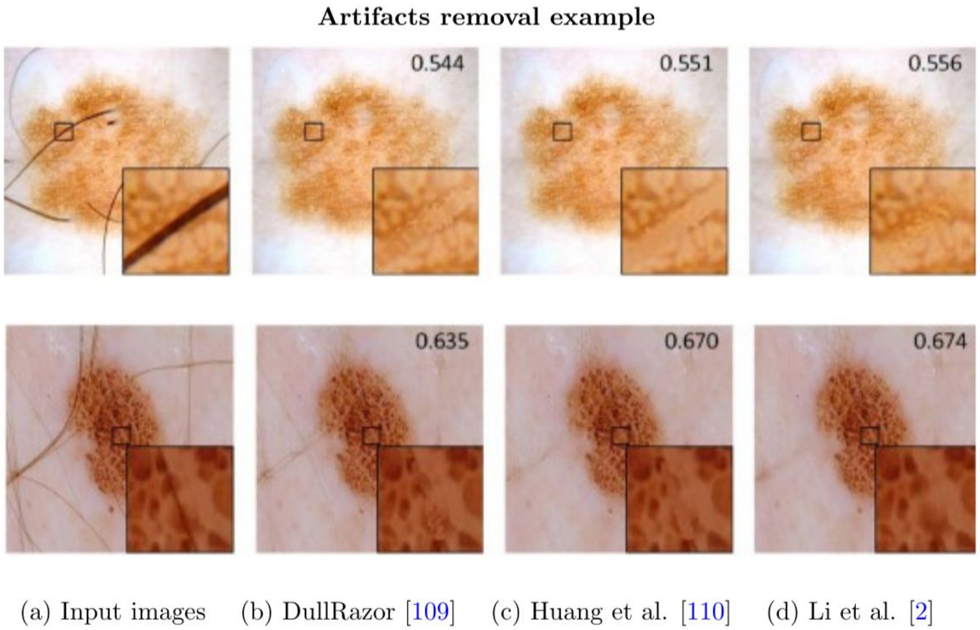
➔ With statistical analysis and Global Explanations we discovered potential sources of bias.

➔ **Next step:** evaluating how it affects the model

Counterfactual bias insertion

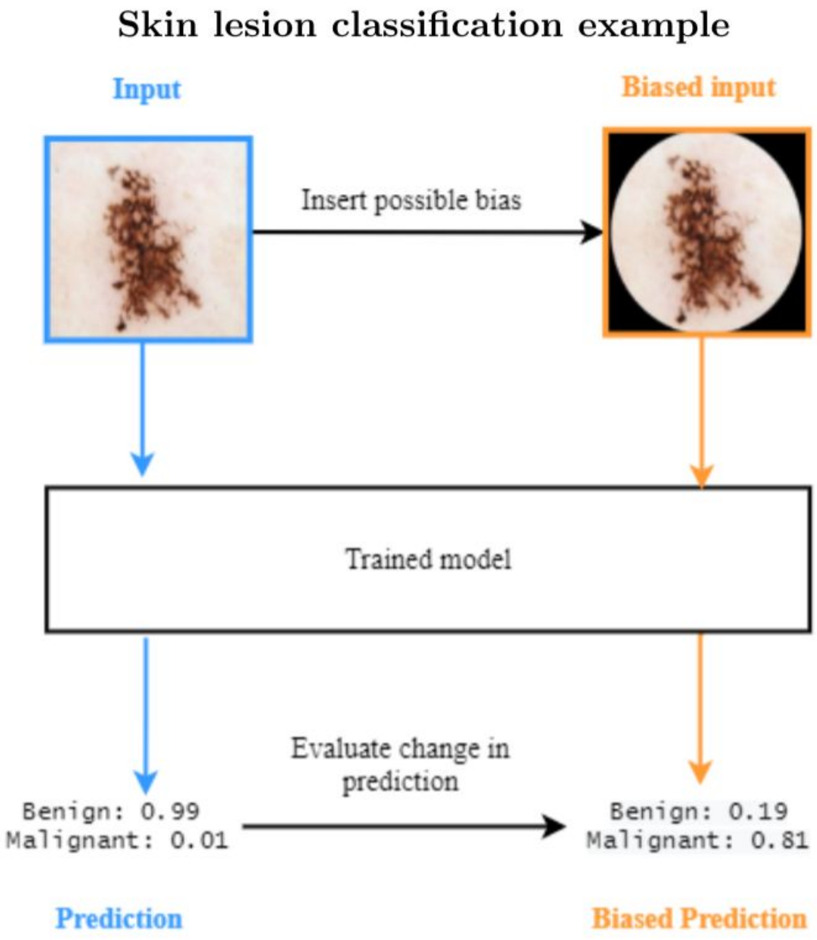
We could remove bias and check how prediction changed.

