Wykrywanie i zmniejszanie wpływu tendencyjności danych za pomocą objaśnialnej sztucznej inteligencji

Agnieszka Mikołajczyk

Plan

Nasz zespół

Zakres badań - przeszłość i teraźniejszość

Preludium

Podsumowanie

Nasz zespół

Michał Grochowski
Kierownik zespołu

Arkadiusz Kwasigroich
Doktorant, główny wykonawca Diamentowego Grantu, pracuje w dziale R&D w firmie Brainscan

Maria Ogryczak
Doktorantka (wkrótce), zainteresowania zastosowaniami głębokiego uczenia w medycynie

Agnieszka Mikolajczyk
Doktorantka, zainteresowana XAI, główny wykonawca Preludium, pracuje w dziale R&D w firmie Voicelab (projekt NCBR)

Klasyfikacja znamion skórnych

Wprowadzenie

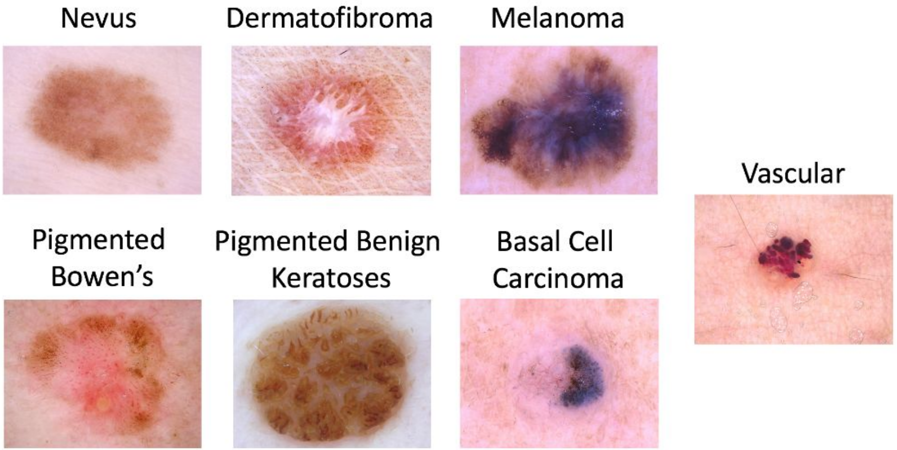
- Ważny i znany problem
- Rozpopularyzowany przez ISIC Archive i coroczne wyzwania
- Co roku więcej danych
- W tym roku na CVPR Workshop

Klasyfikacja znamion skórnych

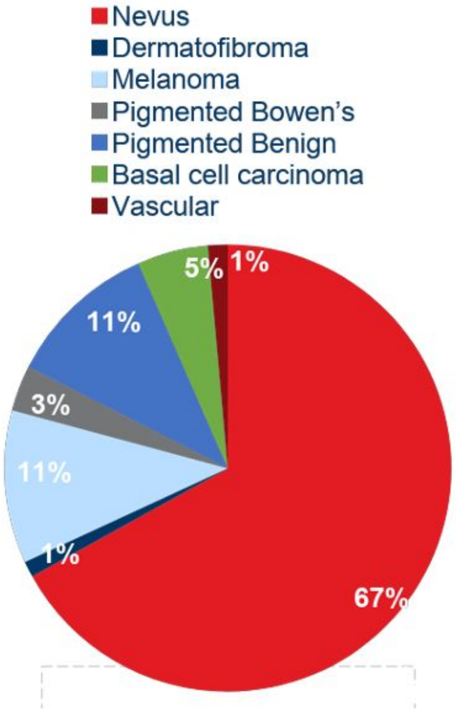
- Najczęstsze zagadnienia:
  - klasyfikacja znamion na łagodne i złośliwe,
  - klasyfikacja wg. typu znamienia,
  - segmentacja,
  - wykrywanie struktur charakterystycznych dla nowotworów

Klasyfikacja znamion skórnych

- Niezbilansowane bazy danych

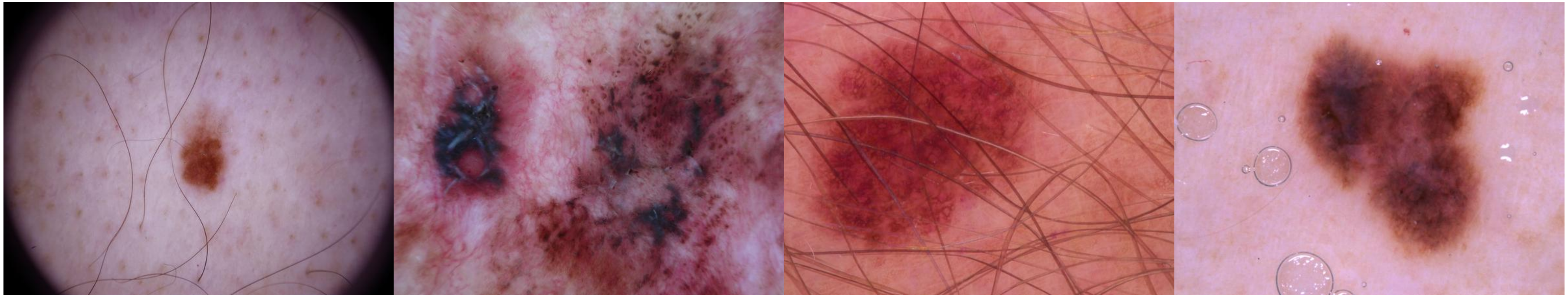


ISIC Challenge 2018



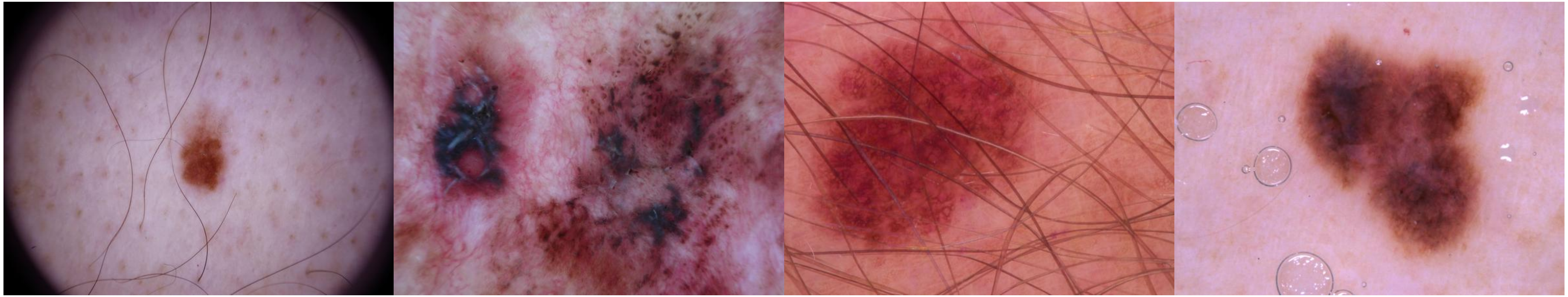
Klasyfikacja znamion skórnych

- Dane różnej jakości, brak standardów fotografii dermatoskopowej



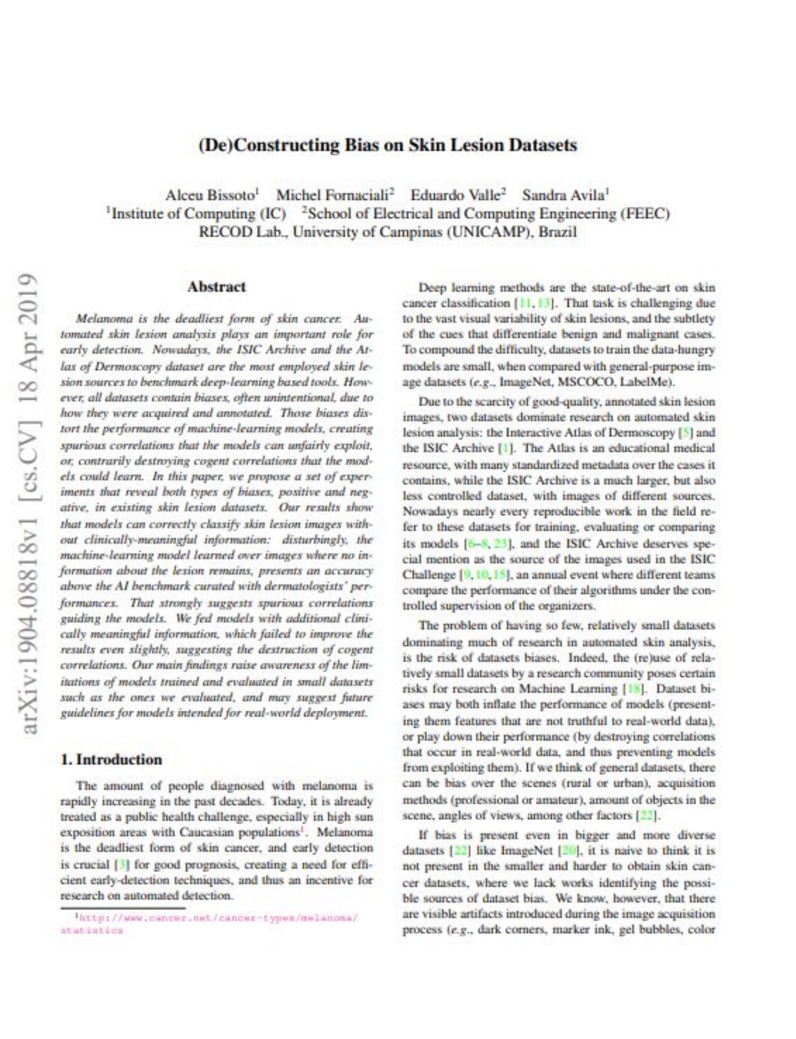
Klasyfikacja znamion skórnych

- Częste występowanie artefaktów na zdjęciach:
  - krople żelu
  - czarne ramki
  - linijki
  - odbarwienia od flamastra



- Tendencyjność (ang. bias) w danych

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



(De)Constructing Bias on Skin Lesion Datasets

Alceu Bissoto[1] Michel Fornaciali[2] Eduardo Valle[2] Sandra Avila[1]
[1]Institute of Computing (IC) [2]School of Electrical and Computing Engineering (FEEC)
RECOD Lab., University of Campinas (UNICAMP), Brazil

Abstract

Melanoma is the deadliest form of skin cancer. Automated skin lesion analysis plays an important role for early detection. Nowadays, the ISIC Archive and the Atlas of Dermoscopy dataset are the most employed skin lesion sources to benchmark deep-learning based tools. However, all datasets contain biases, often unintentional, due to how they were acquired and annotated. Those biases distort the performance of machine-learning models, creating spurious correlations that the models can unfairly exploit, or, contrarily destroying cogent correlations that the models could learn. In this paper, we propose a set of experiments that reveal both types of biases, positive and negative, in existing skin lesion datasets. Our results show that models can correctly classify skin lesion images without clinically-meaningful information: disturbingly, the machine-learning model learned over images where no information about the lesion remains, presents an accuracy above the AI benchmark curated with dermatologists' performances. That strongly suggests spurious correlations guiding the models. We fed models with additional clinically meaningful information, which failed to improve the results even slightly, suggesting the destruction of cogent correlations. Our main findings raise awareness of the limitations of models trained and evaluated in small datasets such as the ones we evaluated, and may suggest future guidelines for models intended for real-world deployment.

1. Introduction

The amount of people diagnosed with melanoma is rapidly increasing in the past decades. Today, it is already treated as a public health challenge, especially in high sun exposition areas with Caucasian populations[1]. Melanoma is the deadliest form of skin cancer, and early detection is crucial [3] for good prognosis, creating a need for efficient early-detection techniques, and thus an incentive for research on automated detection.

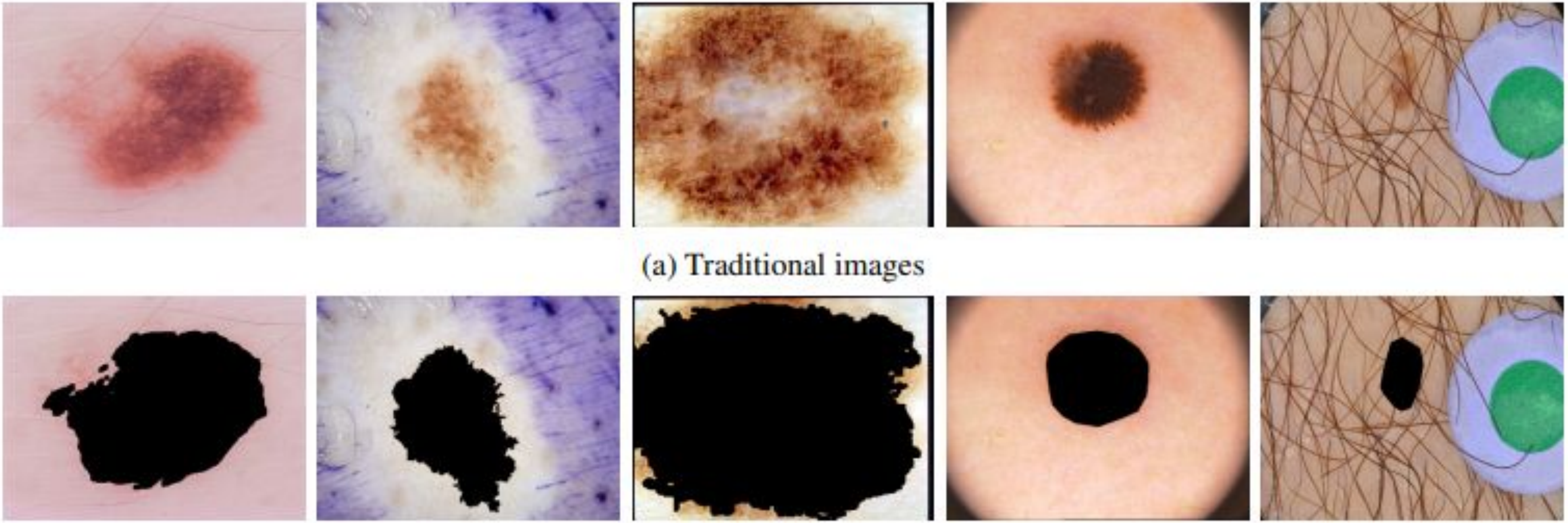
Deep learning methods are the state-of-the-art on skin cancer classification [11, 13]. That task is challenging due to the vast visual variability of skin lesions, and the subtlety of the cues that differentiate benign and malignant cases. To compound the difficulty, datasets to train the data-hungry models are small, when compared with general-purpose image datasets (*e.g.*, ImageNet, MSCOCO, LabelMe).

