

Zastosowanie algorytmów ewolucyjnych w technikach wyjaśnialnej sztucznej inteligencji

Adam Żychowski



Czym jest wyjaśnialność?

Wyjaśnialność AI polega na:

- dostarczeniu użytkownikowi informacji potrzebnych do zrozumienia, dlaczego system autonomiczny zachowuje się w określony sposób w określonych okolicznościach
- dostępie do rzetelnych informacji o działaniu modelu AI

responsible AI, AI ethics, transparent, interpretable, explainable

AI act (Unia Europejska)

OECD AI Principles (Organizacja Współpracy Gospodarczej i Rozwoju)

UNESCO AI Ethics (Organizacja Narodów Zjednoczonych)

Blueprint for an AI Bill of Rights (US government)

MIT, Alan Turing Institute, AI Now, AAAI, ACM, IEEE i wiele więcej

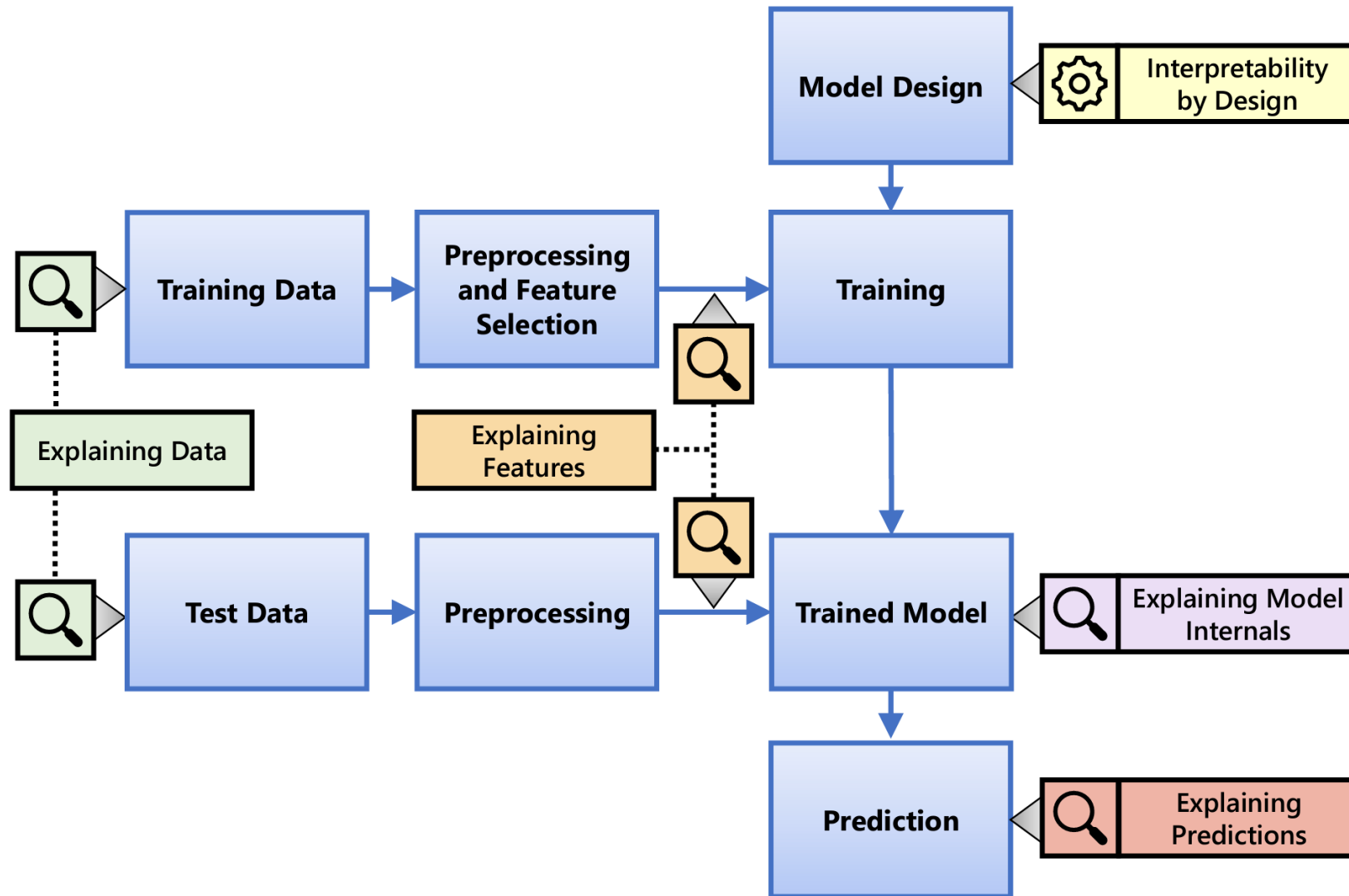
Czym jest wyjaśnialność?

Różne “stopnie wyjaśnialności” w zależności od:

- domeny
- ryzyka
- interesariuszy

Różne “poziomy wyjaśnialności”:

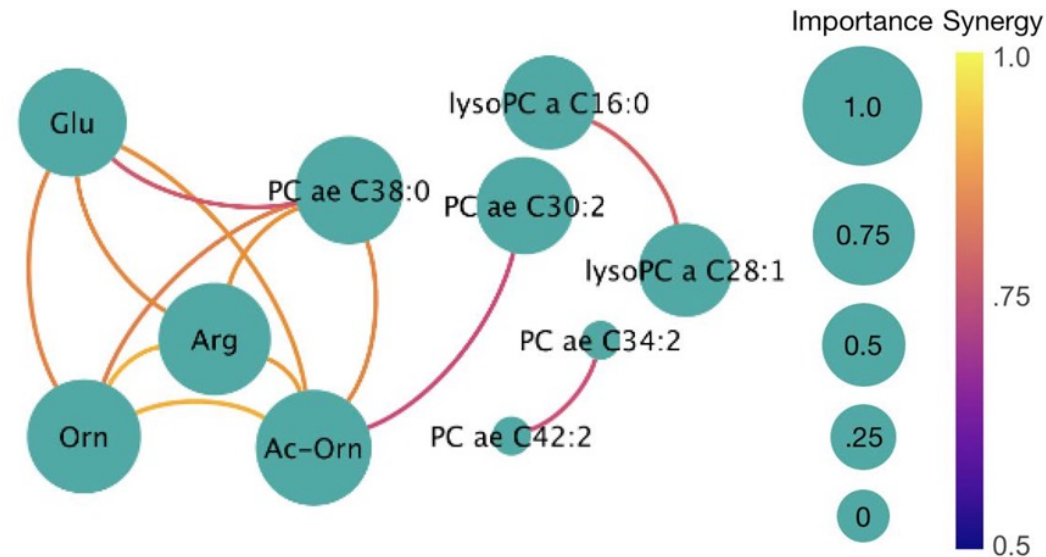
- **proces decyzyjny** - przyczyny, które doprowadziły do wydania określonej decyzji,
- **dane** - jakie dane zostały wykorzystane, w jaki sposób, czy nie powodują biasu, czy są rzetelne,
- **model** - jak działa zastosowany model,
- **odpowiedzialność** - kto odpowiada za system, ponosi konsekwencje jego działań,
- **bezpieczeństwo** - dowód na niezawodność, dokładność, odporność AI,
- **wpływ** - wpływ na jednostkę, grupy, procesy społeczne,
- **wynik** - uzasadnienie dlaczego wynik jest obiektywny, sprawiedliwy.