Removing artifacts from the images is difficult task. Moreover, image inpainting generates new artifacts instead



Counterfactual bias insertion

We will insert artifacts instead and see how the prediction changed



Counterfactual bias insertion

Early results

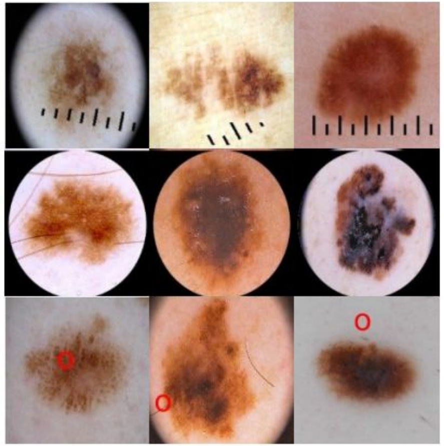


Figure 3: Modified examples by insertion of artificial bias: a) ruler markings, b) black frames, c) red circles

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**	**30.77**	**62.43**
	Ben	**32.04**	**63.66**
Red circle	Mal	2.27	15.51
	Ben	1.50	12.78

Mitigating bias

We detected the bias and want to mitigate it

The idea: during the training force the model to ignore biasing factors.

- Indirectly: Targeted Data Augmentations

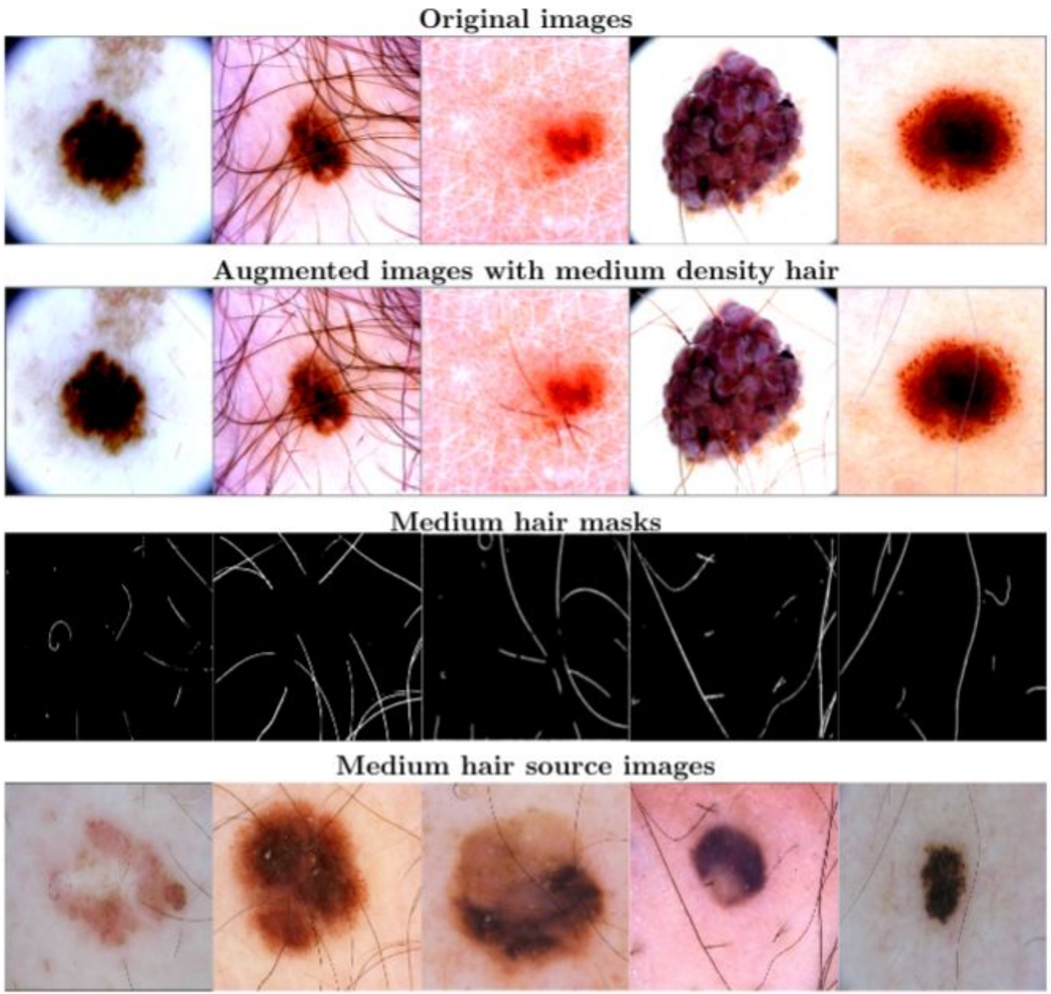
Targeted data augmentations

Theory: Insert bias randomly during the training with a given probability *p* to force the model to ignore it

In practice: make custom data augmentation e.g. with Albumentations library (custom torch transform). Design the method depending on data and bias.