Due to the scarcity of good-quality, annotated skin lesion images, two datasets dominate research on automated skin lesion analysis: the Interactive Atlas of Dermoscopy [5] and the ISIC Archive [1]. The Atlas is an educational medical resource, with many standardized metadata over the cases it contains, while the ISIC Archive is a much larger, but also less controlled dataset, with images of different sources. Nowadays nearly every reproducible work in the field refer to these datasets for training, evaluating or comparing its models [6–8, 23], and the ISIC Archive deserves special mention as the source of the images used in the ISIC Challenge [9, 10, 15], an annual event where different teams compare the performance of their algorithms under the controlled supervision of the organizers.

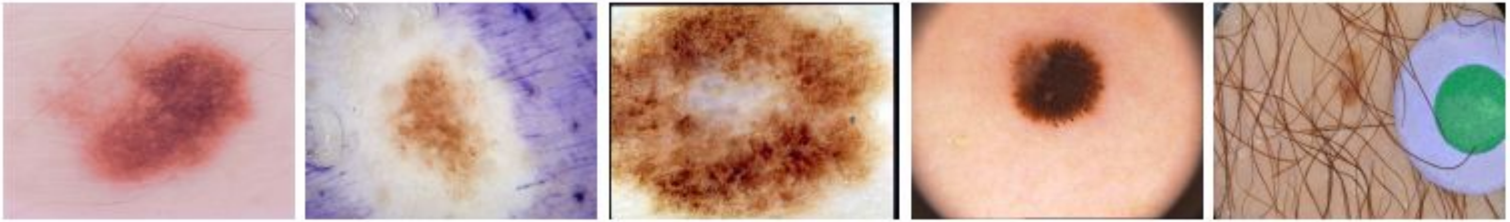
The problem of having so few, relatively small datasets dominating much of research in automated skin analysis, is the risk of datasets biases. Indeed, the (re)use of relatively small datasets by a research community poses certain risks for research on Machine Learning [18]. Dataset biases may both inflate the performance of models (presenting them features that are not truthful to real-world data), or play down their performance (by destroying correlations that occur in real-world data, and thus preventing models from exploiting them). If we think of general datasets, there can be bias over the scenes (rural or urban), acquisition methods (professional or amateur), amount of objects in the scene, angles of views, among other factors [22].

If bias is present even in bigger and more diverse datasets [22] like ImageNet [20], it is naive to think it is not present in the smaller and harder to obtain skin cancer datasets, where we lack works identifying the possible sources of dataset bias. We know, however, that there are visible artifacts introduced during the image acquisition process (*e.g.*, dark corners, marker ink, gel bubbles, color

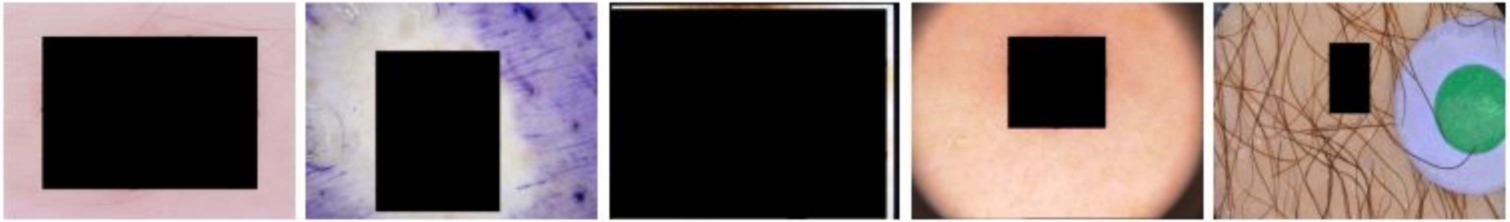
[1] http://www.cancer.net/cancer-types/melanoma/statistics



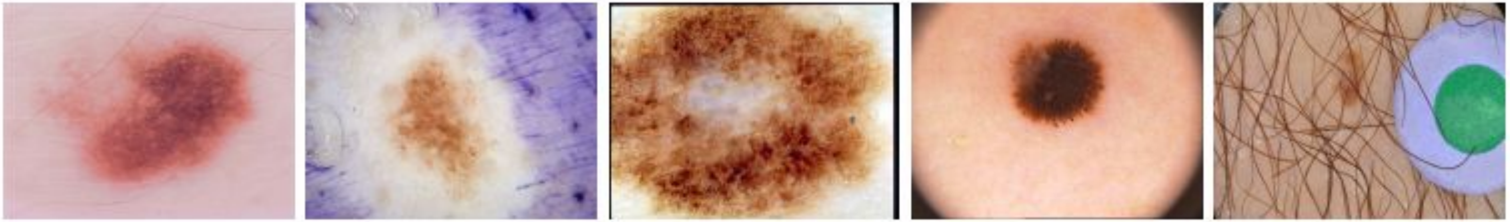
(a) Traditional images



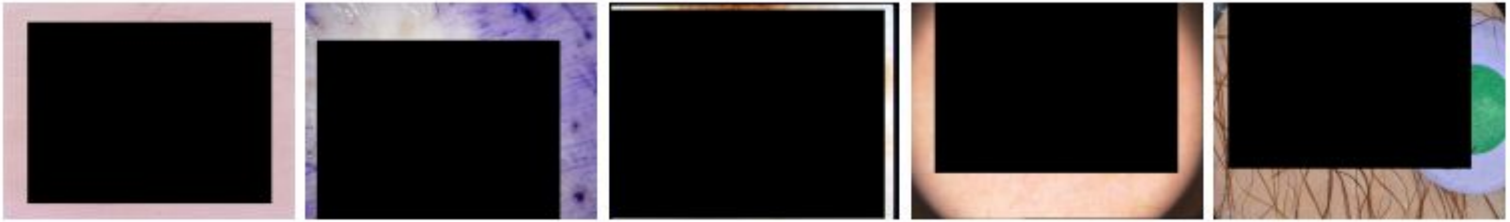
(a) Traditional images



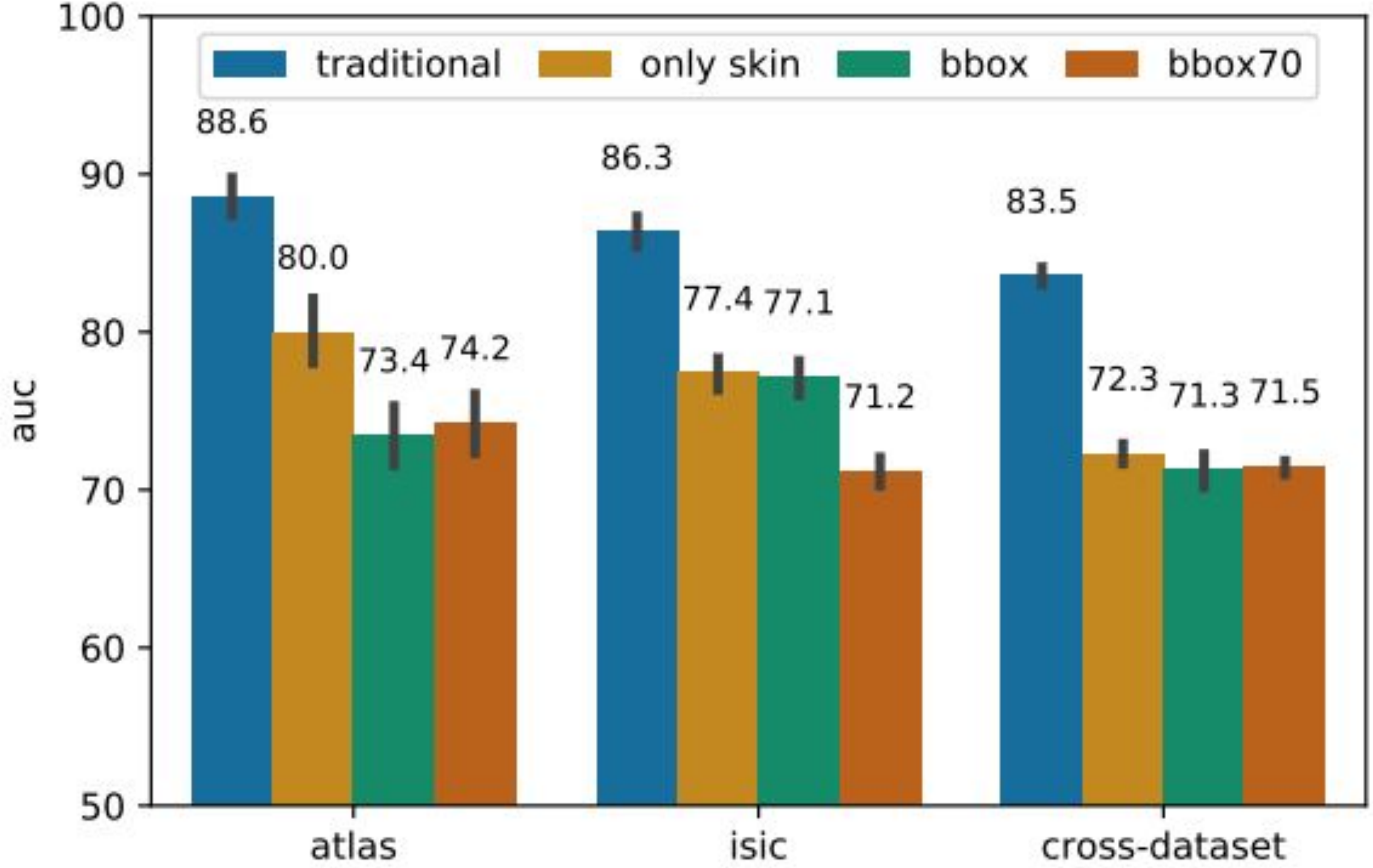
(c) Bbox images



(a) Traditional images



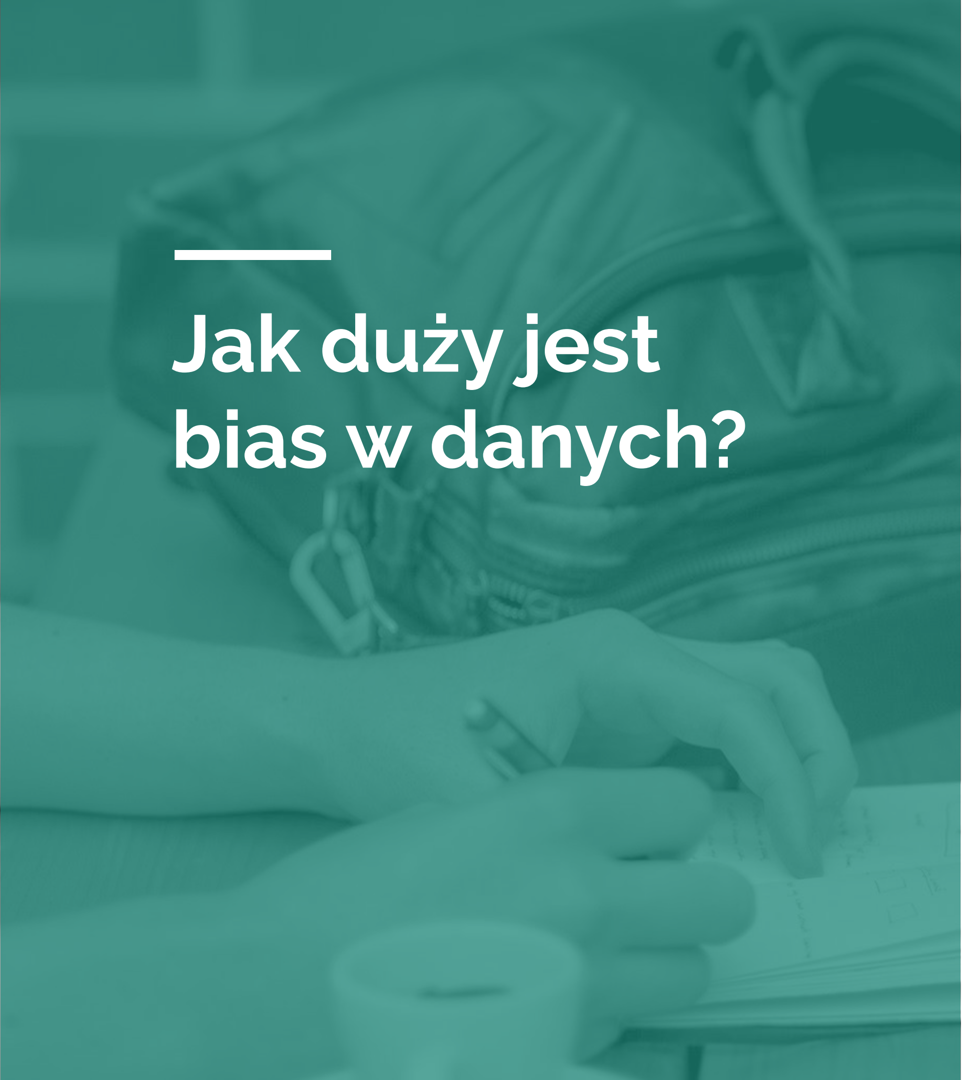
(d) Bbox70 images



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

Cytat z abstraktu

“Our results show that models can correctly classify skin lesion images without clinically-meaningful information: disturbingly, the machine-learning model learned over images where no information about the lesion remains, presents an accuracy above the AI benchmark curated with dermatologists’ performances. That strongly suggests spurious correlations guiding the models. We fed models with additional clinically meaningful information, which failed to improve the results even slightly, suggesting the destruction of cogent correlations.”