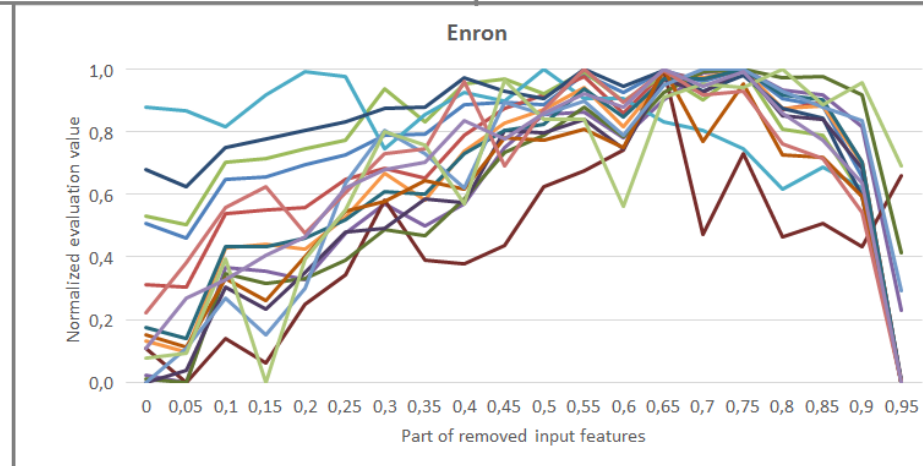
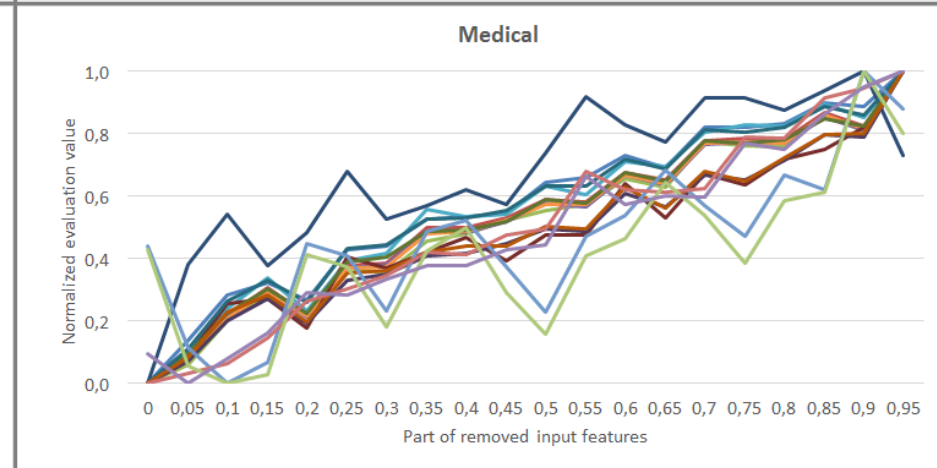
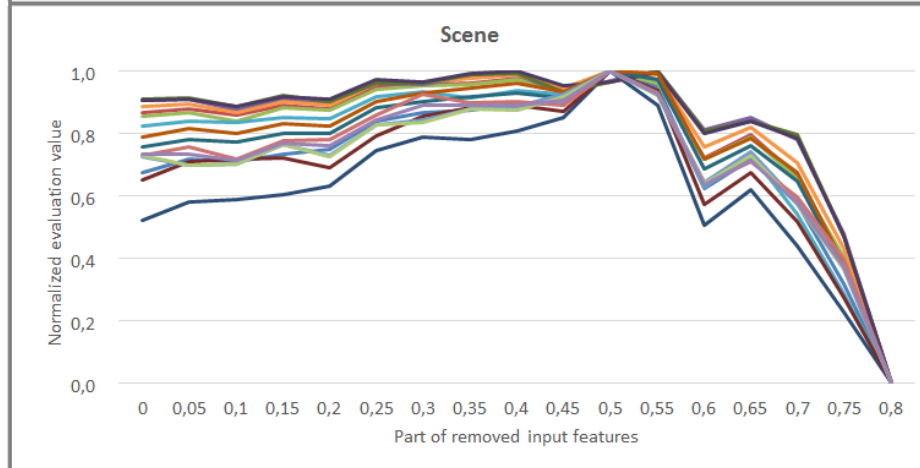
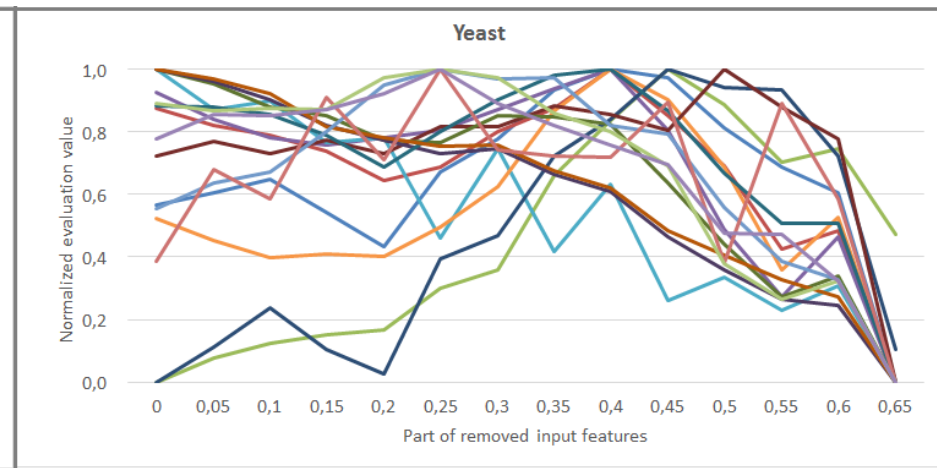
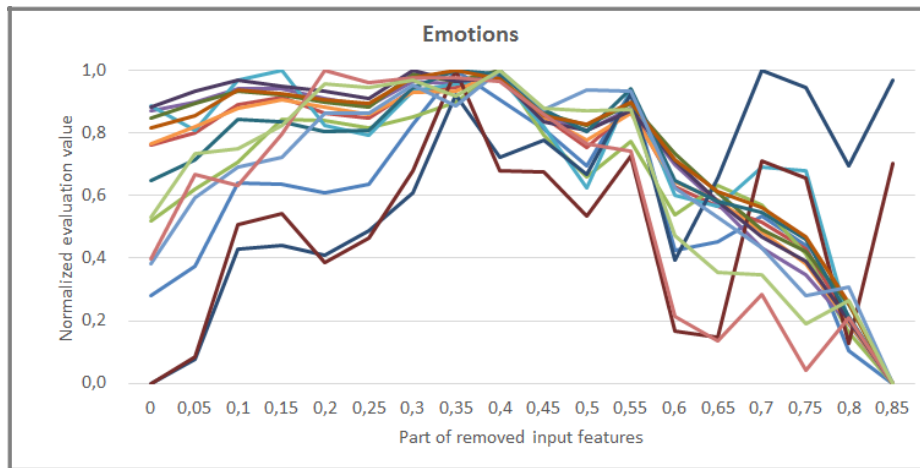
[1]

Wybór cech wejściowych

Przykład: wektor 0-1, gdzie każdy element odpowiada jednej cenie wejściowej i informuje czy powinna ona zostać włączona do uczenia



Robertson, Jake, and Ting Hu. "An evolutionary approach to interpretable learning." *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 2021.



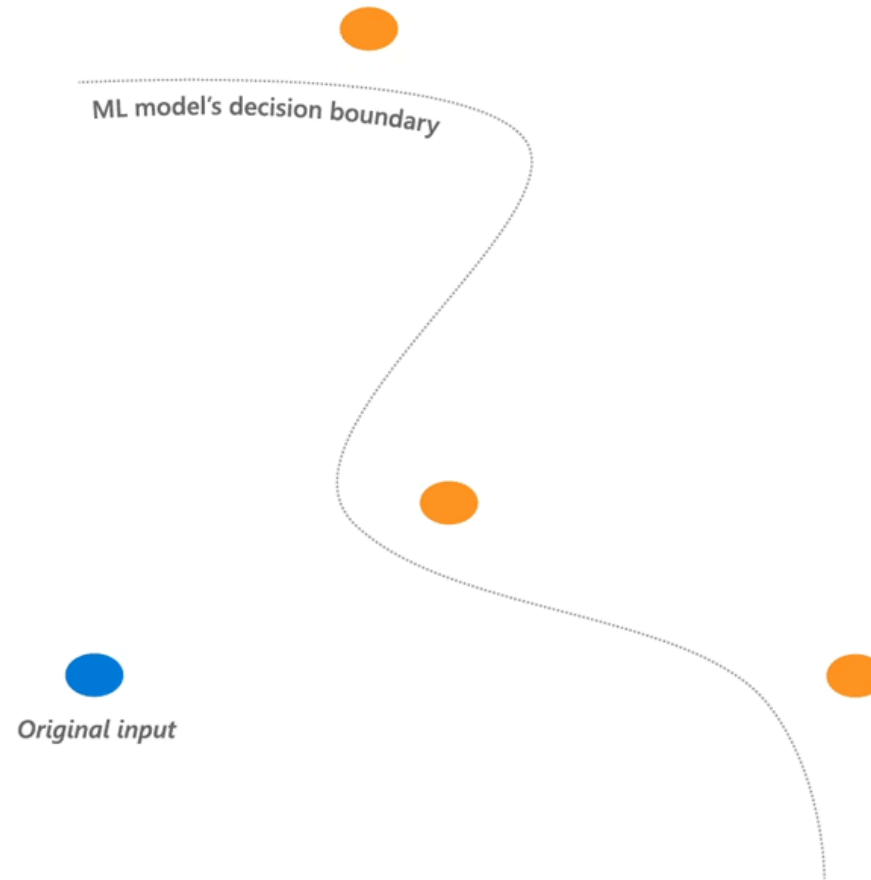
Objaśnienia oparte na przykładach

- **Counterfactual explanations** – mówią, jak i jaka instancja musi się zmienić, aby model znacząco zmienił swoją predykcję.
- **Adversarial examples** – przykłady używane do oszukiwania modeli uczenia maszynowego. W ich przypadku nacisk kładziony jest na odwrócenie przewidywań modelu, a nie na ich wyjaśnienie.
- **Prototypes** – prototypy są wyborem reprezentatywnych instancji z danych.
- **Influential instances** – wpływowe instancje to instancje danych treningowych, które miały największy wpływ na parametry modelu lub na same predykcje.
- **k-nearest neighbors** – wybór z danych treningowych k najbardziej podobnych przykładów.

Counterfactual Examples

**Original class:
Loan rejected**

**Desired class:
Loan approved**



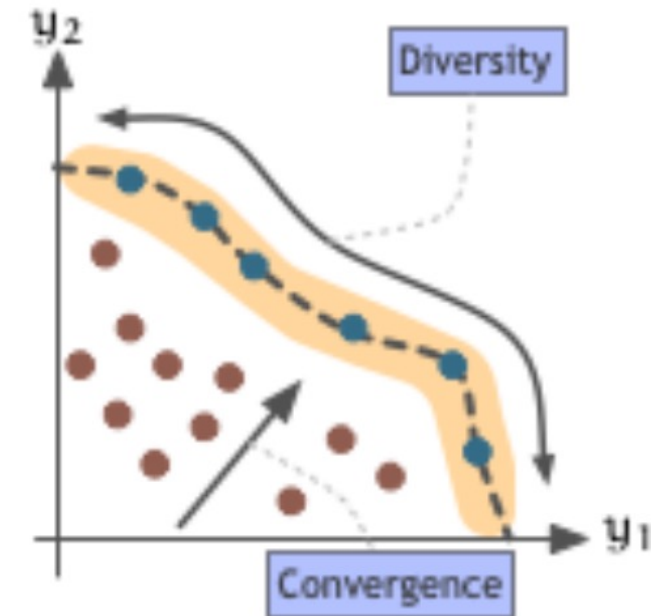
Counterfactual explanations

Kilka celów:

- zmiana predykcji w określony sposób – minimalizacja odległości predykcji od tej oczekiwanej
- minimalizacja odległości od modyfikowanej instancji danych wejściowych
- minimalizacja podobieństwa do innych danych wejściowych

Algorytmy ewolucyjne dobrze sprawdzają się w optymalizacji wielokryterialnej, np. NSGA-II.

Dandl, Susanne, et al. "Multi-objective counterfactual explanations." *International Conference on Parallel Problem Solving from Nature*. Cham: Springer International Publishing, 2020.