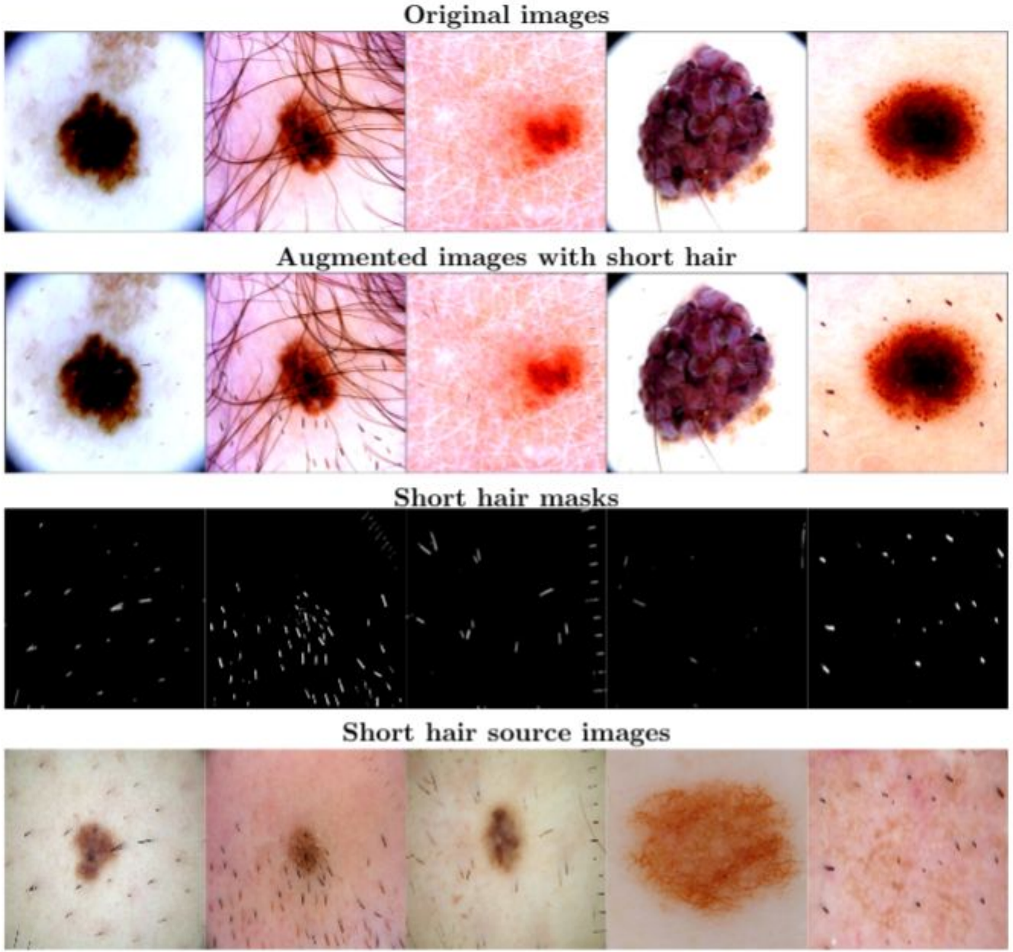
Hair augmentation

- Normal hair



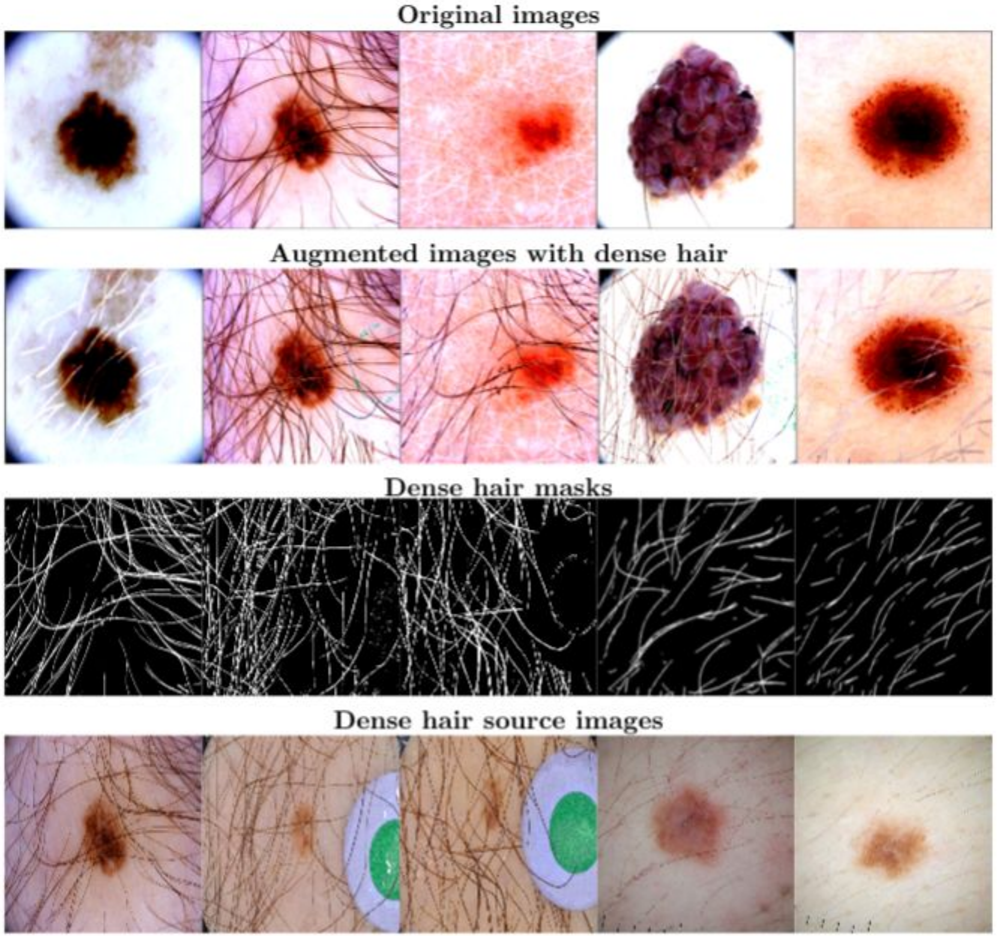
Hair augmentation

- Normal hair
- Short hair

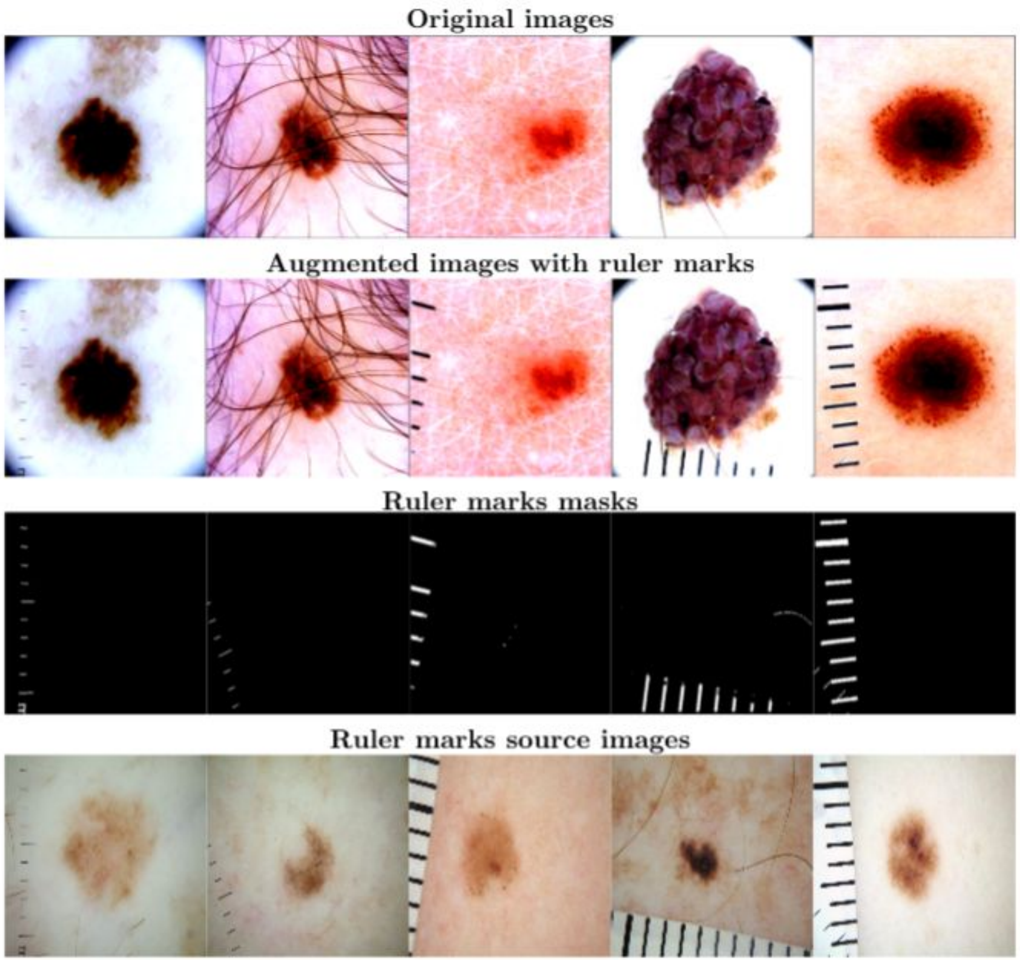


Hair augmentation

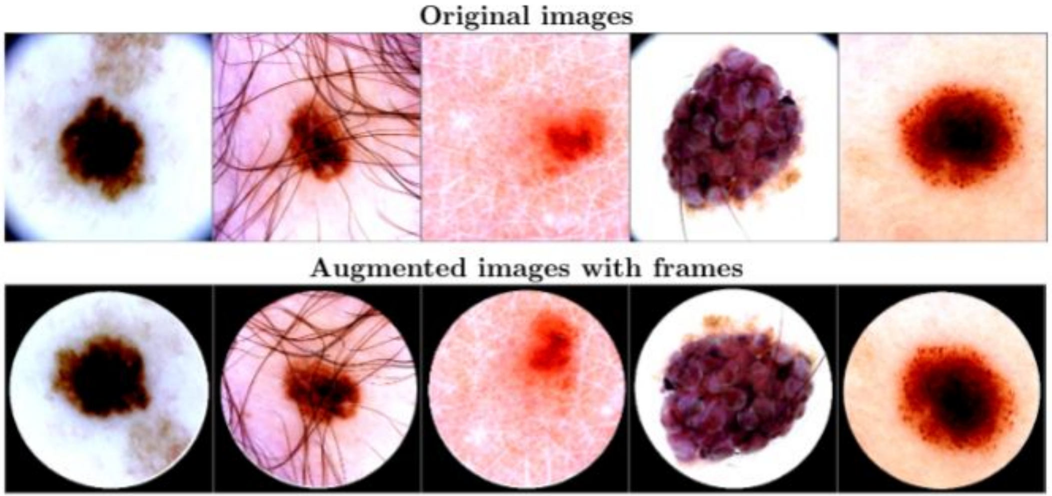
- Normal hair
- Short hair
- Dense hair



Ruler augmentation



Frame augmentation