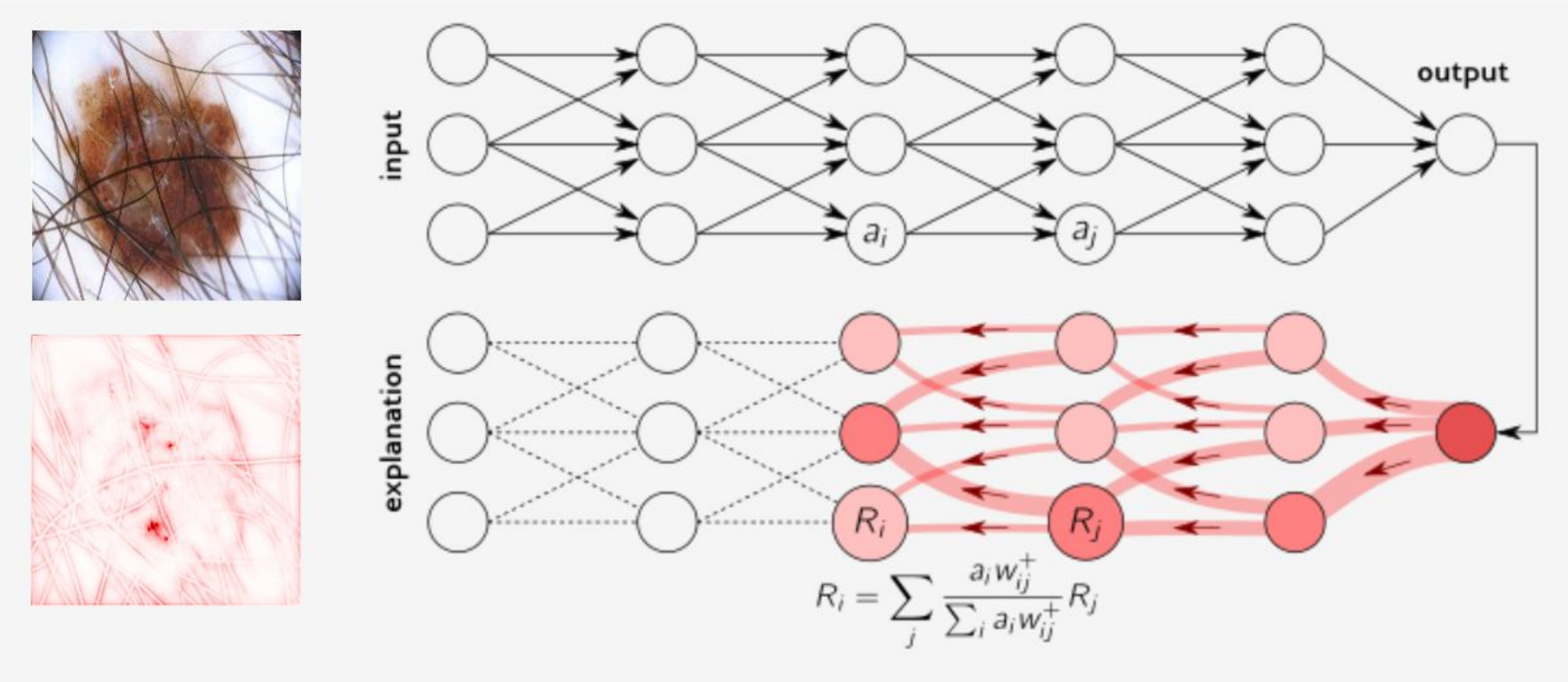


Jak duży jest bias w danych?

Czy możemy ufać naszym modelom?

Przeprowadziliśmy własne eksperymenty

Layer-wise Relevance Propagation - LRP



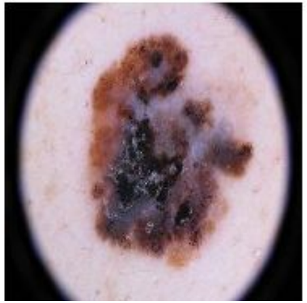
$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

Layer-wise Relevance Propagation - LRP

1 Prepare trained model and instance which you want to explain

Trained model: DenseNet 121

Input data



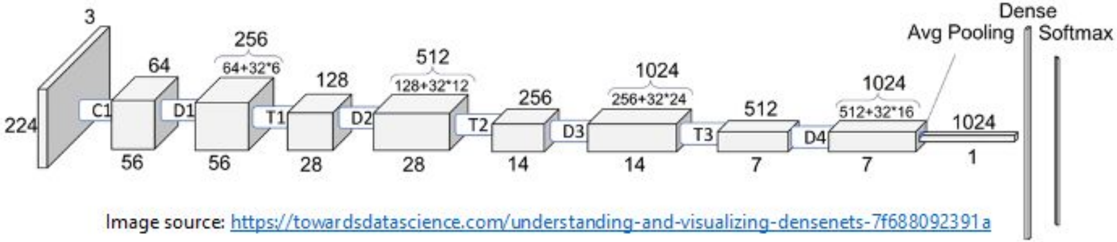


Image source: https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a

Prediction

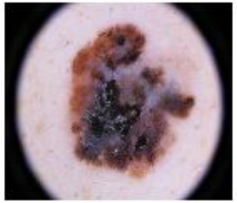
Malignant

?

Layer-wise Relevance Propagation - LRP

2 Calculate predictions for one instance and save neuron's activations

Input data



Trained model: DenseNet 121

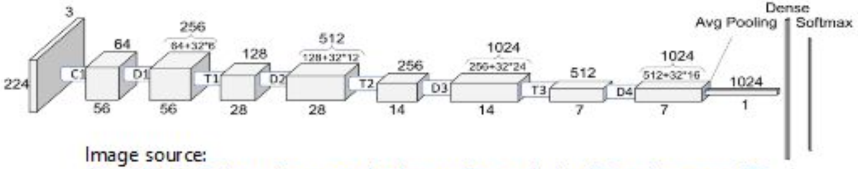


Image source:
https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a

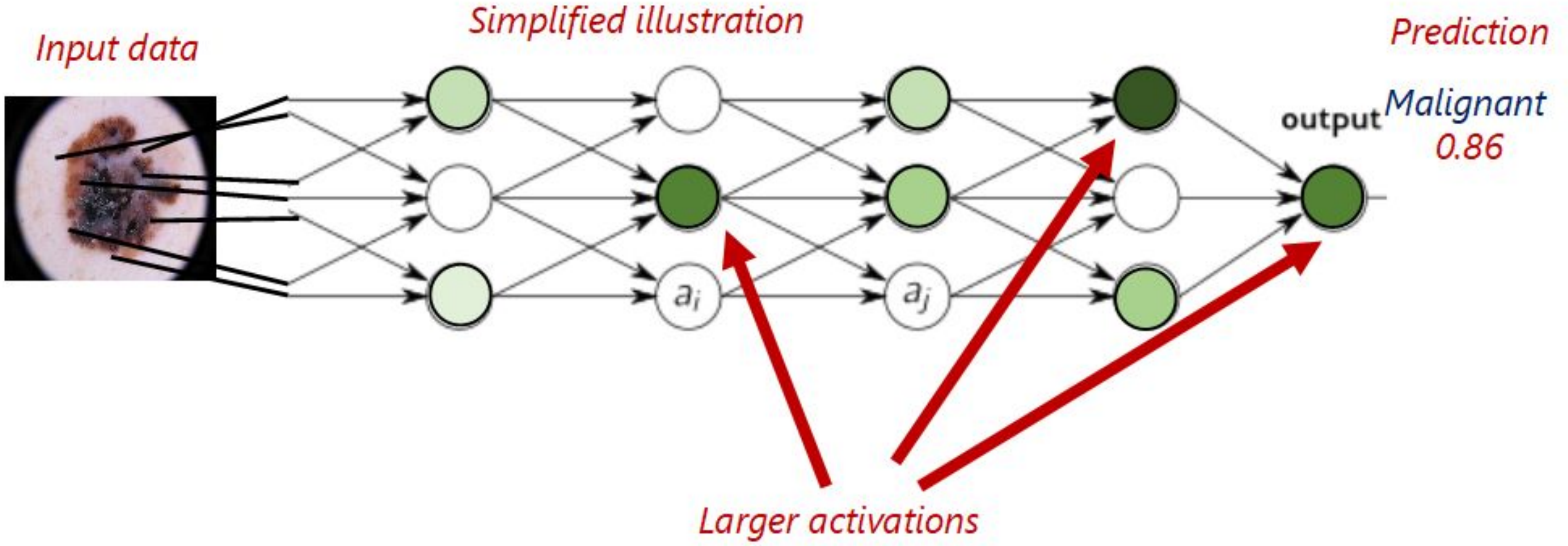
Prediction

Malignant
0.86

Activations will be used to calculate the relevance in the next step

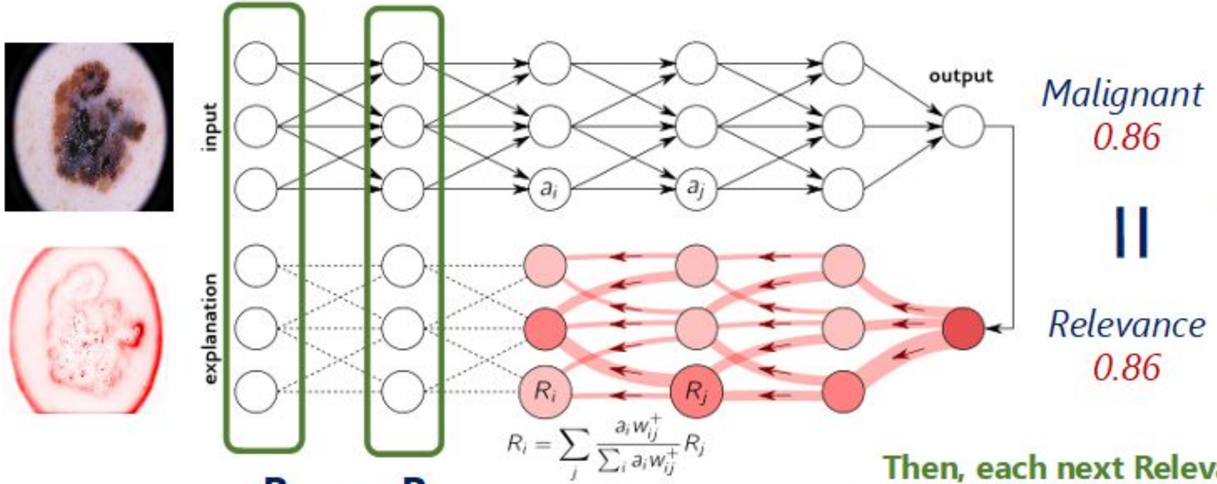
Layer-wise Relevance Propagation - LRP

2 Calculate predictions for one instance and save neuron's activations



Layer-wise Relevance Propagation - LRP

3 Backpropagate Relevance through the network



$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

Intialization: Output becomes first Relevance

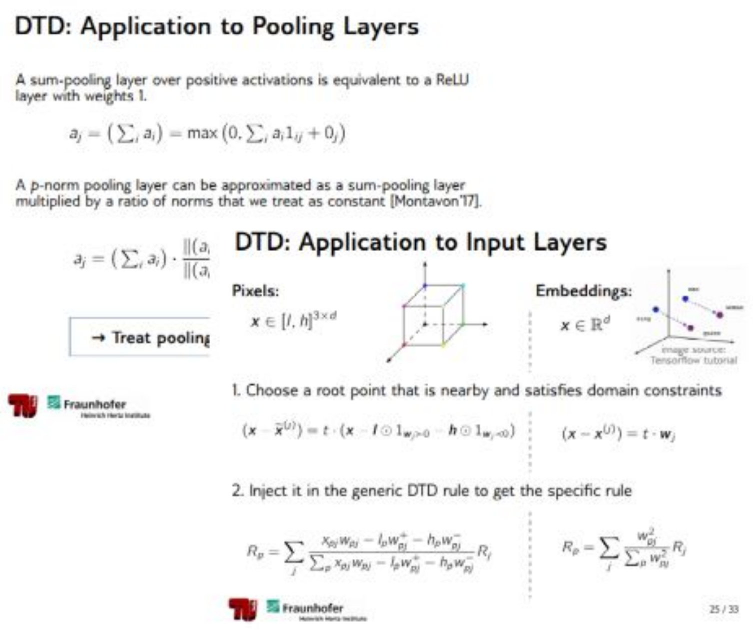
$\mathbf{R_0} = \mathbf{R_1}$

Total sum of relevances in each layer is equal

Then, each next Relevance is calculated based on: previous Relevance, weights, and activations

Layer-wise Relevance Propagation - LRP

- Each type of layer have its own Rules of how to backpropagate: Check the original paper!