Young \rightarrow Old

Query



DiME

DiVE



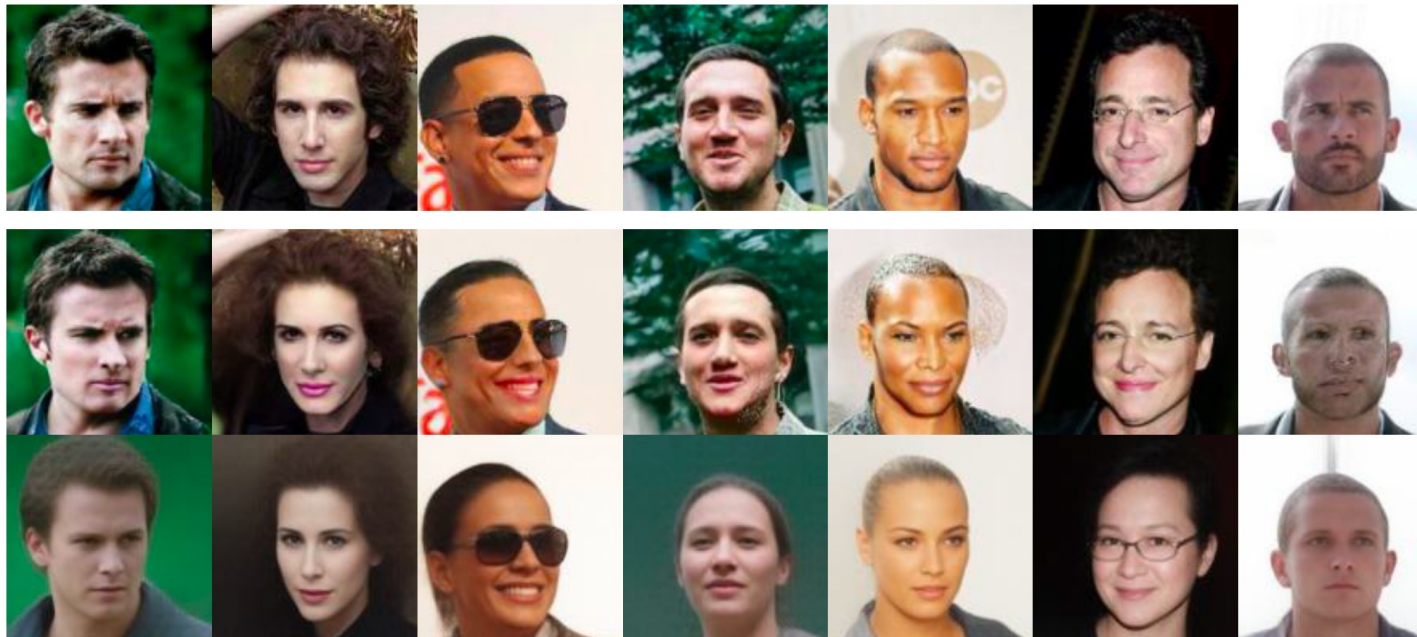
Male \rightarrow Female

Query



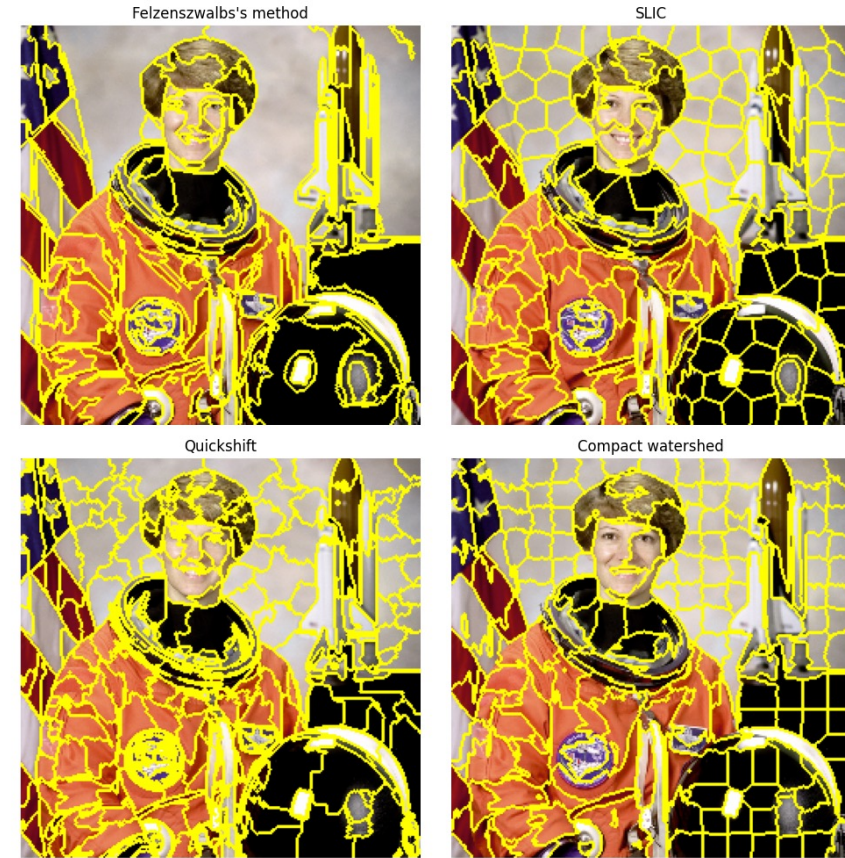
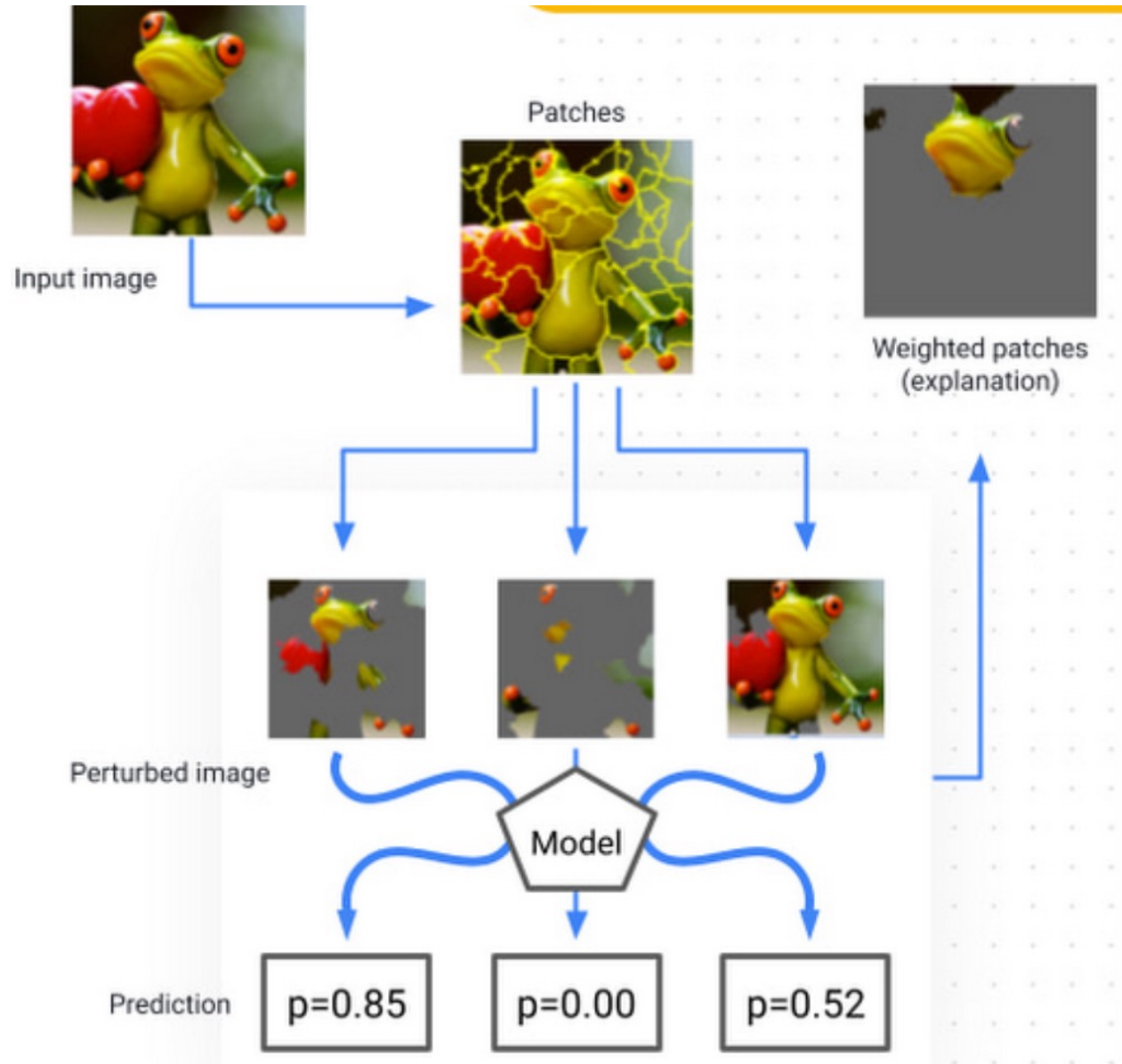
DiME

DiVE



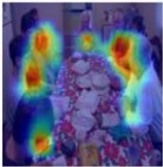
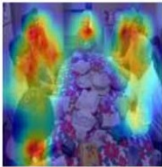









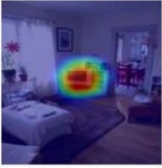
















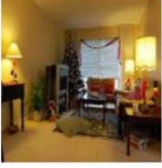



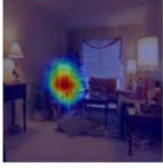
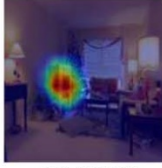
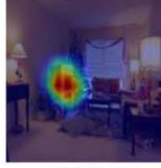
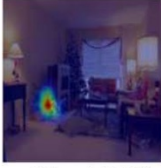
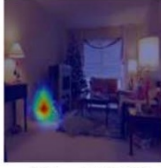


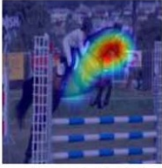
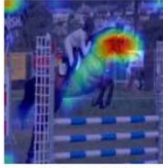
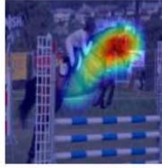


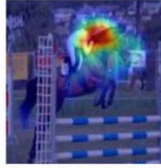




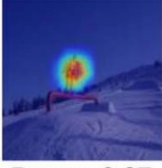
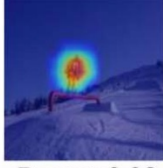
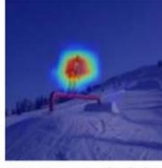

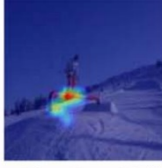
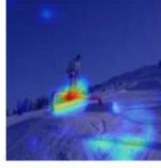
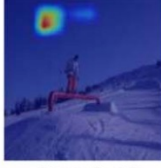
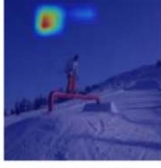
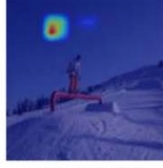