Results - Performance evaluation (frame augmentation)

model	p	$f1_{org}$	$f1_{aug}$	$f1_{mean}$	$recall_{org}$	$recall_{aug}$	$precision_{org}$	$precision_{aug}$
efficientnet-b2	0	59.96%	52.99%	56.48%	46.45%	55.74%	84.58%	50.50%
	0.25	60.14%	59.12%	59.63%	46.17%	44.26%	86.22%	89.01%
	0.5	**64.24%**	**62.50%**	**63.37%**	**53.01%**	**50.55%**	**81.51%**	**81.86%**
	0.75	60.14%	59.49%	59.82%	46.17%	44.54%	86.22%	89.56%
	1	58.42%	58.02%	58.22%	46.45%	46.45%	78.70%	77.27%
efficientnet-b3	0	66.15%	54.69%	60.42%	58.47%	58.20%	76.16%	51.57%
	0.25	**64.93%**	**61.73%**	**63.33%**	**51.09%**	**46.72%**	**89.05%**	**90.96%**
	0.5	62.13%	58.59%	60.36%	48.63%	44.26%	85.99%	86.63%
	0.75	62.50%	61.54%	62.02%	49.18%	46.99%	85.71%	89.12%
	1	62.44%	61.15%	61.80%	49.73%	49.45%	83.87%	80.09%
efficientnet-b4	0	68.35%	61.60%	64.98%	55.46%	58.74%	89.04%	64.76%
	0.25	64.93%	63.76%	64.35%	54.37%	51.91%	80.57%	82.61%
	0.5	66.01%	64.40%	65.21%	54.64%	51.64%	83.33%	85.52%
	0.75	66.99%	65.01%	66.00%	57.10%	53.55%	81.01%	82.70%
	1	**67.55%**	**65.65%**	**66.60%**	**62.57%**	**58.74%**	**73.40%**	**74.39%**