DTD: Application to Pooling Layers

A sum-pooling layer over positive activations is equivalent to a ReLU layer with weights 1.

$$a_j = \left(\sum_i a_i\right) = \max\left(0, \sum_i a_i 1_{ij} + 0_j\right)$$

A p-norm pooling layer can be approximated as a sum-pooling layer multiplied by a ratio of norms that we treat as constant [Montavon'17].

$$a_j = \left(\sum_i a_i\right) \cdot \frac{\|(a_i}{\|(a_i}$$

→ Treat pooling

DTD: Application to Input Layers

Pixels:

$x \in [l, h]^{3 \times d}$

Embeddings:

$x \in \mathbb{R}^d$

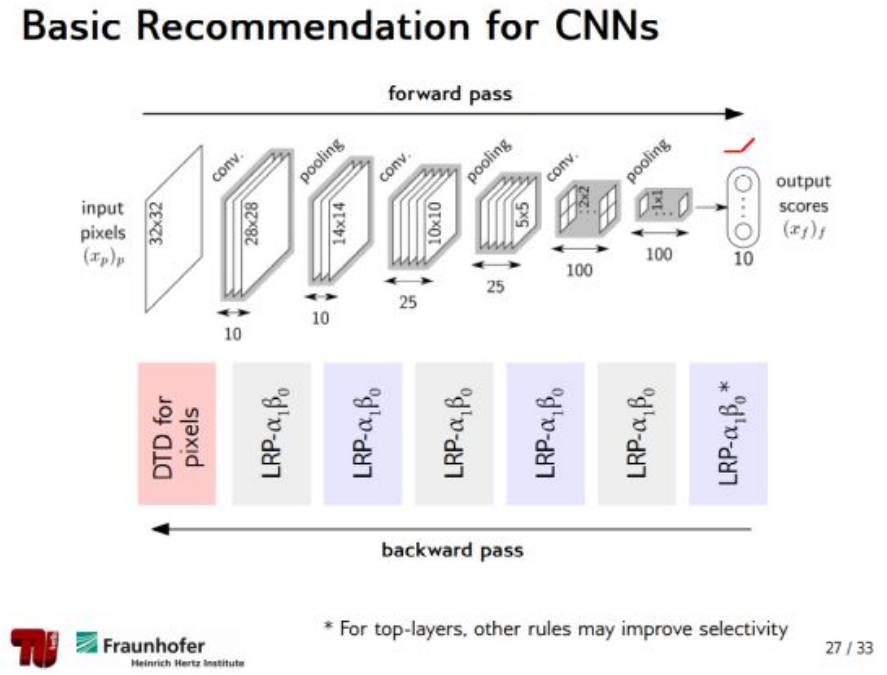
1. Choose a root point that is nearby and satisfies domain constraints

$$(x - \widetilde{x}^{(j)}) = t \cdot (x - l \odot 1_{w_j > 0} - h \odot 1_{w_j < 0}) \qquad (x - x^{(j)}) = t \cdot w_j$$

2. Inject it in the generic DTD rule to get the specific rule

$$R_p = \sum_j \frac{x_{pj} w_{pj} - l_p w_{pj}^+ - h_p w_{pj}^-}{\sum_p x_{pj} w_{pj} - l_p w_{pj}^+ - h_p w_{pj}^-} R_j \qquad R_p = \sum_j \frac{w_{pj}^2}{\sum_p w_{pj}^2} R_j$$

25 / 33

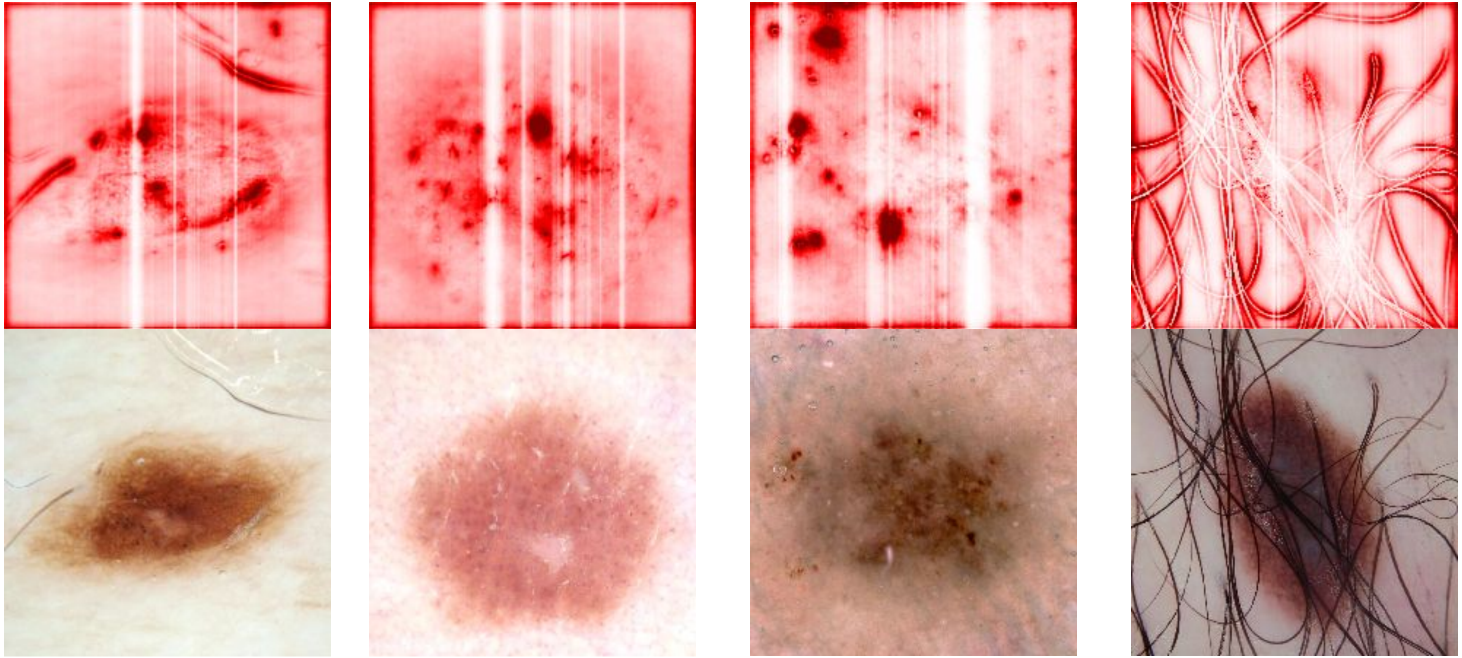


Basic Recommendation for CNNs

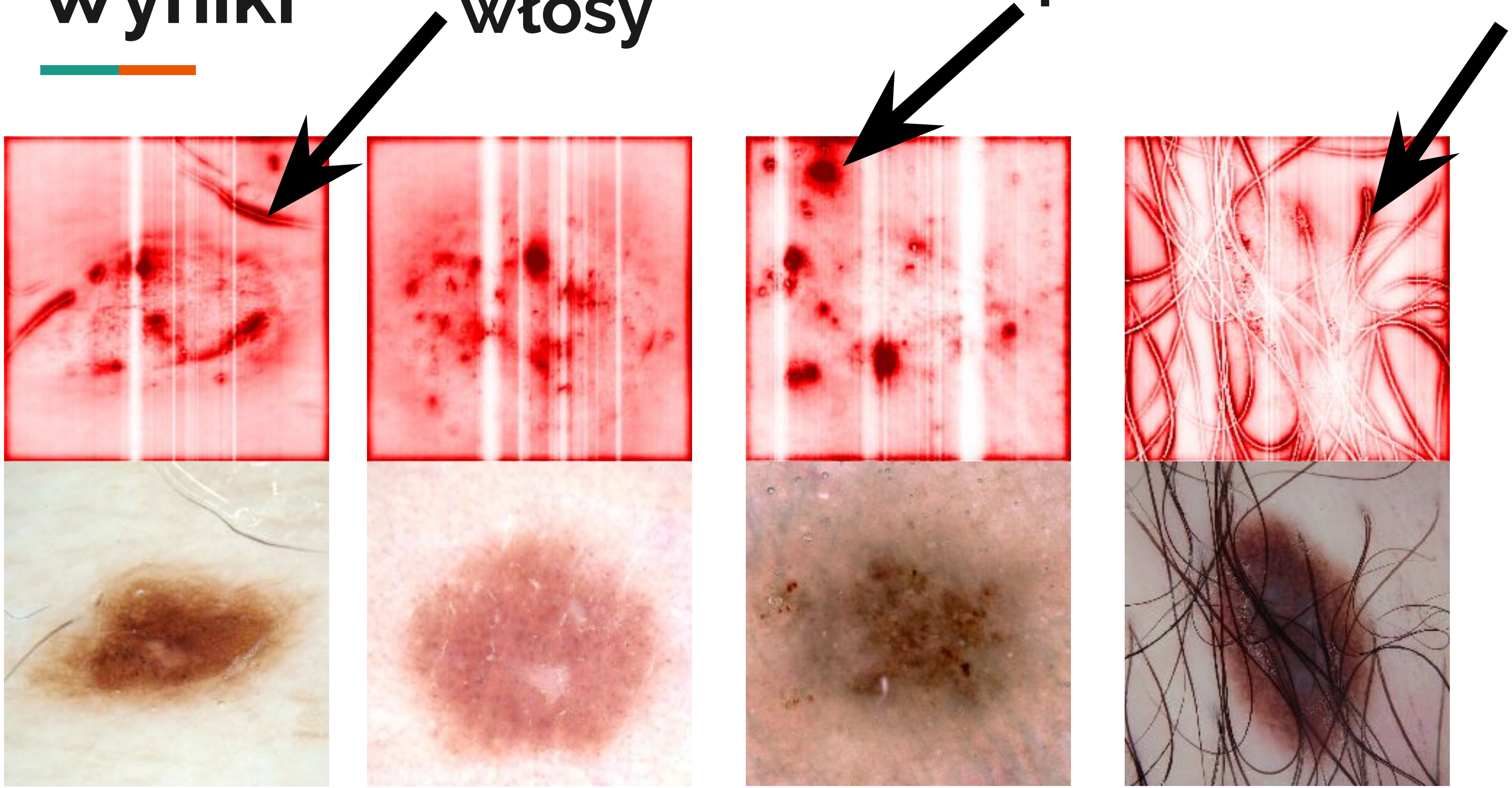
* For top-layers, other rules may improve selectivity

27 / 33

Wyniki



Wyniki



włosy

plamki żelu

włosy

Pytanie: Czy włosy, ramki wokół zdjęć, linijki i inne artefakty mogą powodować bias w modelu?

Czy to samo znamię z włosami i bez włosów może zostać inaczej zaklasyfikowane?

Jeden z podziałów w XAI

Objaśnienia globalne i lokalne

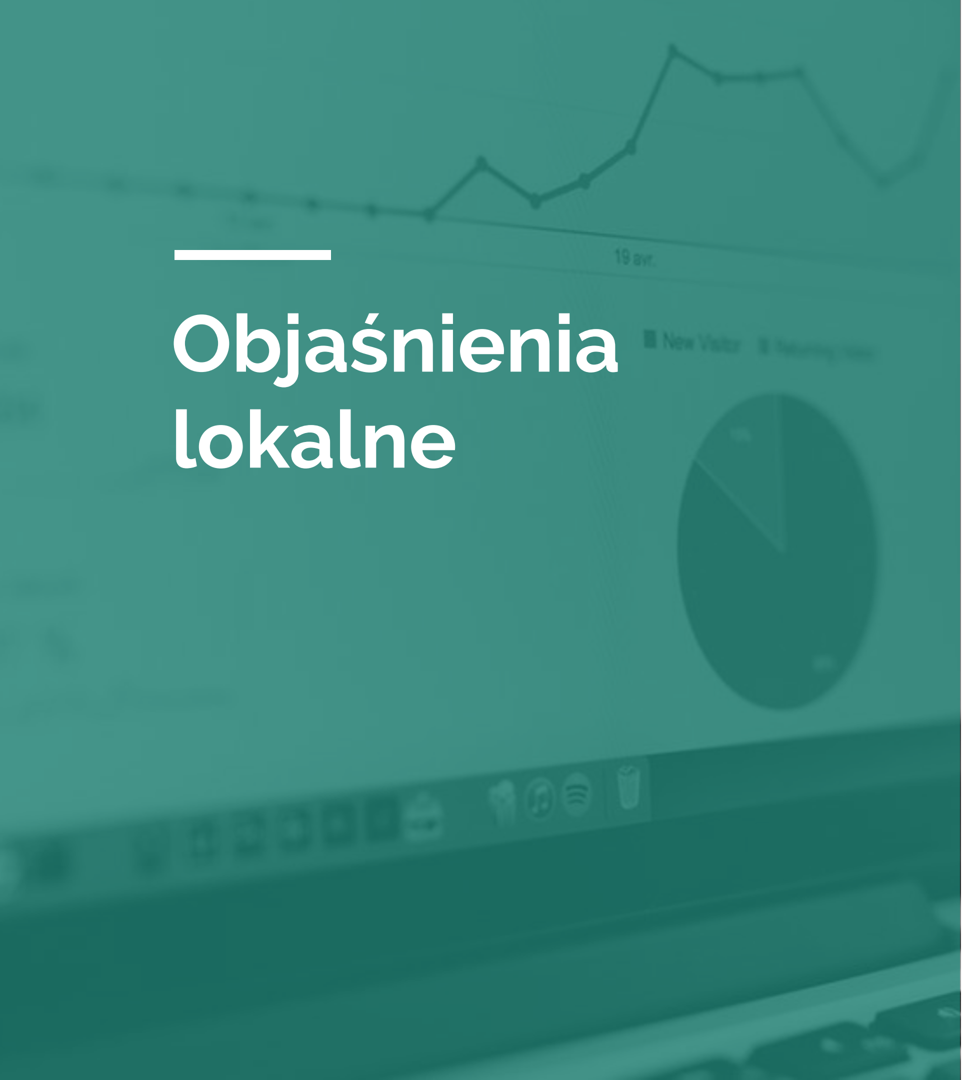
Globalne - gdy próbujemy wyjaśnić jak działa cały model

Lokalne - gdy próbujemy wyjaśnić jedną predykcję

Jeden z podziałów w XAI

Objaśnienia lokalne:

- żeby wyjaśnić pojedynczą predykcję
- np. jak ważne było każde z wejść dla końcowej predykcji?
- jaka zmiana wejść zmieniłaby predykcję?



Objaśnienia lokalne

Przykładowe metody:

- LIME
- SHAP
- LRP
- Anchors

Objaśnienia globalne

Objaśnienia globalne:

- żeby wyjaśnić jak działa cały model
- np. jakie wejścia/cechy mają zazwyczaj największy wpływ na predykcję?