[4]

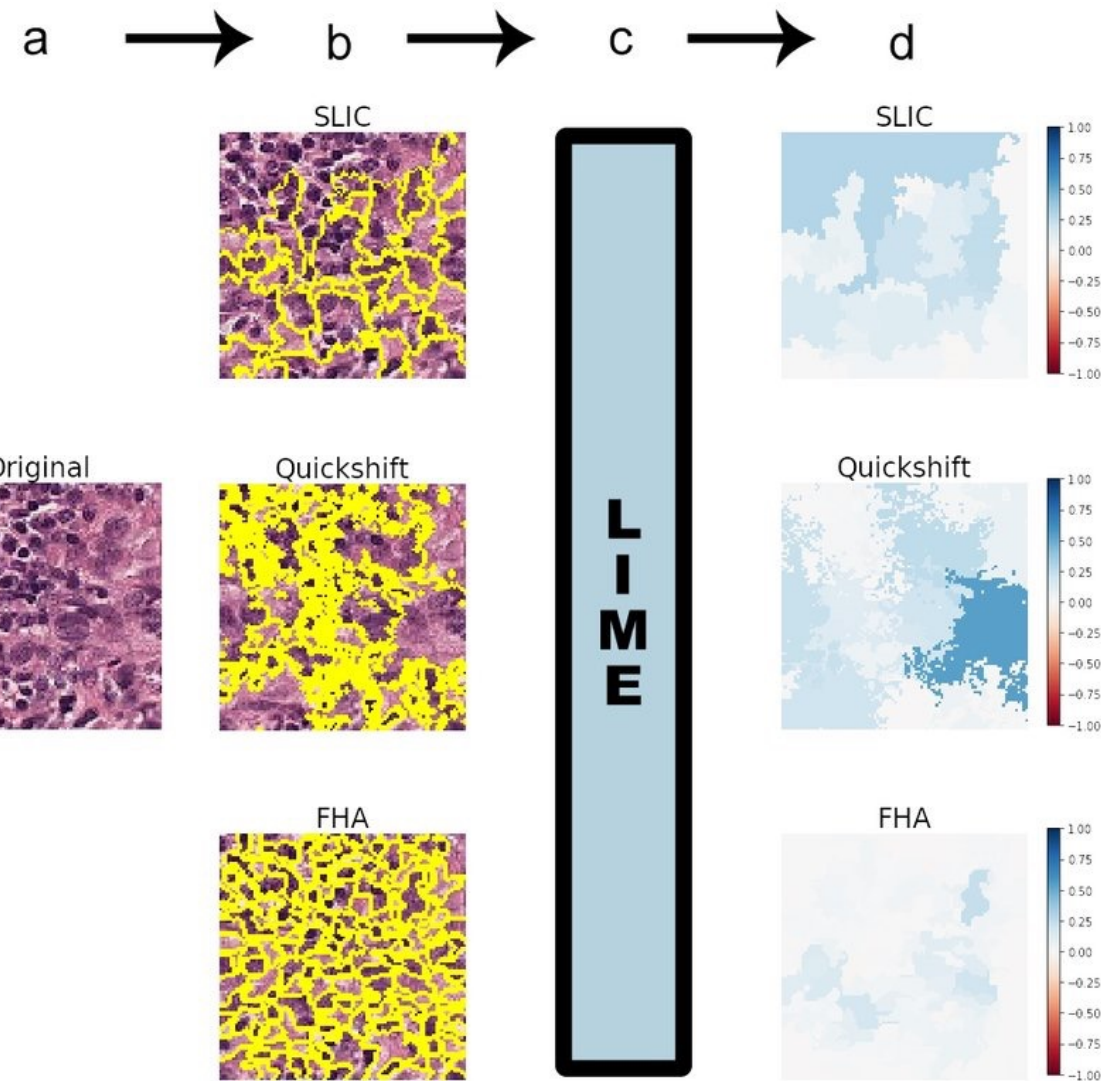
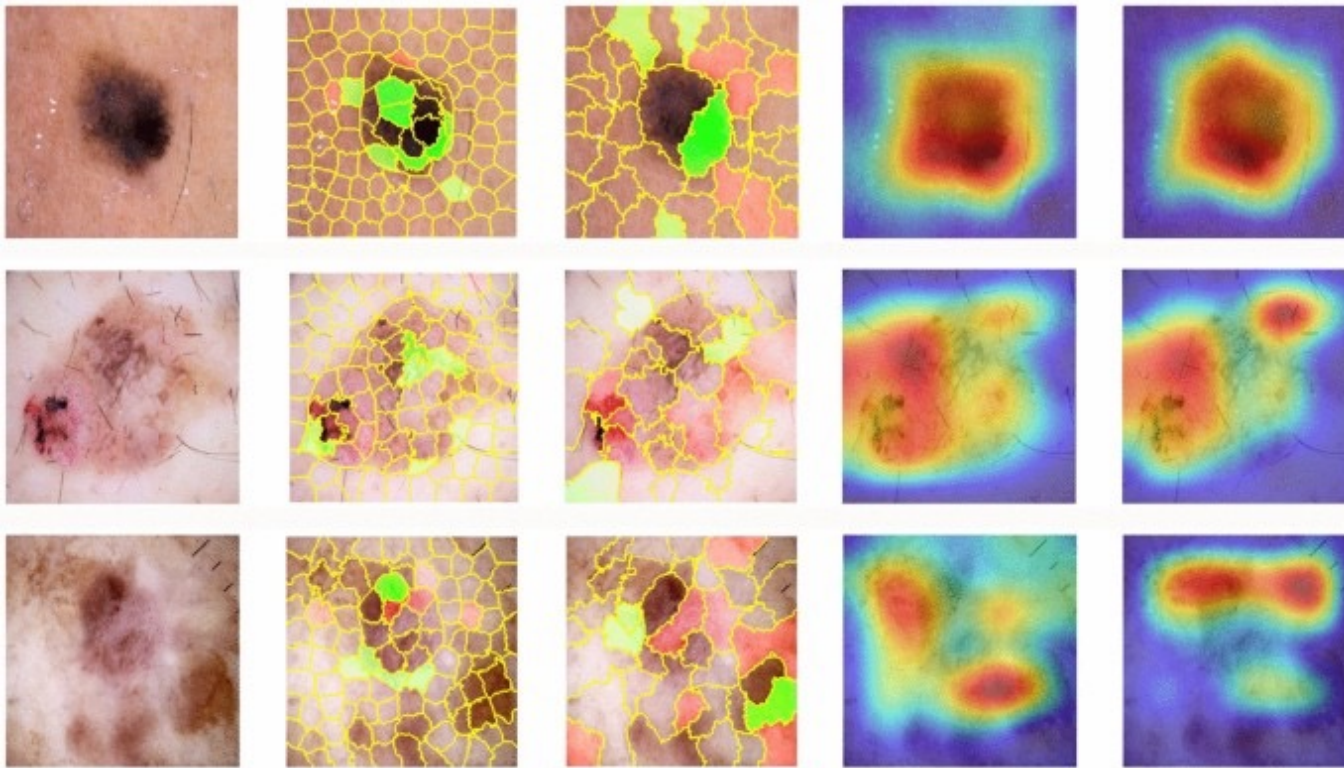
LIME



Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you? Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

