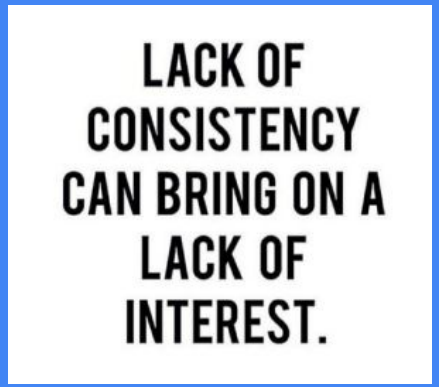
Agree to Disagree: When Deep Learning Models With Identical Architectures Produce Distinct Explanations

“We conclude that current trends in model explanation are not sufficient to mitigate the risks of deploying models in real life healthcare applications.”
Paper Authors



Jak poprzeć konkluzję?

1. Sposoby zaburzania modeli
2. Miara spójności wyjaśnień zaburzeń modelu
3. Miara rozróżnialności pary wyjaśnień
4. Eksperymenty
5. Wnioski

1. Sposoby zaburzania modeli

Micro level:
- architektura modelu
- zbiór danych
- loss

Macro level:
+ random seed
+ kolejność danych
+ dropout

2. Miara spójności wyjaśnień zaburzeń modelu

Gdybym miał miarę pozwalającą określić jak bardzo wyjaśnienia dwóch zaburzeń modelu się różnią, to pewnie wziąłbym średnią tych miar dla różnych par a, b zaburzeń danego modelu.

$$C = 1 - \frac{\sum_{(a,b)} S_{(a,b)}}{\alpha}$$

C - miara spójności
S(a, b) - ta miara
alfa - ilość porównanych par

3. Miara rozróżnialności pary wyjaśnień

Jakie własności powinna mieć ta miara?
- niezależna od modelu
- niezależna od metody wyjaśnienia
- dla identycznych modeli powinna zwracać 0

$$S_{(a,b)} = \mathbb{E}_i\left[D\left(E(Y^a(x_i)), E(Y^b(x_i))\right)\right]$$

(w skrócie: wartość oczekiwana odległości od siebie wyjaśnień dwóch różnych modeli na tym samym przykładzie)

3.1 Jakich miar próbowano użyć?

1. Testy statystyczne (czy dwa zbiory próbek pochodzą z tego samego rozkładu)
2. Mierzenie podobieństwa dwóch rozkładów
3. Skuteczność modelu próbującego określić od którego z dwóch rozpatrywanych modeli pochodzi wyjaśnienie.

S(a,b) na wartościach SHAP modelu CNN4 wytrenowanego na MNIST z różnym random seed

if x > 0,5:
  S(a, b) = 2 * (x - 0.5)
else :
  S(a, b) = 2 * (0.5 - x)

M1 Seed	M2 Seed	JSD	KS	Wilcoxon	LR
1	1	0	0	0	0.5
1	12303	0.8062	0.9744	7.877e+09	0.973
1	15135	0.8012	0.9690	1.738e+10	0.978
1	16959	0.7346	0.8890	2.464e+11	0.975
12303	12303	0	0	0	0.5
12303	15135	0.8228	0.9913	4.350e+08	0.979
12303	16959	0.7900	0.9567	3.316e+10	0.974
15135	15135	0	0	0	0.5
15135	16959	0.8122	0.9810	6.611e+09	0.975

a

b

S(a, b) = x

3.1.1 Która metoda wygrała?

$$2 * |M_{(a,b)} - 0.5|$$

M(a, b) - accuracy modelu regresji liniowej klasyfikującego, czy dane wyjaśnienie pochodzi od modelu a, czy od modelu b