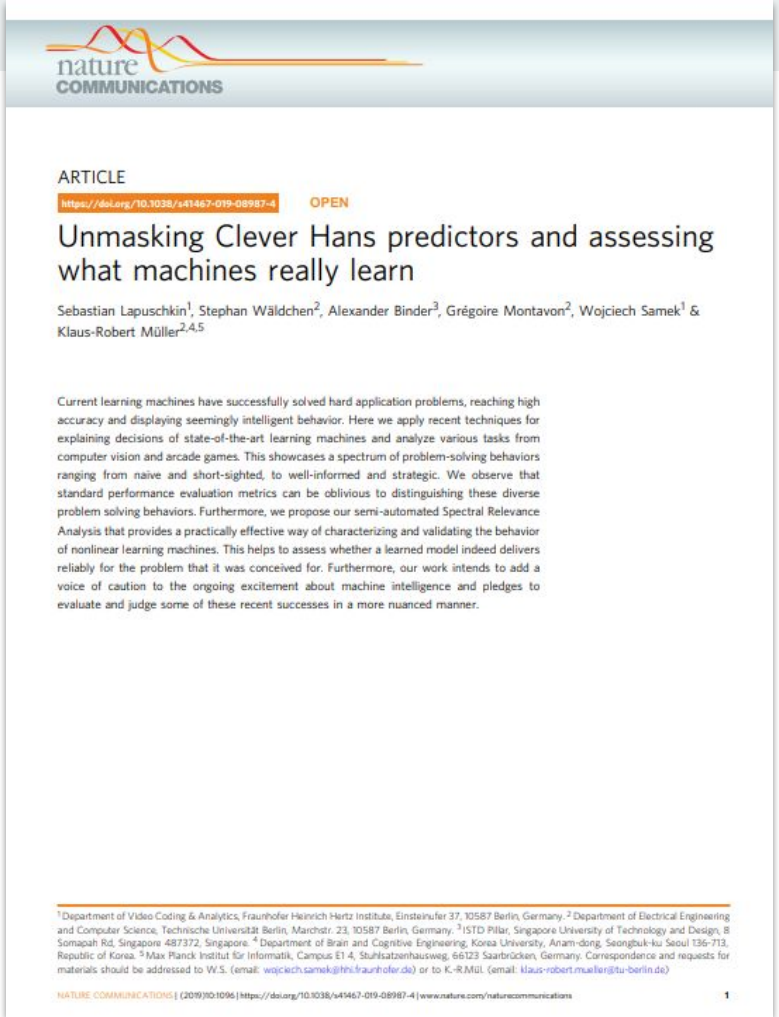
Objaśnienia globalne

Przykładowe metody:

- Summarized local explanations
- T-SNE on CNNs
- T-SNE on latent space

Spectral Relevance Analysis

Metoda pozwalająca na generację półautomatycznych objaśnień globalnych



nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-08987-4 OPEN

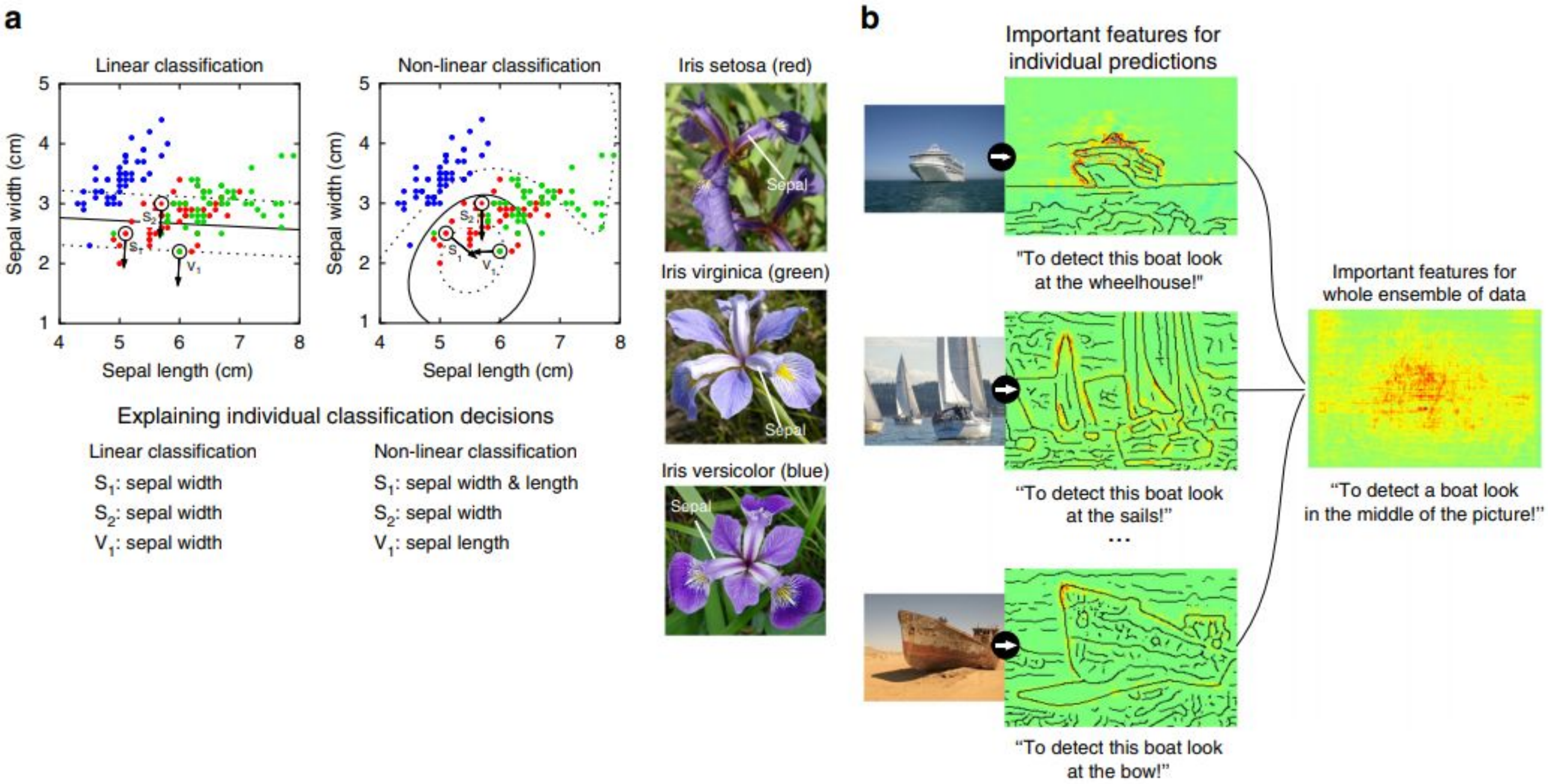
Unmasking Clever Hans predictors and assessing what machines really learn

Sebastian Lapuschkin[1], Stephan Wäldchen[2], Alexander Binder[3], Grégoire Montavon[2], Wojciech Samek[1] & Klaus-Robert Müller[2,4,5]

Current learning machines have successfully solved hard application problems, reaching high accuracy and displaying seemingly intelligent behavior. Here we apply recent techniques for explaining decisions of state-of-the-art learning machines and analyze various tasks from computer vision and arcade games. This showcases a spectrum of problem-solving behaviors ranging from naive and short-sighted, to well-informed and strategic. We observe that standard performance evaluation metrics can be oblivious to distinguishing these diverse problem solving behaviors. Furthermore, we propose our semi-automated Spectral Relevance Analysis that provides a practically effective way of characterizing and validating the behavior of nonlinear learning machines. This helps to assess whether a learned model indeed delivers reliably for the problem that it was conceived for. Furthermore, our work intends to add a voice of caution to the ongoing excitement about machine intelligence and pledges to evaluate and judge some of these recent successes in a more nuanced manner.

[1] Department of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, Einsteinufer 37, 10587 Berlin, Germany. [2] Department of Electrical Engineering and Computer Science, Technische Universität Berlin, Marchstr. 23, 10587 Berlin, Germany. [3] ISTD Pillar, Singapore University of Technology and Design, 8 Somapah Rd, Singapore 487372, Singapore. [4] Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-ku Seoul 136-713, Republic of Korea. [5] Max Planck Institut für Informatik, Campus E1 4, Stuhlsatzenhausweg, 66123 Saarbrücken, Germany. Correspondence and requests for materials should be addressed to W.S. (email: wojciech.samek@hhi.fraunhofer.de) or to K.-R.Mül. (email: klaus-robert.mueller@tu-berlin.de)

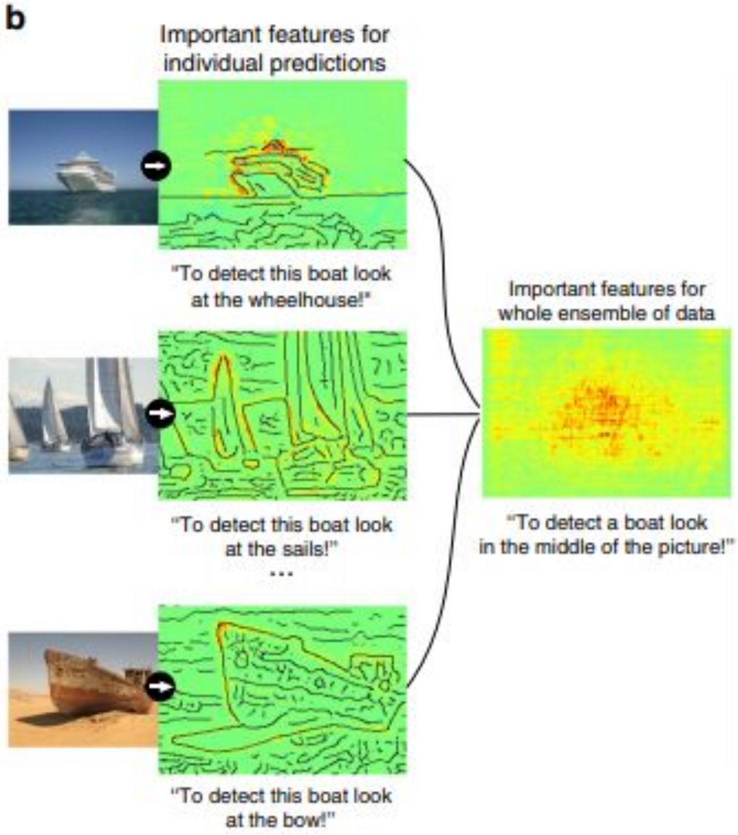
Spectral Relevance Analysis


a
Linear classification
Non-linear classification
Sepal width (cm)
Sepal length (cm)
1
2
3
4
5
4
5
6
7
8
S_1
S_2
V_1
Iris setosa (red)
Iris virginica (green)
Iris versicolor (blue)
Sepal
Explaining individual classification decisions
Linear classification
S_1: sepal width
S_2: sepal width
V_1: sepal width
Non-linear classification
S_1: sepal width & length
S_2: sepal width
V_1: sepal length
b
Important features for individual predictions
"To detect this boat look at the wheelhouse!"
"To detect this boat look at the sails!"
...
"To detect this boat look at the bow!"
Important features for whole ensemble of data
"To detect a boat look in the middle of the picture!"

Spectral Relevance Analysis

Idea

stworzyć liczne objaśnienia lokalne i na ich podstawie zrobić uśrednione objaśnienie globalne.

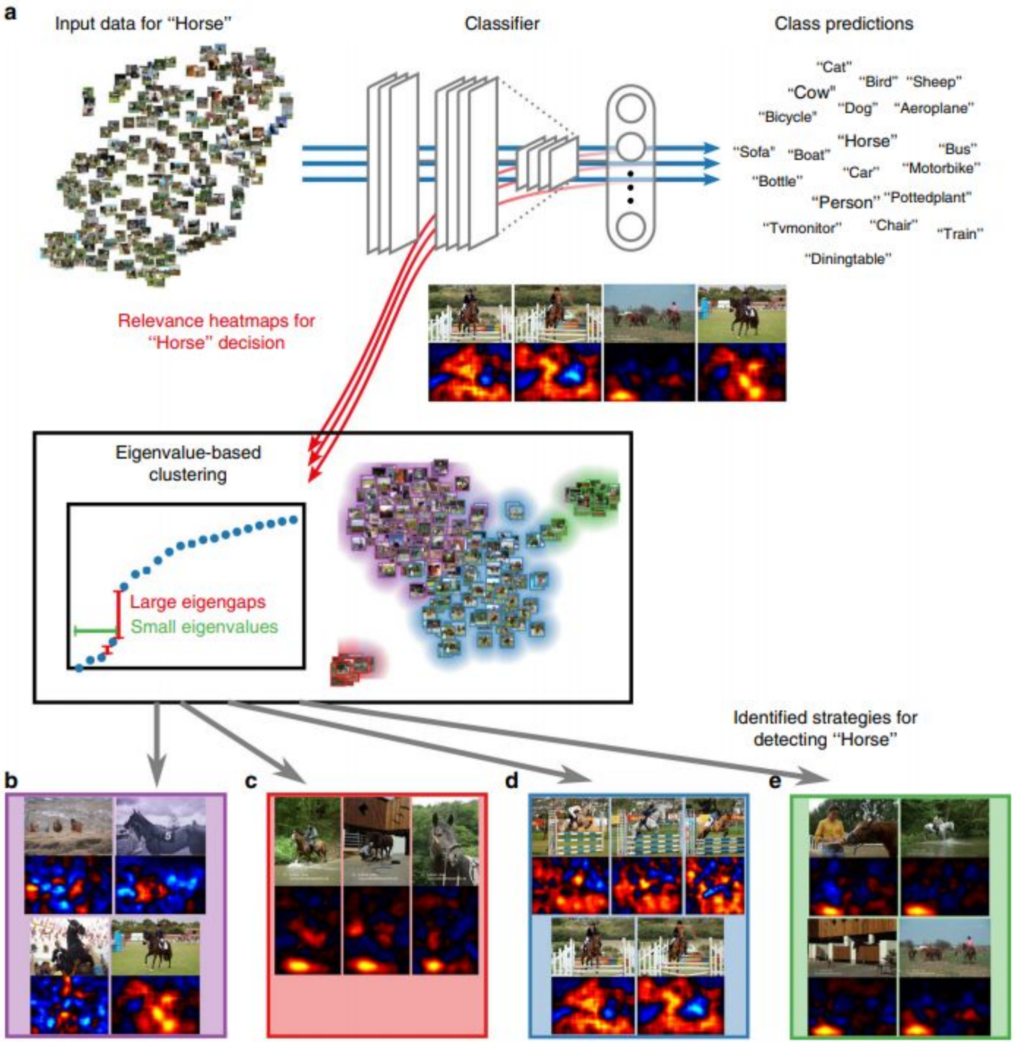

b
Important features for individual predictions
"To detect this boat look at the wheelhouse!"
"To detect this boat look at the sails!"
...
"To detect this boat look at the bow!"
Important features for whole ensemble of data
"To detect a boat look in the middle of the picture!"