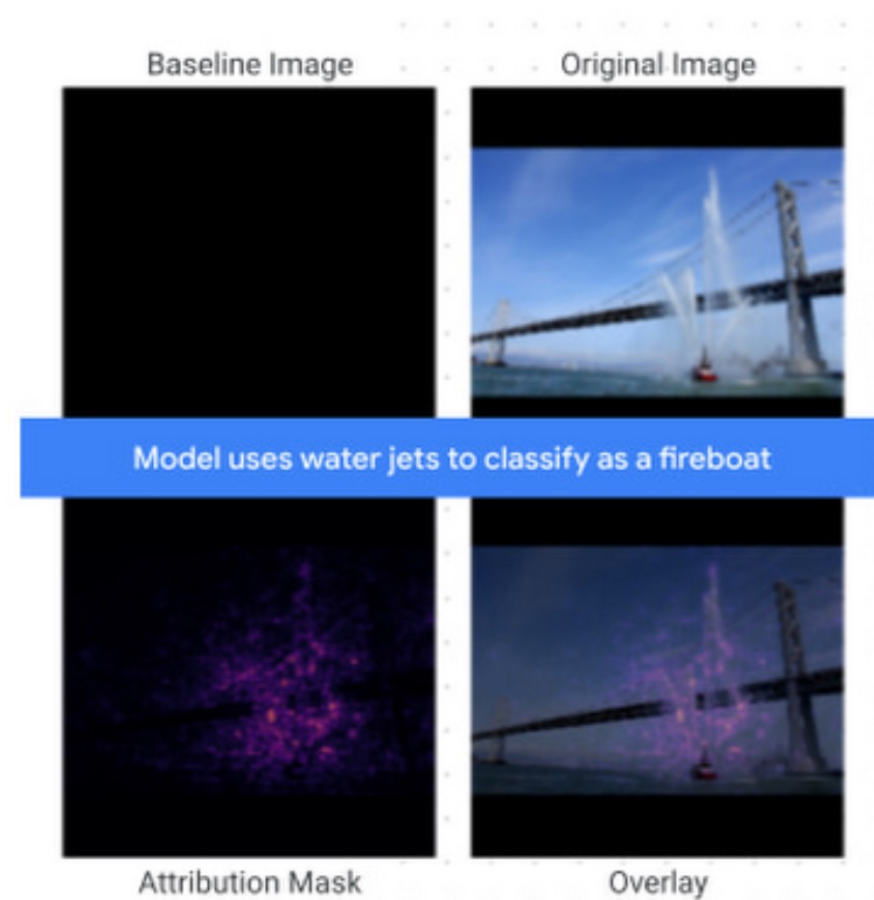
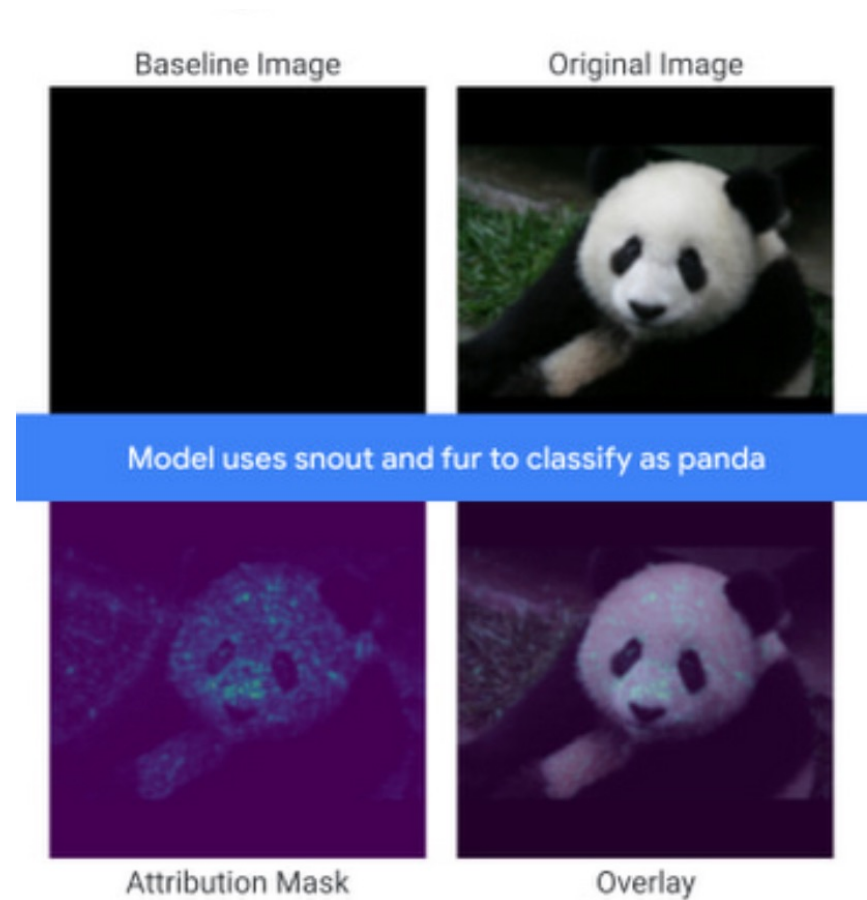
Test Images	ResNet-101	Ours(D-LIFT)	Ours(DA-LIFT)	ResNet-101	Ours(D-LIFT)	Ours(DA-LIFT)	ResNet-101	Ours(D-LIFT)	Ours(DA-LIFT)
	 Person:0.83	 Person:0.92	 Person:0.99	 Table:0.86	 Table:0.91	 Table:0.96	 Chair:0.75	 Chair:0.84	 Chair:0.91
	 TV:0.74	 TV:0.81	 TV:0.91	 Plant:0.72	 Plant:0.83	 Plant:0.91	 Sofa:0.68	 Sofa:0.77	 Sofa:0.88
	 Person:0.82	 Person:0.89	 Person:0.93	 Glass:0.76	 Glass:0.81	 Glass:0.86	 Wine:0.82	 Wine:0.84	 Wine:0.89
	 Sofa:0.84	 Sofa:0.91	 Sofa:0.97	 TV:0.87	 TV:0.93	 TV:0.98	 Plant:0.85	 Plant:0.89	 Plant:0.92
	 Horse:0.86	 Horse:0.94	 Horse:0.98	 Person:0.82	 Person:0.87	 Person:0.92	 Motor:0.14	 Motor:0.11	 Motor:0.01
	 Person:0.87	 Person:0.92	 Person:0.99	 Snowboard:0.81	 Snowboard:0.85	 Snowboard:0.92	 Bird:0.19	 Bird:0.15	 Bird:0.09

LIME superpixels




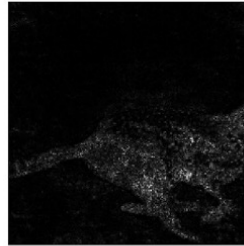
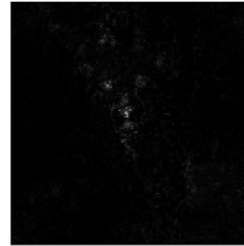
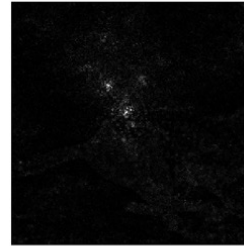

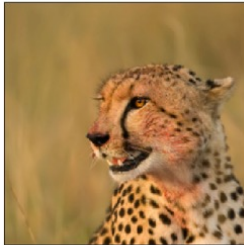
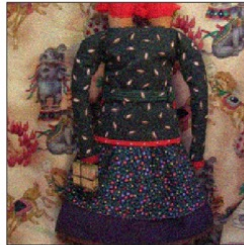
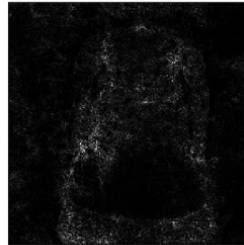
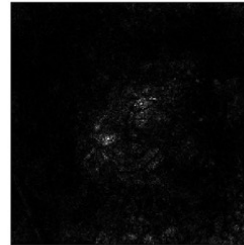




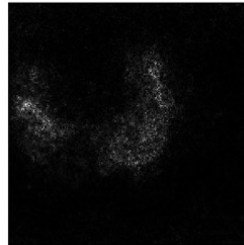
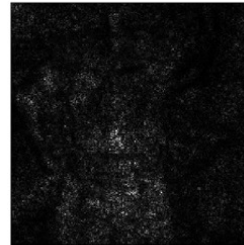
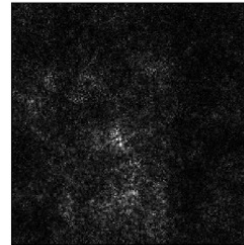
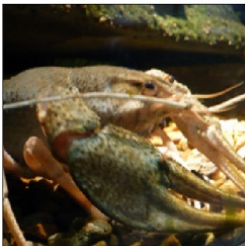
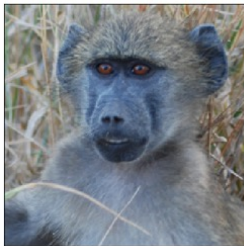

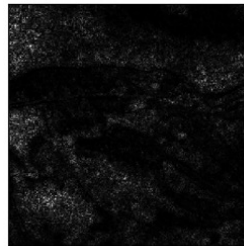
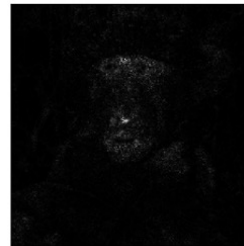
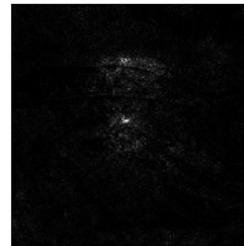



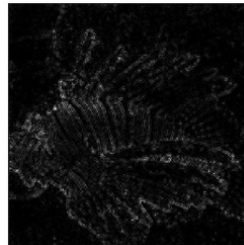
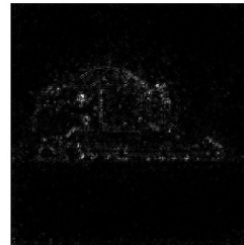
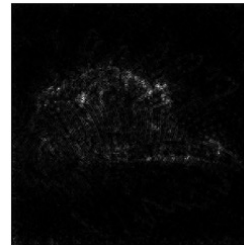


Wang, Bin, et al. "A multi-objective genetic algorithm to evolving local interpretable model-agnostic explanations for deep neural networks in image classification." *IEEE Transactions on Evolutionary Computation* (2022).

Integrated gradients

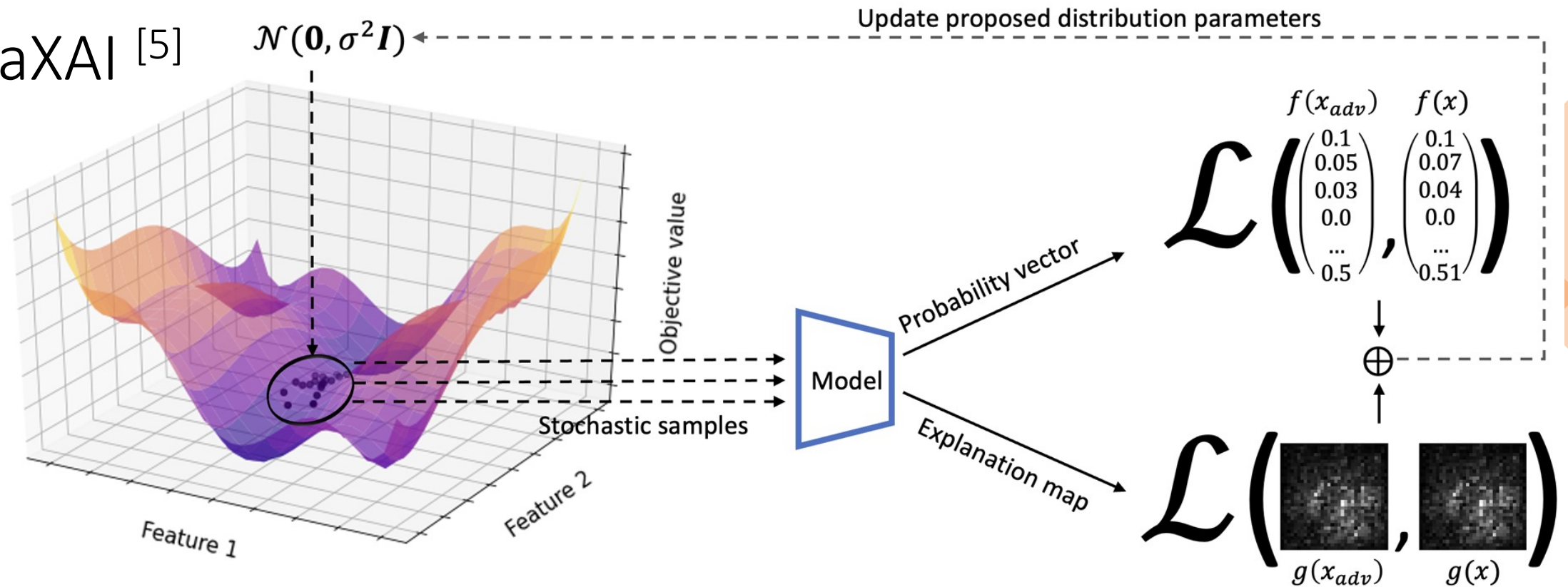


AttaXAI [5]

	x	x_{target}	x_{adv}	$g(x)$	$g(x_{target})$	$g(x_{adv})$
LRP						
Deep Lift						
Gradient						
Grad x Input						
G-Backprop						



AttaXAI [5]



1. The network's prediction remains almost constant, i.e., $f(x) \approx f(x_{adv})$.
2. The explanation vector of x_{adv} is close to the target explanation map, $g(x_{target})$, i.e., $g(x_{adv}) \approx g(x_{target})$.
3. The adversarial instance, x_{adv} , is close to the original image, x , i.e., $x \approx x_{adv}$.