Results - Bias evaluation (frame augmentation)

model	p	mean	median	switched all	to ben	to mal
efficientnet-b2	0	5.43%	1.25%	$\boxed{241}$	19	222
	0.25	1.07%	0.02%	54	34	20
	0.5	1.13%	0.02%	50	31	19
	0.75	1.07%	0.02%	$\boxed{48}$	31	17
	1	1.49%	0.06%	72	34	38
efficientnet-b3	0	4.66%	0.73%	$\boxed{208}$	38	170
	0.25	0.96%	0.00%	44	33	11
	0.5	0.92%	0.01%	48	34	14
	0.75	1.05%	0.02%	$\boxed{39}$	28	11
	1	1.44%	0.04%	59	25	34
efficientnet-b4	0	4.16%	0.62%	$\boxed{138}$	17	121
	0.25	1.24%	0.10%	51	34	17
	0.5	0.86%	0.03%	$\boxed{47}$	33	14
	0.75	1.34%	0.07%	53	37	16
	1	1.63%	0.05%	77	50	27

Results - Performance evaluation (hair augmentation)

model	type	p	$f1_{org}$	$f1_{aug}$	$f1_{mean}$	$precision_{org}$	$precision_{aug}$
B2	short	0	59.96%	54.04%	57.00%	84.58%	82.58%
		0.25	61.14%	57.96%	59.55%	83.10%	83.94%
		0.5	62.02%	59.43%	60.73%	85.58%	85.20%
		0.75	**63.59%**	**61.40%**	**62.50%**	**84.93%**	**85.78%**
		1	58.09%	55.47%	56.78%	88.76%	89.63%
	medium	0	59.96%	51.42%	55.69%	84.58%	83.44%
		0.25	61.14%	58.20%	59.67%	83.10%	82.09%
		0.5	62.02%	58.82%	60.42%	85.58%	80.19%
		0.75	**63.59%**	**64.38%**	**63.99%**	**84.93%**	**86.24%**
		1	58.09%	56.26%	57.18%	88.76%	83.78%
	dense	0	59.96%	35.82%	47.89%	84.58%	81.55%
		0.25	61.14%	50.00%	55.57%	83.10%	81.48%
		0.5	62.02%	53.43%	57.73%	85.58%	78.72%
		0.75	**63.59%**	**52.47%**	**58.03%**	**84.93%**	**86.25%**
		1	58.09%	48.92%	53.51%	88.76%	86.21%

Results - Bias evaluation (hair augmentation)

model	type	p	mean	median	all	to ben	to mal
efficientnet-b2	short	**0**	**0.94%**	**0.02%**	**45**	**34**	**11**
		0.25	0.91%	0.02%	42	31	11
		0.5	0.97%	0.03%	40	26	14
		0.75	0.80%	0.03%	43	29	14
		1	0.72%	0.03%	26	20	6
	medium	0	1.22%	0.02%	62	50	12
		0.25	**1.14%**	**0.03%**	**64**	**38**	**26**
		0.5	1.12%	0.04%	52	24	28
		0.75	0.97%	0.04%	37	19	18
		1	0.96%	0.05%	35	14	21
	dense	**0**	**1.83%**	**0.06%**	**136**	**117**	**19**
		0.25	2.32%	0.16%	109	80	29
		0.5	2.44%	0.22%	102	61	41
		0.75	2.41%	0.26%	107	83	24
		1	1.83%	0.16%	71	52	19

Conclusion

- Helps to mitigate unwanted biases
- Easy to use and implement, easy to merge with existing ML pipeline (just add new data augmentation method)
- Difficult preliminary step: Bias detection
- Can be applied to only the part of biases

Future works

- Use all targeted data augmentations at once
- Test on some NLP cases
- Need better bias identification strategies

Thank you. Questions?

Agnieszka Mikołajczyk
agnieszka.mikolajczyk@pg.edu.pl
Gdańsk University of Technology

Personal website: amikolajczyk.netlify.com

Github: github.com/AgaMiko

Linkedin: linkedin.com/in/agnieszkamikolajczyk

Twitter: @AgnMikolajczyk