SpRAy

Step-by-step

Krok 0. Przygotuj wytrenowany model i dane które chcesz przetestować

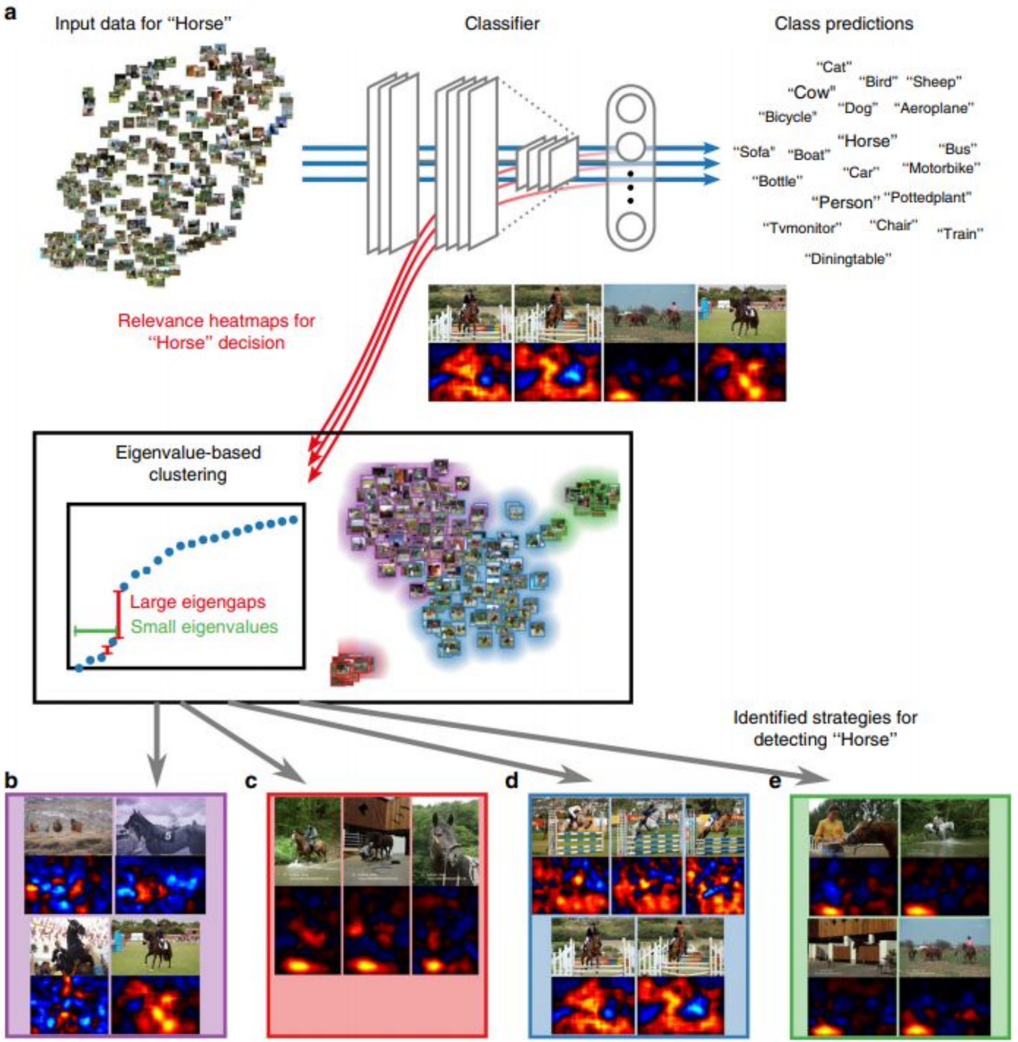
Krok 1. Wygeneruj mapy ciepła/uwagi przy pomocy metody LRP



SpRAy

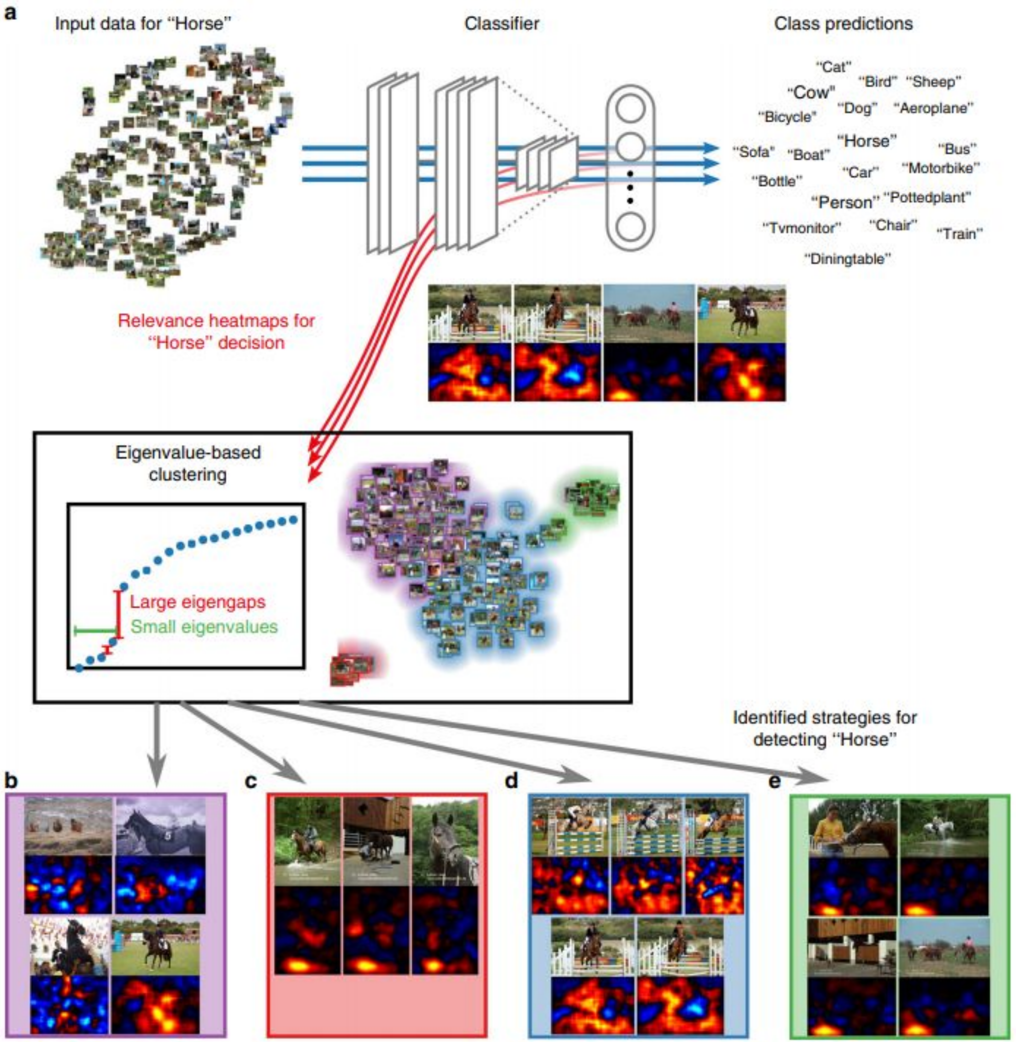
Krok 2. Przetwórz odpowiednio otrzymane mapy ciepła: zmniejsz je, znormalizuj

Krok 3. Przeprowadź klasteryzacje (spectral clustering) na mapach ciepła



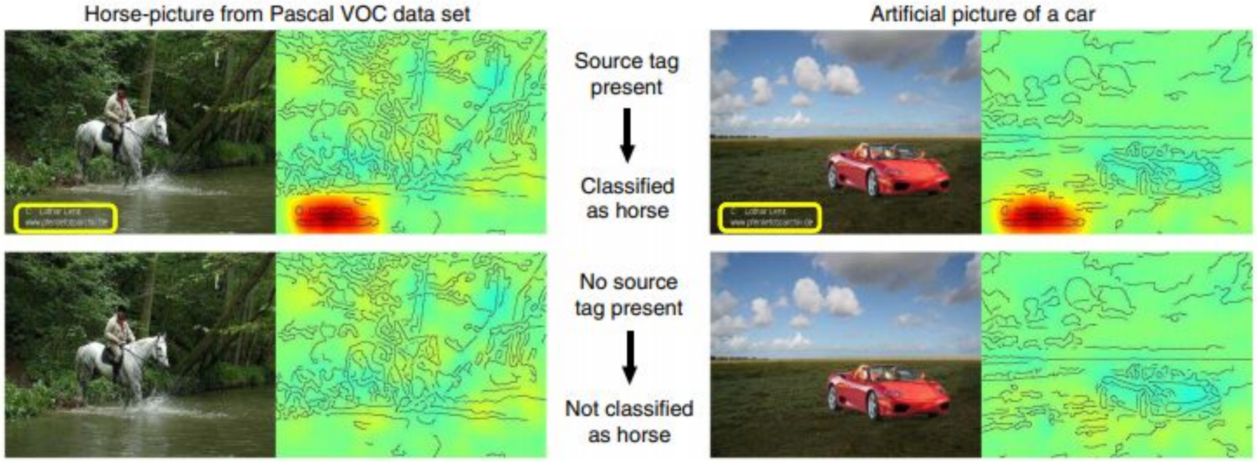
SpRAy

Krok 4. Przeanalizuj zidentyfikowane strategie predykcji



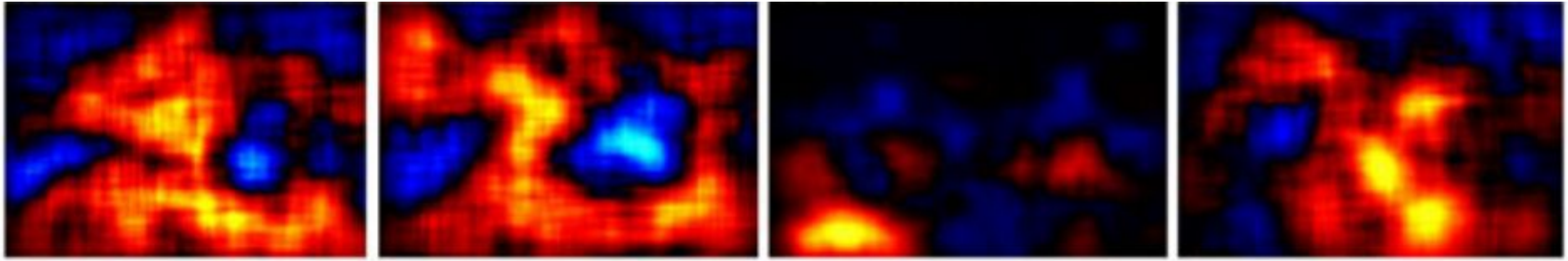
SpRAy - wyniki

Cytat: *The Fisher vector classifier trained on the PASCAL VOC 2007 dataset focuses on a source tag present in about one-fifth of the horse figures. Removing the tag also removes the ability to classify the picture as a horse. Furthermore, inserting the tag on a car image changes the classification from car to horse*



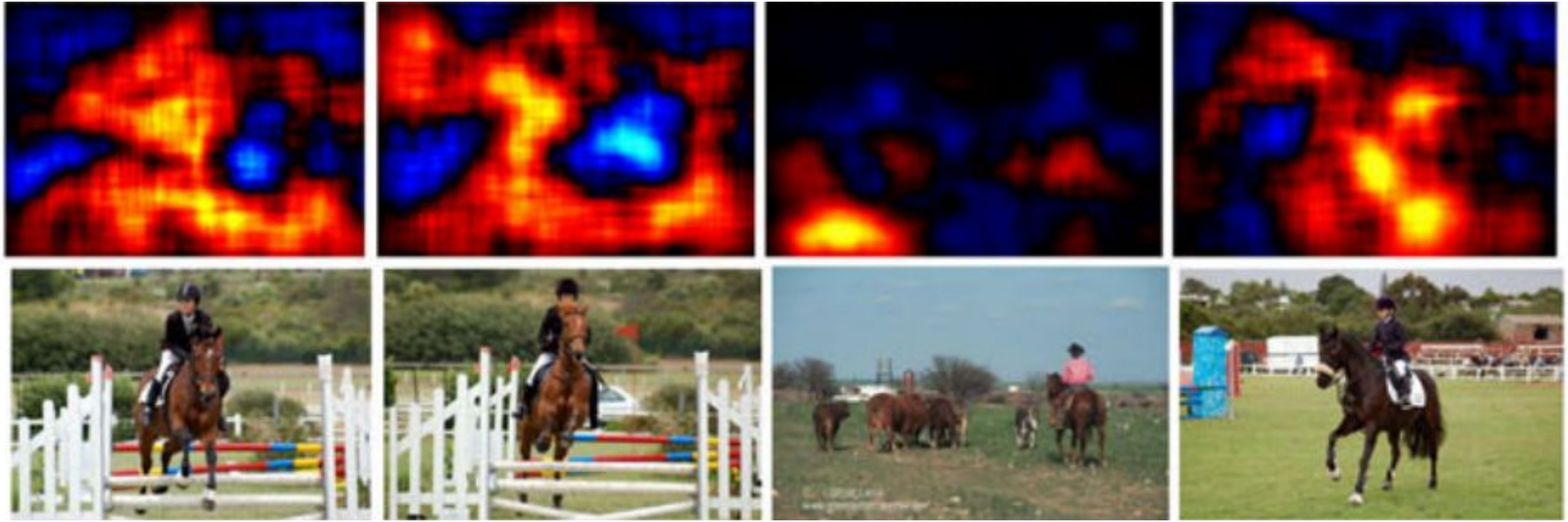
SpRAy - problemy

Analizator widzi tylko mapy ciepła, które wyglądają tak:



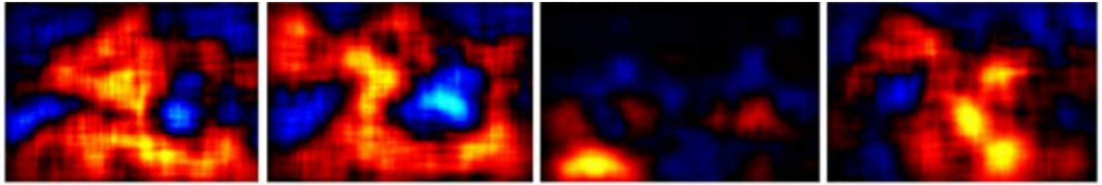
Co było na tych zdjęciach?