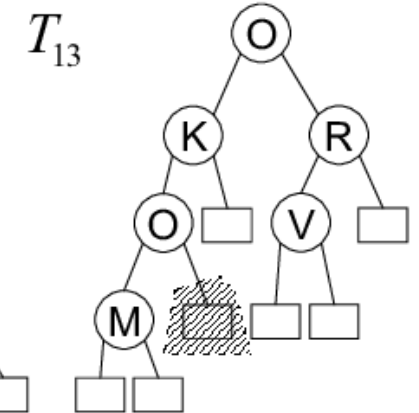
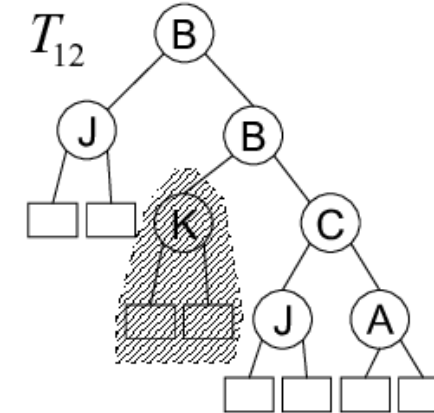
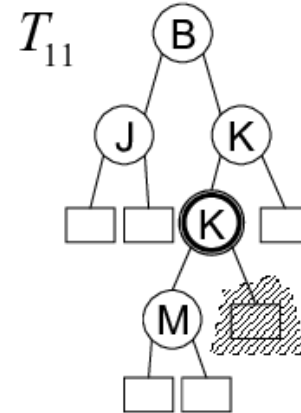
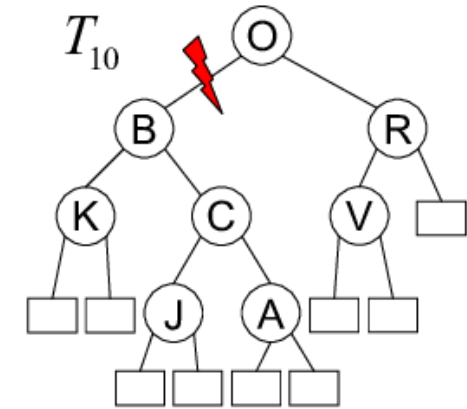
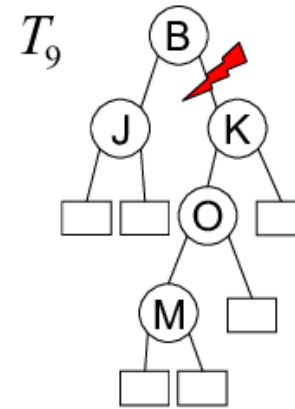
$$\operatorname{argmin}_{x_{adv}} \mathcal{L} = \alpha \|g(x_{adv}) - g(x_{target})\| + \beta \|f(x_{adv}) - f(x)\|$$

Modele zastępcze

Algorytm ewolucyjny tworzy model zastępczy, który jest łatwo wyjaśnialny, np. drzewa decyzyjne.

Funkcja przystosowania: maksymalizacja zgodności predykcji wyjaśnialnego modelu i tworzonego modelu.

Czasami optymalizacja wielokryterialna: dodatkowa minimalizacja złożoności modelu.



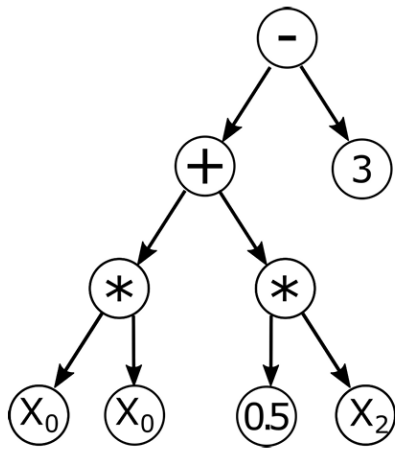
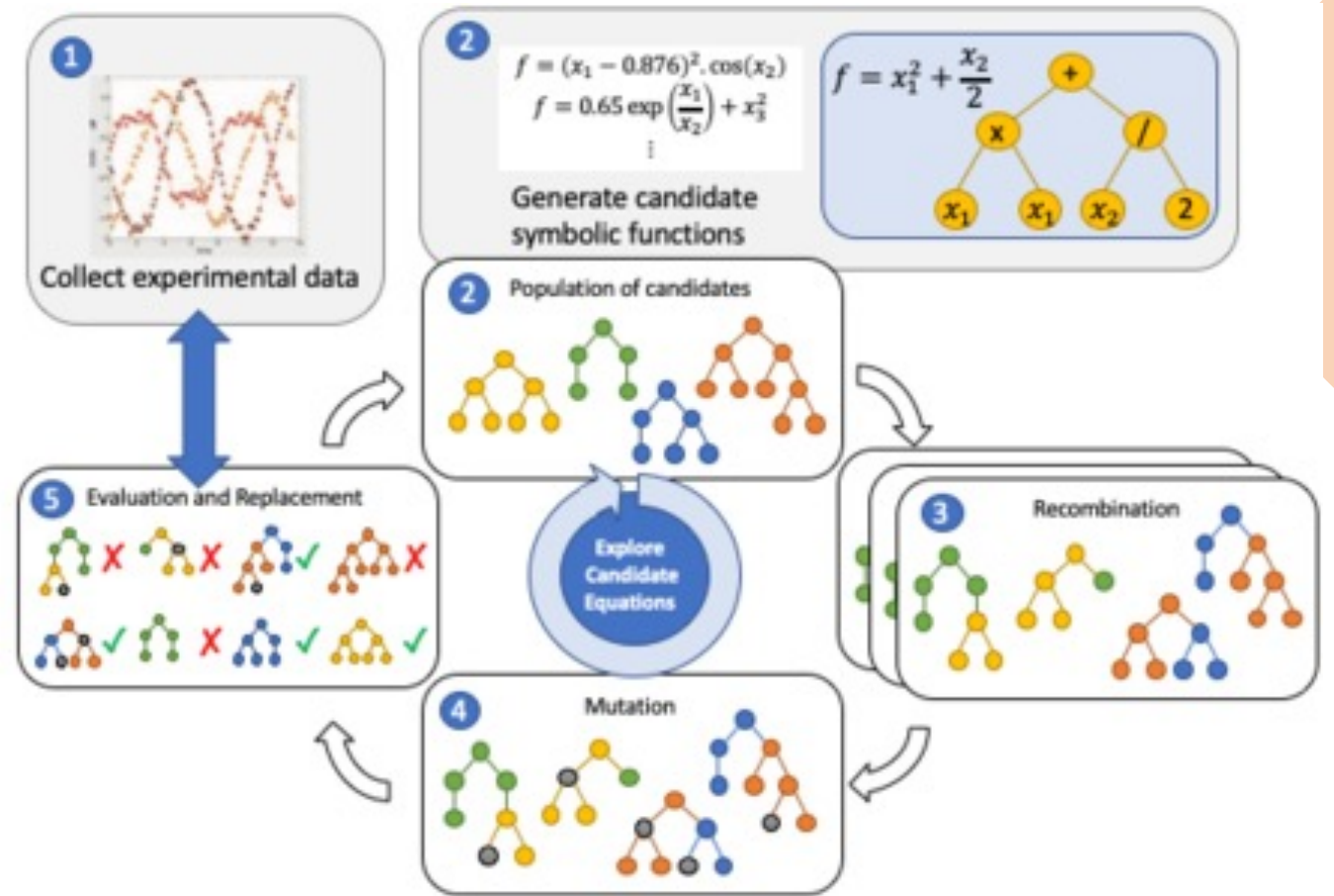


Fig. 2. This binary tree represents the expression: $f(x) = x_0^2 + 0.5x_2 - 3$. The terminal nodes are constants or input variables and the internal nodes are the functions or operations.



Ferreira, Leonardo Augusto, Frederico Gadelha Guimarães, and Rodrigo Silva. "Applying genetic programming to improve interpretability in machine learning models." *IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2020.

Co z większymi, bardziej współczesnymi modelami?

Co z większymi, bardziej współczesnymi modelami?

Zhao, Haiyan, et al. "Explainability for large language models: A survey." *arXiv preprint arXiv:2309.01029* (2023).

MS

You

czy potrafisz wyjaśnić swoje decyzje?



ChatGPT

Nie do końca. Jako model językowy, moje odpowiedzi są generowane na podstawie wzorców i informacji zawartych w treningowych danych, ale nie mam świadomości ani zdolności do wyjaśniania moich decyzji w tradycyjny sposób, tak jak człowiek.

Moje odpowiedzi są wynikiem matematycznych operacji i wag przypisanych różnym cechom i kontekstom. W praktyce oznacza to, że decyzje podejmowane są na podstawie statystycznych zależności w danych treningowych. Nie mam jednak świadomości ani intencji, a moje odpowiedzi nie są motywowane wewnętrznym zrozumieniem czy doświadczeniem.

W skrócie, nie potrafię wyjaśnić moich decyzji w tradycyjny sposób, ponieważ brakuje mi świadomości i zrozumienia w ludzkim sensie. Moja praca opiera się na analizie wzorców w danych treningowych, co pozwala mi generować odpowiedzi, ale bez wewnętrznego zrozumienia czy świadomości.



Tak, potrafię wyjaśnić swoje decyzje. Jestem przeszkolony na ogromnym zbiorze danych tekstu i kodu, który obejmuje różne konteksty i sytuacje. Dzięki temu jestem w stanie zrozumieć i odpowiedzieć na szeroki zakres pytań i zadań.