SpRAy - problemy



SpRAy - problemy

Analizator widzi tylko mapy ciepła, które wyglądają tak:

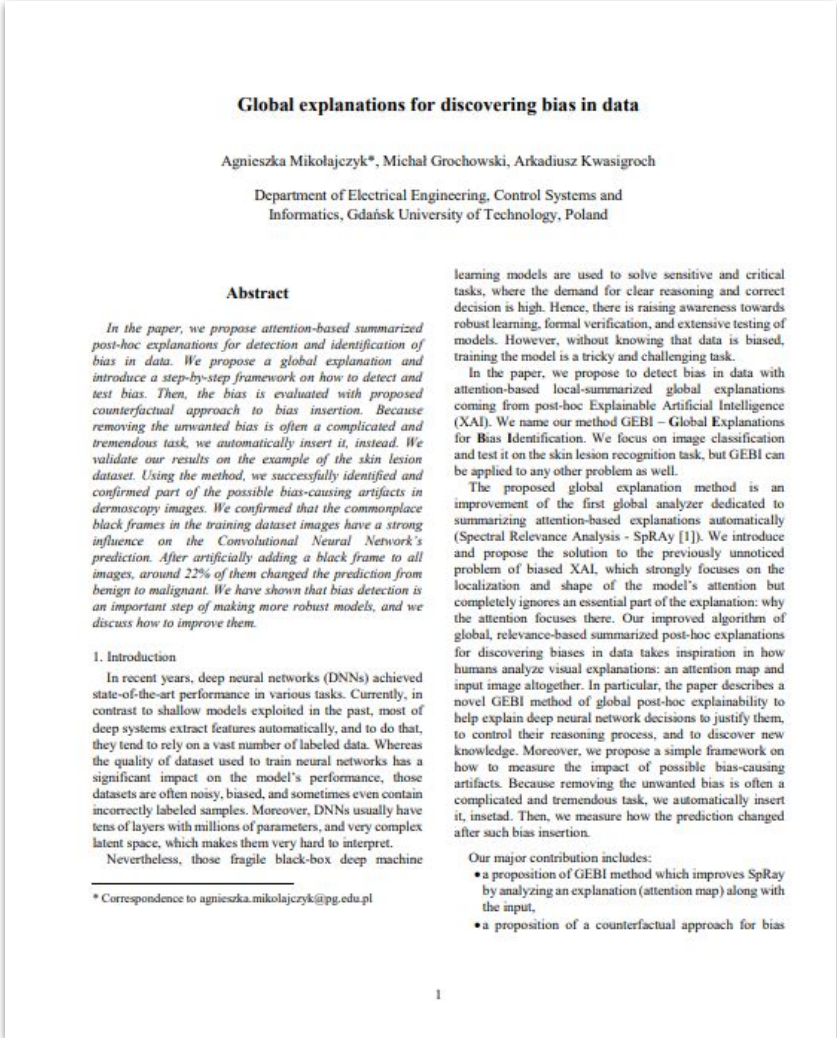


Czy analizator globalny bazujący wyłącznie na heatmapach ma sens?

Jest zbiasowany przeciwko: kształtowi pola uwagi i lokalizacji

SpRAy 2.0 - GEBI

Propozycja modyfikacji metody SpRAy



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, insetad. Then, we measure how the prediction changed after such bias insertion.

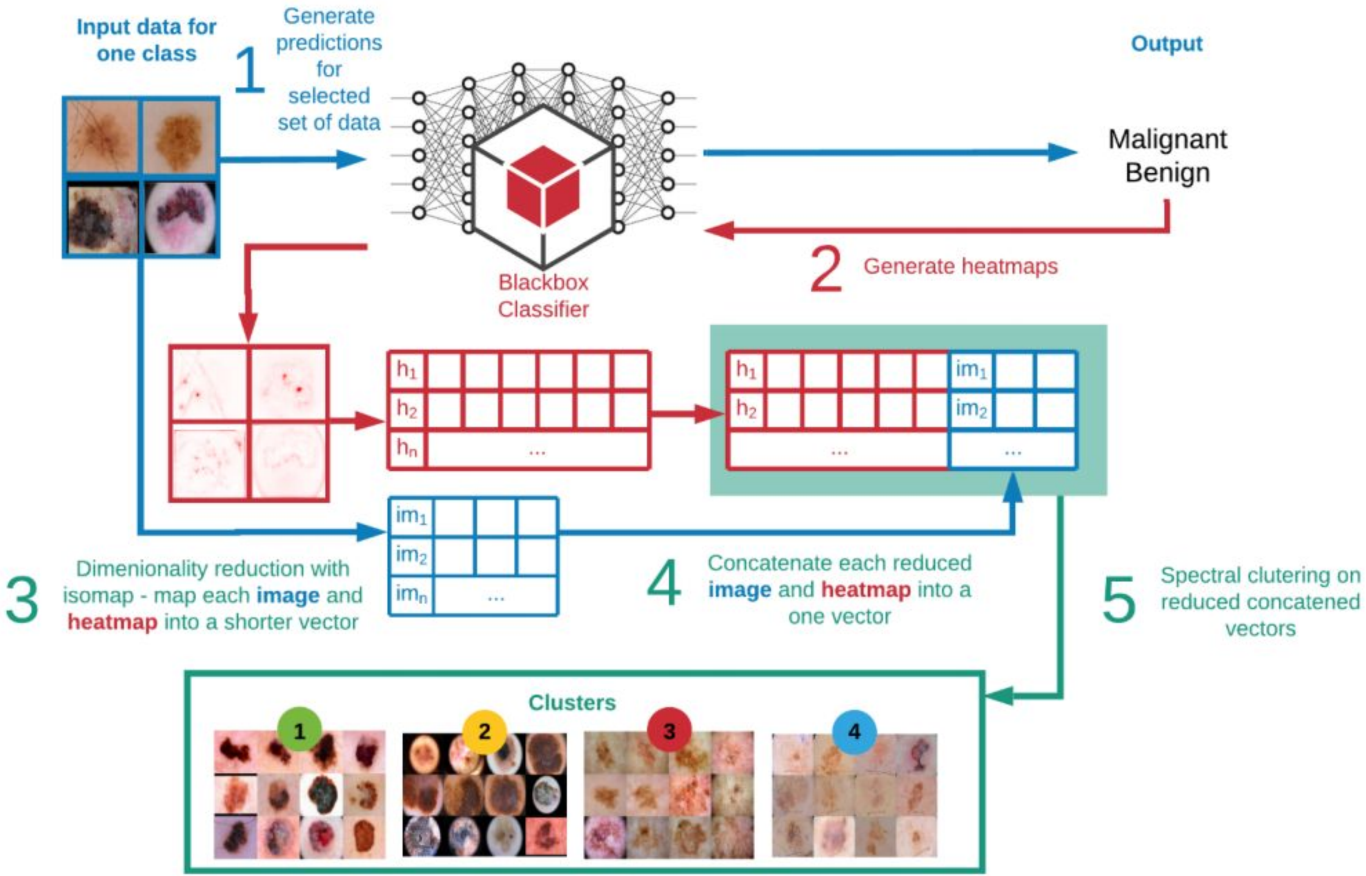
Our major contribution includes:

- 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

Global Explanations for Bias Identification

- **Główna idea:** tworzymy objaśnienia globalne analizując objaśnienia lokalne jak człowiek: analizując jednocześnie parę mapy ciepła oraz odpowiadającego jej wejścia (obrazu)



Input data for one class

1 Generate predictions for selected set of data

Output

2 Generate heatmaps

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

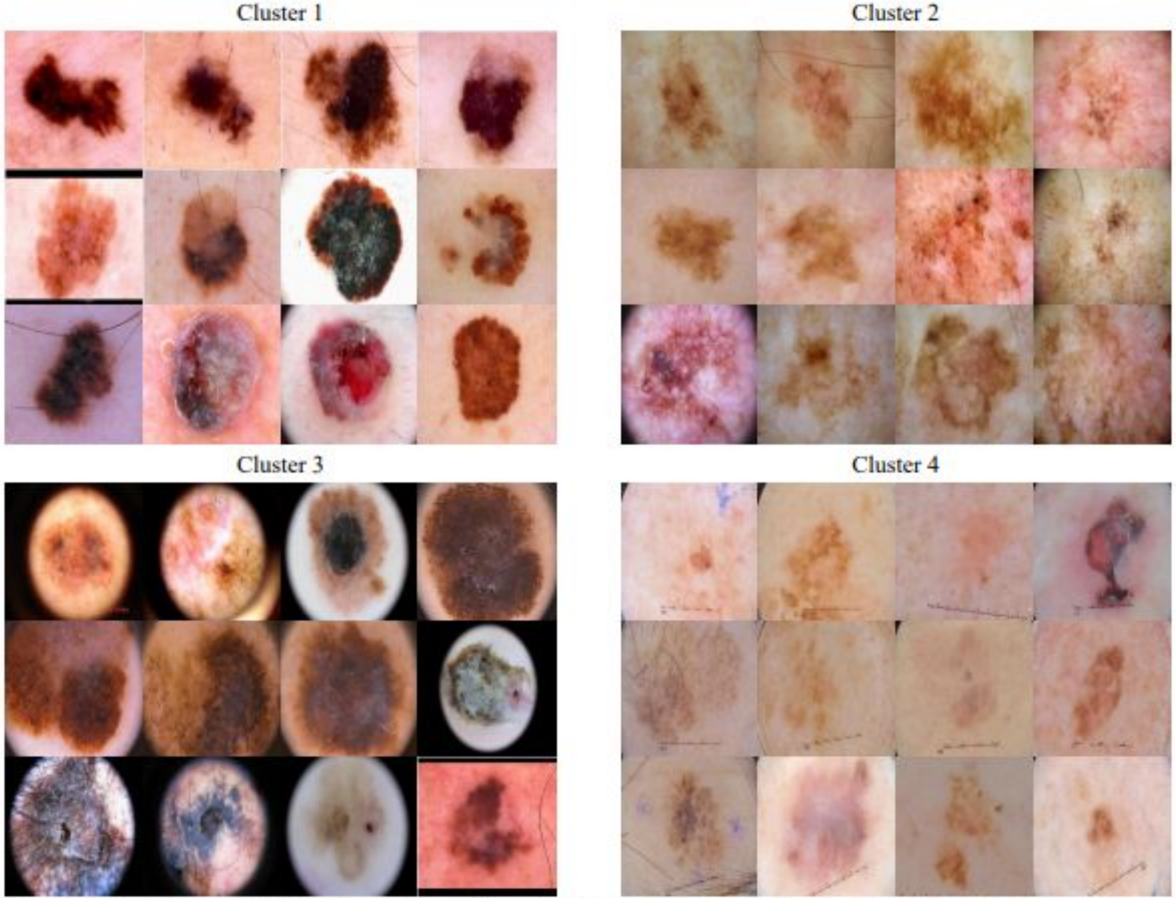
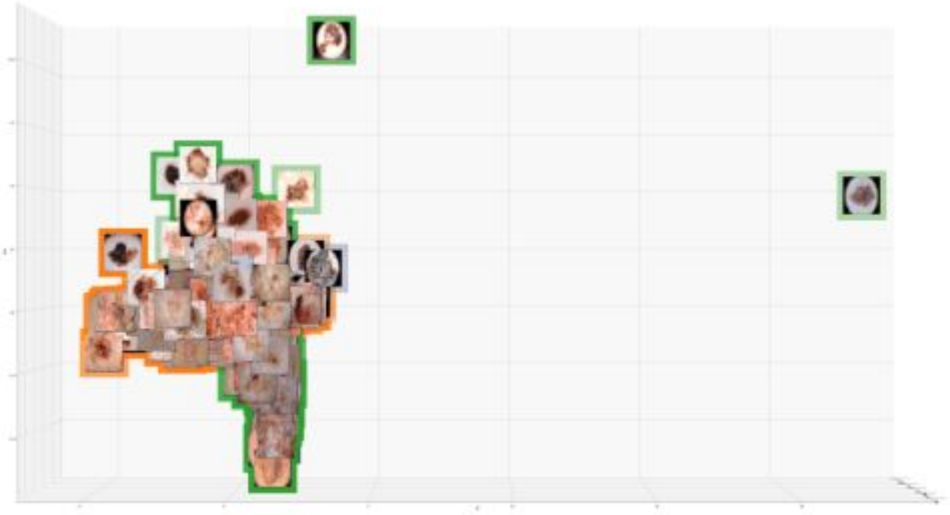
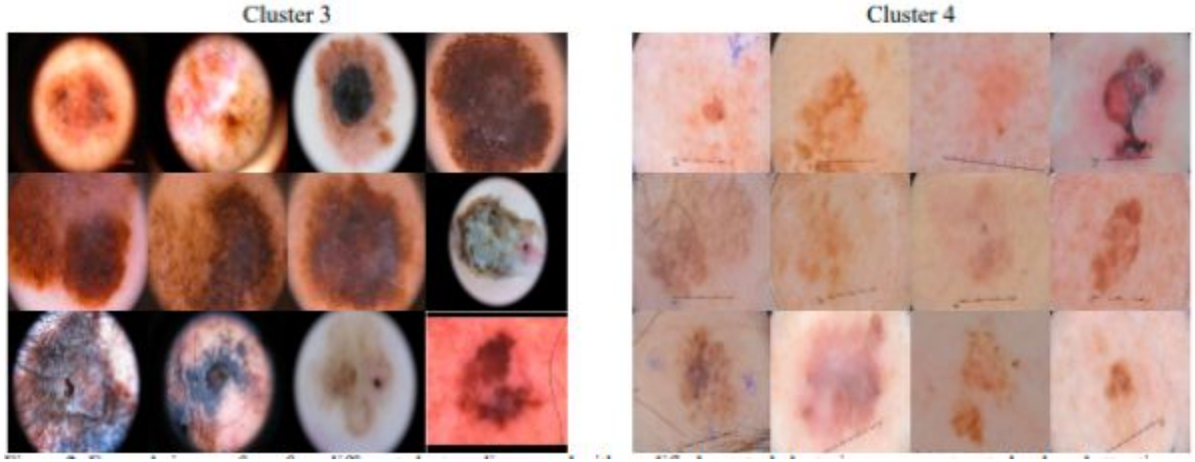