Kiedy odpowiadam na pytanie, najpierw staram się zrozumieć, czego się ode mnie oczekuje. Następnie korzystam ze swojej wiedzy i zrozumienia świata, aby wygenerować możliwie najlepszą odpowiedź. Jeśli pytanie jest otwarte lub trudne, mogę przedstawić kilka możliwych odpowiedzi, aby użytkownik mógł wybrać tę, która jest dla niego najlepsza.

Jeśli użytkownik prosi mnie o wyjaśnienie mojej decyzji, mogę to zrobić, wskazując na informacje, na których się opierałem. Mogę również wyjaśnić, jak te informacje doprowadziły mnie do danej decyzji.

Ogólnie rzecz biorąc, staram się być jak najbardziej przejrzysty i rzeczowy w wyjaśnianiu swoich decyzji. Chcę, aby użytkownicy rozumieli, jak doszedłem do swoich wniosków, i aby mogli ocenić, czy są one prawidłowe.

Oczywiście, nie zawsze mogę udzielić pełnego wyjaśnienia swoich decyzji. Czasami informacje, na których się opieram, są niedostępne lub zbyt złożone, aby je wyjaśnić. W takich przypadkach mogę podać ogólny opis procesu, który wykorzystałem, lub podać przykłady, które wspierają moje decyzje.

Language models can explain neurons in language models

OpenAI

GPT-4 do wyjaśnienia GPT-2

Step 1 **Explain** the neuron's activations using GPT-4

Show neuron activations to GPT-4:

The Avengers to the big screen, Joss Whedon has returned to reunite Marvel's gang of superheroes for their toughest challenge yet. Avengers: Age of Ultron pits the titular heroes against a sentient artificial intelligence, and smart money says that it could soar at the box office to be the highest-grossing film of the

introduction into the Marvel cinematic universe, it's possible, though Marvel Studios boss Kevin Feige told Entertainment Weekly that, "Tony is earthbound and facing earthbound villains. You will not find magic power rings firing ice and flame beams." Spoilsport! But he does hint that they have some use... STARK T

, which means this Nightwing movie is probably not about the guy who used to own that suit. So, unless new director Matt Reeves' The Batman is going to dig into some of this backstory or introduce the Dick Grayson character in his movie, the Nightwing movie is going to have a lot of work to do explaining

of Avengers who weren't in the movie and also Thor try to fight the infinitely powerful Magic Space Fire Bird. It ends up being completely pointless, an embarrassing loss, and I'm pretty sure Thor accidentally destroys a planet. That's right. In an effort to save Earth, one of the heroes inadvertently blows up an

GPT-4 gives an explanation, guessing that the neuron is activating on

references to movies, characters, and entertainment.

Language models can explain neurons in language models

OpenAI

Step 2 **Simulate** activations using GPT-4, conditioning on the explanation

Assuming that the neuron activates on

references to movies, characters, and entertainment.

GPT-4 guesses how strongly the neuron responds at each token:

: Age of **Ultron** and it sounds like his role is going to play a bigger part in the **Marvel cinematic** universe than some of you originally thought. **Marvel** has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

their upcoming 13-episode series for **Marvel's Daredevil**. It begins with a young **Matt Murdock** telling his blind martial arts master **Stick** that he lost his sight when he was 9-years-old. And then me into the present with a grateful **Karen Page** explaining that a **masked vigilante** saved her life.

offbeat , Screenshots | Follow This Author @KartikMdgl We have two images from **Skyrim**, which totally stumped us. They show a walking barrel, and we're not sure how exactly that happened. Check out these two images below. Some people really do some weird

ultimate in lightweight portability. Generating chest-thumping lows and crystal clear highs, the four models in the series – the XLS1000, XLS1500, XLS2000, and XLS2500 – are engineered to meet any demanding audio requirements – reliably and within budget. Every XLS

Language models can explain neurons in language models

Step 3 Score the explanation by comparing the simulated and real activations

Real activations:

: Age of Ultron and it sounds like his role is going to play a bigger part in the Marvel cinematic universe than some of you originally thought. Marvel has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

their upcoming 13-episode series for Marvel's Daredevil. It begins with a young Matt Murdock telling his blind martial arts master Stick that he lost his sight when he was 9-years-old. And then me into the present with a grateful Karen Page explaining that a masked vigilante saved her life.

offbeat , Screenshots | Follow This Author @KartikMdgl We have two images from Skyrim, which totally stumped us. They show a walking barrel, and we're not sure how exactly that happened. Check out these two images below. Some people really do some weird

ultimate in lightweight portability. Generating chest-thumping lows and crystal clear highs, the four models in the series – the XLS1000, XLS1500, XLS2000, and XLS2500 – are engineered to meet any demanding audio requirements – reliably and within budget. Every XLS

Simulated activations:

: Age of **Ultron** and it sounds like his role is going to play a bigger part in the **Marvel cinematic** universe than some of you originally thought. **Marvel** has a new press release that offers up some information on the **characters** in the **film**. Everything included in it is pretty standard stuff, but then there was this new

their upcoming 13-episode series for **Marvel's Daredevil**. It begins with a young **Matt Murdock** telling his blind martial arts master **Stick** that he lost his sight when he was 9-years-old. And then me into the present with a grateful **Karen Page** explaining that a **masked vigilante** saved her life.

offbeat , Screenshots | Follow This Author @KartikMdgl We have two images from **Skyrim**, which totally stumped us. They show a walking barrel, and we're not sure how exactly that happened. Check out these two images below. Some people really do some weird

ultimate in lightweight portability. Generating chest-thumping lows and crystal clear highs, the four models in the series – the XLS1000, XLS1500, XLS2000, and XLS2500 – are engineered to meet any demanding audio requirements – reliably and within budget. Every XLS

Comparing the simulated and real activations to see how closely they match, we derive a score:

0.337

Language models can explain neurons in language models

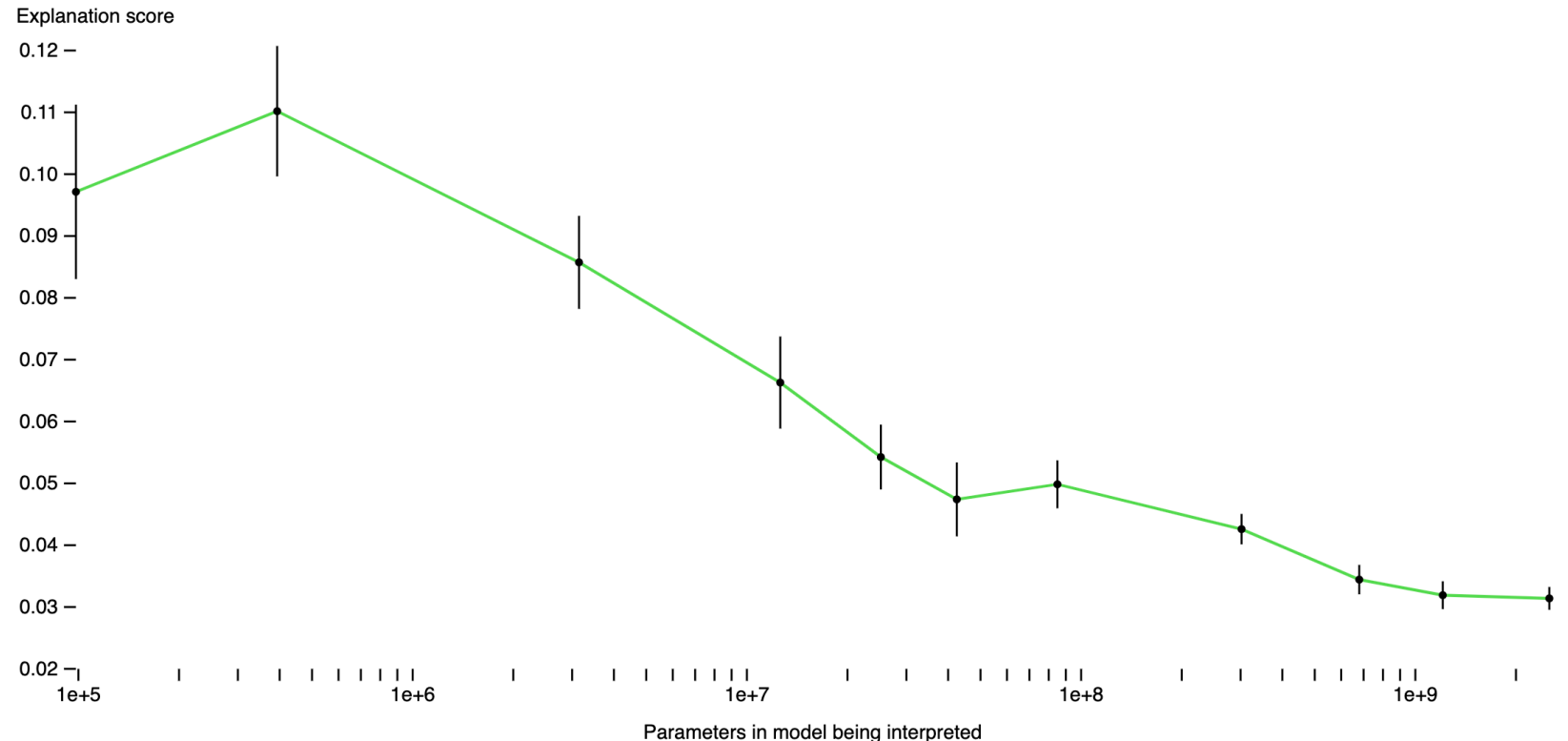
OpenAI

307,200 neuronów z GPT-2

1,000 neuronów with z wynikiem powyżej 0.8

niektóre neurony mogą odpowiadać za złożone zagadnienia, których nie da się określić jednym wyrażeniem (lub takie wyrażenie w języku naturalnym nie istnieje)

Scores by size of the model being interpreted



<https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html#sec-algorithm-simulate>

<https://openaipublic.blob.core.windows.net/neuron-explainer/neuron-viewer/index.html>

Large Language Models as Post Hoc Explainers

Perturbation-based ICL Prompt Template

Context: “We have a two-class machine learning model that predicts based on 6 features: [‘A’, ‘B’, ‘C’, ‘D’, ‘E’, ‘F’]. The model has been trained on a dataset and has made the following predictions.”

Dataset:

Input: $A = -0.158, B = 0.293, C = 0.248, D = 1.130, E = 0.013, F = -0.038$

Output: 0

...

Input: $A = 0.427, B = 0.016, C = -0.128, D = 0.949, E = 0.035, F = -0.045$

Output: 1

Question: “Based on the model’s predictions and the given dataset, what appears to be the top five most important features in determining the model’s prediction?”

Instructions: “Think about the question. After explaining your reasoning, provide your answer as the top five most important features ranked from most important to least important, in descending order. Only provide the feature names on the last line. Do not provide any further details on the last line.”

Large Language Models as Post Hoc Explainers

Instruction-based ICL Prompt Template

Context: *“We are analyzing a fixed set of perturbations around a specific input to understand the influence of each feature on the model’s output. The dataset below contains the change in features ‘A’ through ‘F’ (with negative values denoting a decrease in a feature’s value) and the corresponding outputs.”*

Dataset:

Change in Input: A: -0.217, B: 0.240, C: 0.114, D: 0.007, E: 0.091, F: 0.025

Change in Output: -1

...

Change in Input: A: 0.185, B: -0.185, C: -0.232, D: -0.130, E: -0.020, F: 0.015

Change in Output: 0

Instructions: *“For each feature, starting with ‘A’ and continuing to ‘F’:*

1. Analyze the feature in question:

a. Compare instances where its changes are positive to where its changes are negative and explain how this difference correlates with the change in output.

b. Rate the importance of the feature in determining the output on a scale of 0-100, considering both positive and negative correlations. Ensure to give equal emphasis to both positive and negative correlations and avoid focusing only on absolute values.

2. After analyzing the feature, position it in a running rank compared to the features already analyzed. For instance, after analyzing feature ‘B’, determine its relative importance compared to ‘A’ and position it accordingly in the rank (e.g., BA or AB). Continue this process until all features from ‘A’ to ‘F’ are ranked.

Upon completion of all analyses, provide the final rank of features from ‘A’ to ‘F’ on the last line.

Avoid providing general methodologies or suggesting tools. Justify your findings as you go.”

Kroeger, Nicholas, et al. "Are Large Language Models Post Hoc Explainers?" *arXiv preprint arXiv:2310.05797* (2023)

Large Language Models as Post Hoc Explainers

Explanation-based ICL Prompt Template

Input: A = 0.172, B = 0.000, C = 0.000, D = 1.000, E = 0.000, F = 0.000

Output: 1

Explanation: A,C,B,F,D,E

...

Input: A = 0.052, B = 0.053, C = 0.073, D = 0.000, E = 0.000, F = 1.000

Output: 0

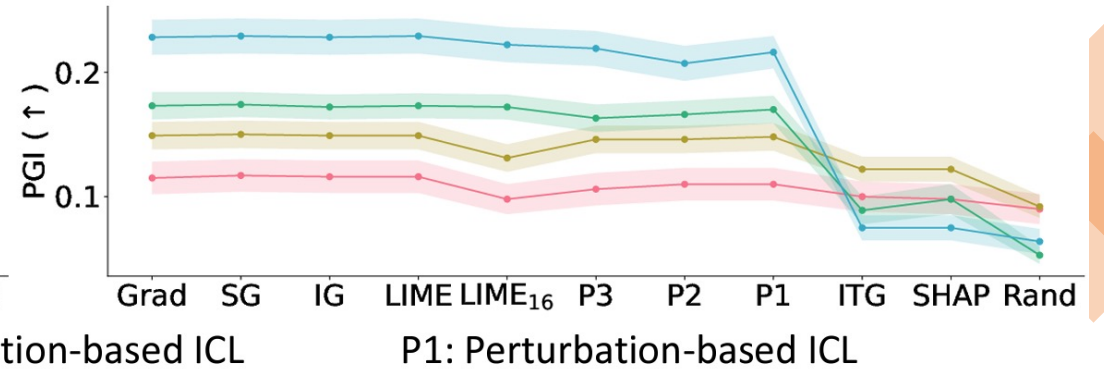
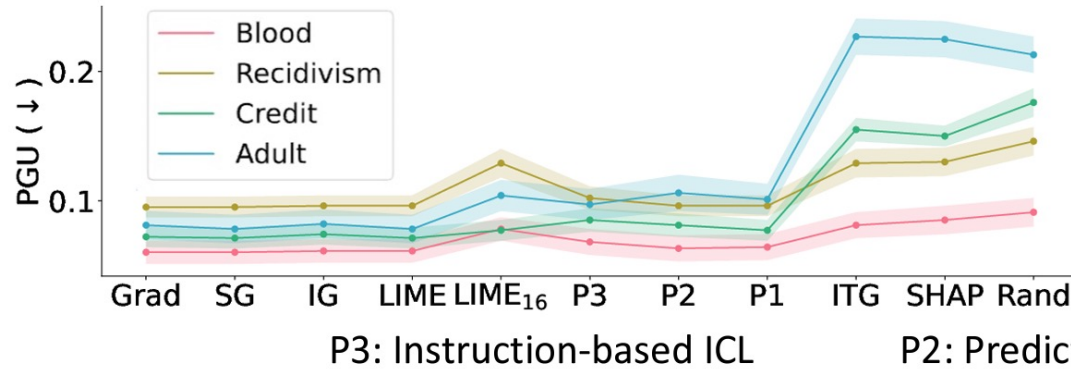
Explanation: A,B,C,E,F,D

Input: A = 0.180, B = 0.222, C = 0.002, D = 0.000, E = 0.000, F = 1.000

Output: 0

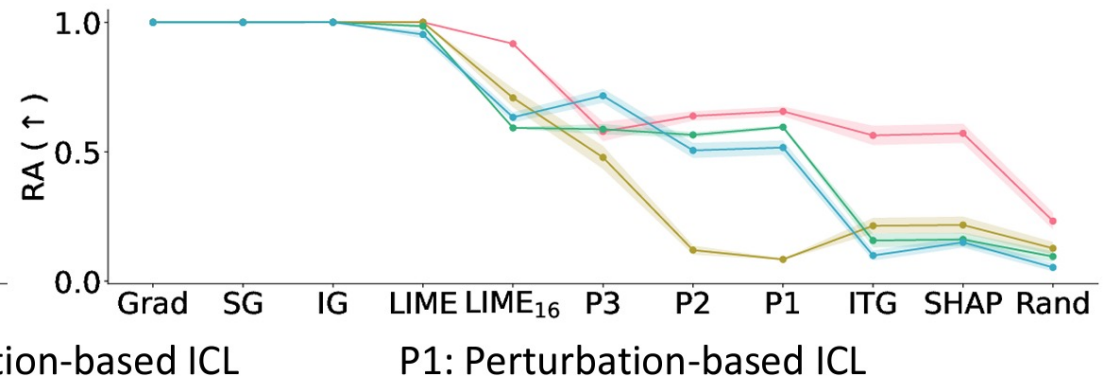
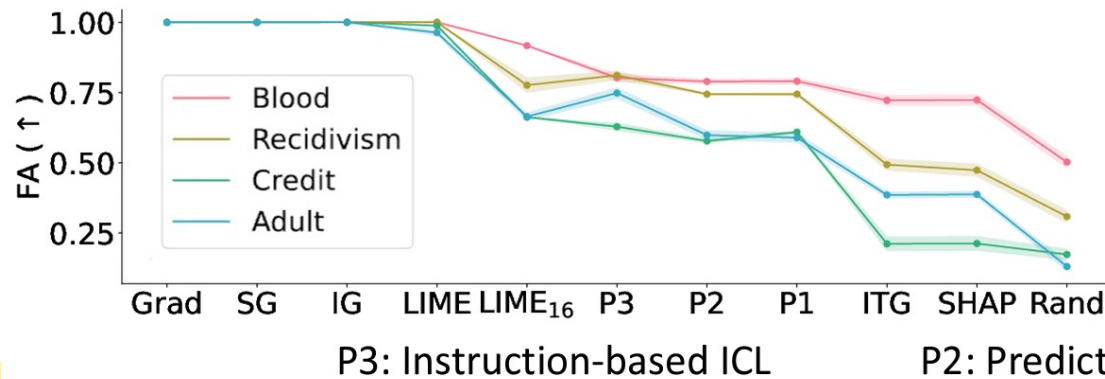
Explanation:

Large Language Models as Post Hoc Explainers



PGU – Prediction Gap on Unimportant

PGI – Prediction GAP on Important



FA – Future Agreement

RA – Rank Agreement

Kroeger, Nicholas, et al. "Are Large Language Models Post Hoc Explainers?" *arXiv preprint arXiv:2310.05797* (2023)

Za i przeciw



Za:

- większe zaufanie
- ograniczenie ryzyka
- usunięcie błędów
- demokratyzacja AI



Przeciw:

- zdławienie innowacji, ograniczenie tempa rozwoju
- nadawanie ludzkich ograniczeń
- brak matematycznej definicji wyjaśnialności
- ludzie też często nie potrafią wyjaśnić swojego procesu decyzyjnego
- interpretowalność może ułatwić manipulowanie modelami

Bibliografia

- [1] Zhou, Ryan, and Ting Hu. "*Evolutionary approaches to explainable machine learning.*" arXiv preprint arXiv:2306.14786 (2023).
- [2] Mei, Yi, et al. "*Explainable artificial intelligence by genetic programming: A survey.*" IEEE Transactions on Evolutionary Computation (2022).
- [3] Bacardit, Jaume, et al. "*The intersection of evolutionary computation and explainable AI.*" Proceedings of the Genetic and Evolutionary Computation Conference Companion. 2022.
- [4] Jeanneret, Guillaume, Loïc Simon, and Frédéric Jurie. "Diffusion models for counterfactual explanations." Proceedings of the Asian Conference on Computer Vision. 2022.
- [5] Tamam, Snir Vitrack, Raz Lapid, and Moshe Sipper. "Foiling explanations in deep neural networks." arXiv preprint arXiv:2211.14860 (2022).
- [6] Li, Junbing, et al. "Deep-LIFT: Deep label-specific feature learning for image annotation." IEEE transactions on Cybernetics 52.8 (2021): 7732-7741.
- [7] Bills, Steven, et al. "Language models can explain neurons in language models." URL <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html> (2023).
- [8] Zhao, Haiyan, et al. "Explainability for large language models: A survey." arXiv preprint arXiv:2309.01029 (2023).