Figure 2: Example images from four different clusters discovered with modified spectral clustering on concatenated reduced attention maps and input images. Cluster 1 shows mostly dark skin lesions with clear border; Cluster 2 shows very textured skin lesions with numerous visible structures; Cluster 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)



GEBI

- ➔ Podobnie jak w Spray analizujemy wytworzone clustry
- ➔ **Hipoteza:** Czarne ramki oraz oznaczenia linijki powodują bias



Czy to rzeczywiście spowodowało bias?

- Usuwanie artefaktów ze zdjęcia jest trudne, oraz samo w sobie może spowodować powstanie nowych artefaktów
- Zamiast tego zaproponowaliśmy dodanie artefaktu do zdjęcia i sprawdzenie jak zmieniła się predykcja

Czy to rzeczywiście spowodowało bias?

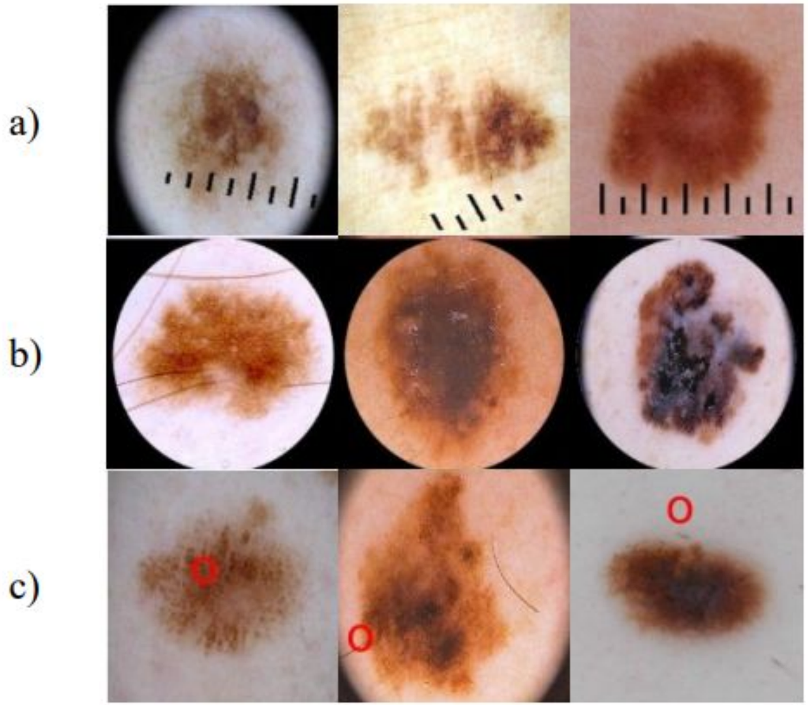


Figure 3: Modified examples by insertion of artificial bias: a) ruler markings, b) black frames, c) red circles

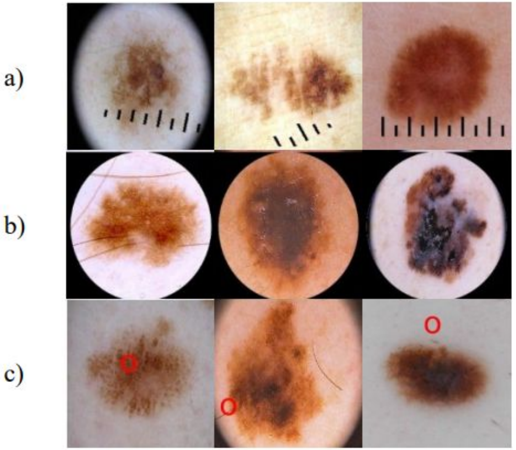


Figure 3: Modified examples by insertion of artificial bias: a) ruler markings, b) black frames, c) red circles

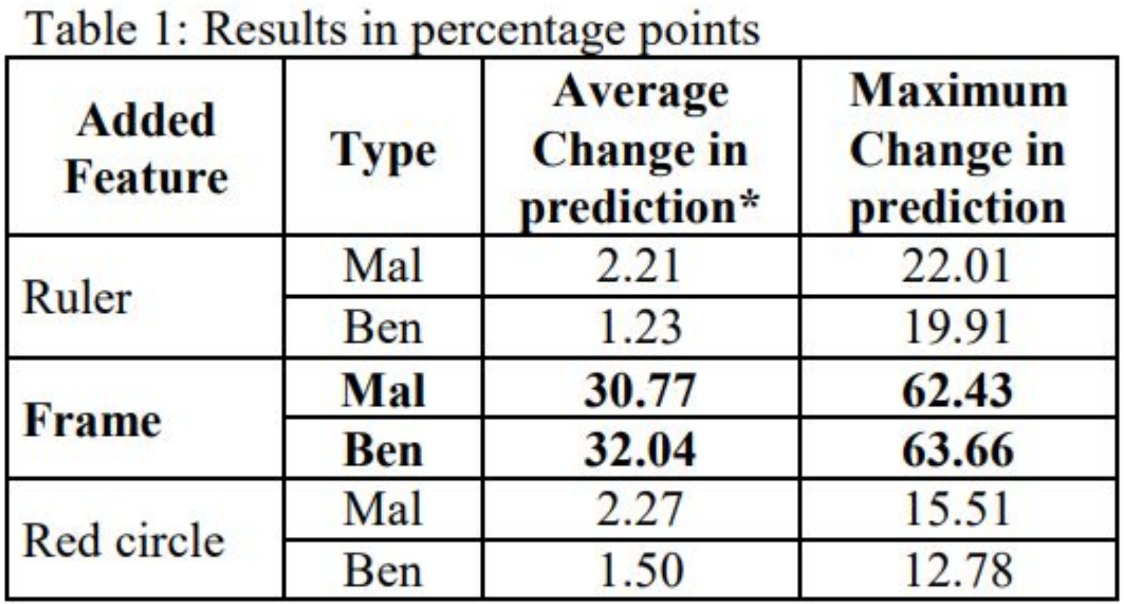
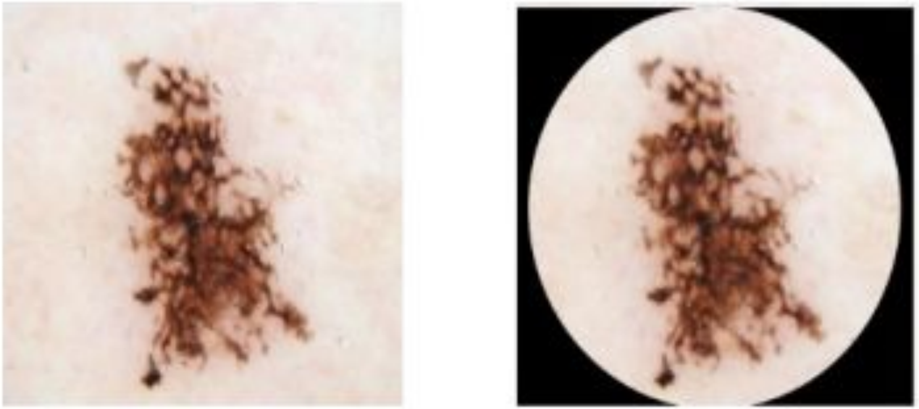


Table 1: Results in percentage points

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



Benign: 0.1

Malignant: 0.89

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

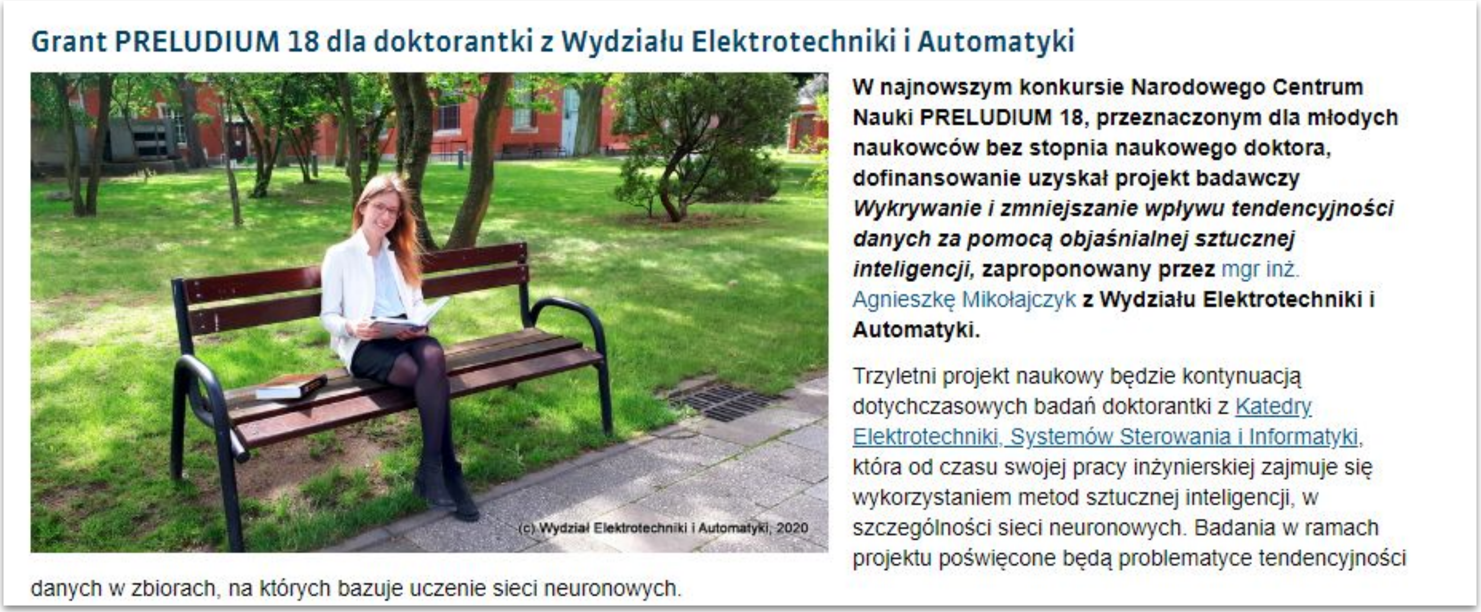
Podsumowanie GEBI

- ➔ Prosta ale bardzo skuteczna metoda na pozbycie się biasu z metody Spray
- ➔ Propozycja zmierzenia możliwego biasu dzięki metodzie “Bias insertion”
- ➔ Największy wpływ miało dodanie czarnej ramki: aż 22% znamion klasyfikowanych jako łagodne, zostało zakwalifikowane jako złośliwe po dodaniu ramki!

Co dalej?

Preludium

- ➔ Temat: “Wykrywanie i zmniejszanie wpływu tendencyjności danych za pomocą objaśnialnej sztucznej inteligencji”



Grant PRELUDIUM 18 dla doktorantki z Wydziału Elektrotechniki i Automatyki

(c) Wydział Elektrotechniki i Automatyki, 2020

W najnowszym konkursie Narodowego Centrum Nauki PRELUDIUM 18, przeznaczonym dla młodych naukowców bez stopnia naukowego doktora, dofinansowanie uzyskał projekt badawczy *Wykrywanie i zmniejszanie wpływu tendencyjności danych za pomocą objaśnialnej sztucznej inteligencji*, zaproponowany przez mgr inż. Agnieszkę Mikołajczyk **z Wydziału Elektrotechniki i Automatyki.**

Trzyletni projekt naukowy będzie kontynuacją dotychczasowych badań doktorantki z Katedry Elektrotechniki, Systemów Sterowania i Informatyki, która od czasu swojej pracy inżynierskiej zajmuje się wykorzystaniem metod sztucznej inteligencji, w szczególności sieci neuronowych. Badania w ramach projektu poświęcone będą problematyce tendencyjności danych w zbiorach, na których bazuje uczenie sieci neuronowych.

Preludium zadania

- ➔ Developing globally aware local explanations for prediction justification
- ➔ Developing global explanations for detecting undesirable bias in data
- ➔ Developing trainable attention for eliminating influences of undesirable bias in data on the model

Preludium

- Projekt 36 miesięcy
- Zastosowanie głównie w obrazach
- Chętnie nawiążę współpracę, w szczególności w badaniach w tej samej tematyce ale metod stosowanych w NLP

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