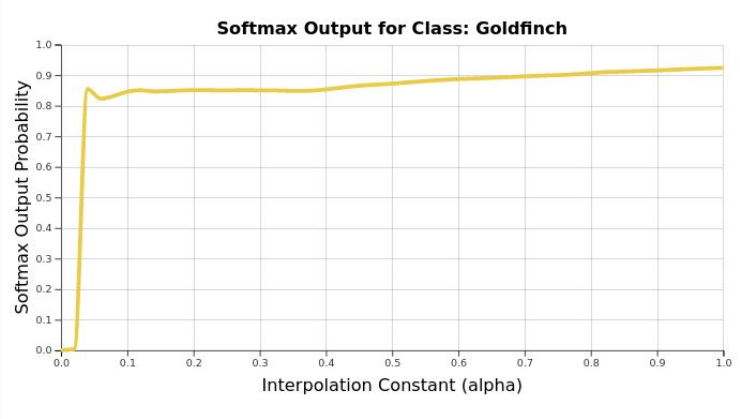
Zbiory, modele i metody wyjaśnienia

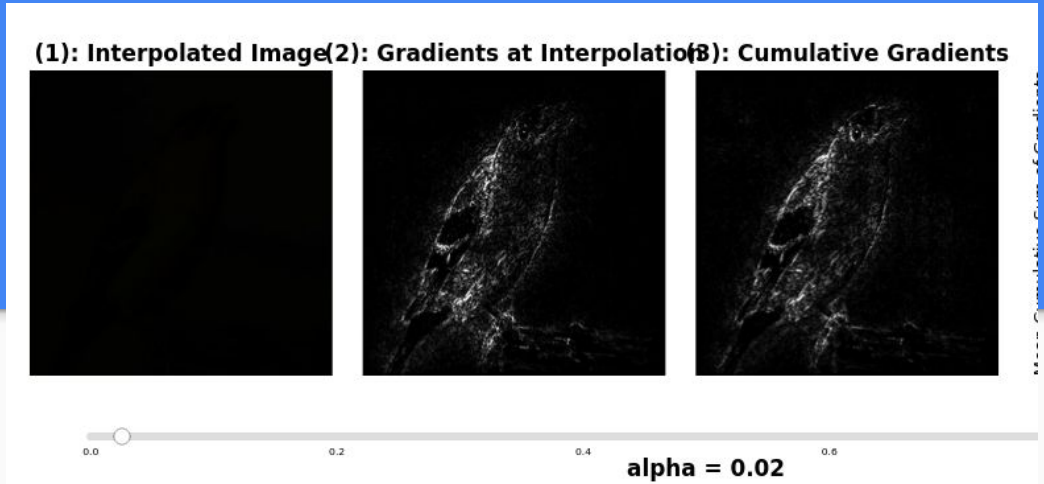
- Modele trenowane na MNIST:
  - MLP
  - Small-CNN, CNN, GaborNet
  - ResNet18
  - SVM z funkcją jądrową RBF
  - Ensemble 10 sub-ResNet, Hyperensemble
- Modele trenowane na zdjęciach rentgenowskich (MIMIC-CXR-JPG):
  - Densenet-121
  - Ensemble 3 Densenet-121
- Wyjaśnienia: SHAP i Integrated Gradients

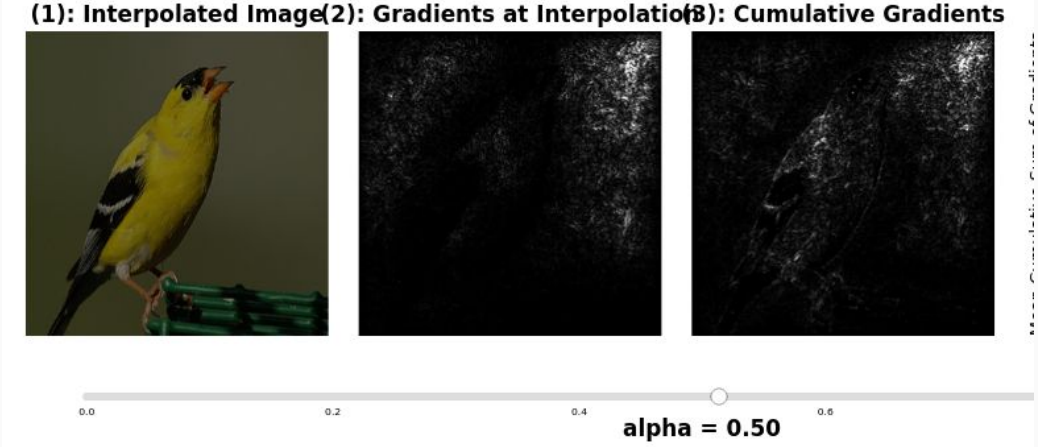
Integrated Gradients - przypomnienie

- Metoda stosowana do analizy DNN
- Bazuje na sumowaniu gradientów dla różnych wejść sieci
- Każde wejście jest pewną średnią ważoną z wejścia analizowanego i szumu

Integrated Gradients dla klasyfikacji obrazów







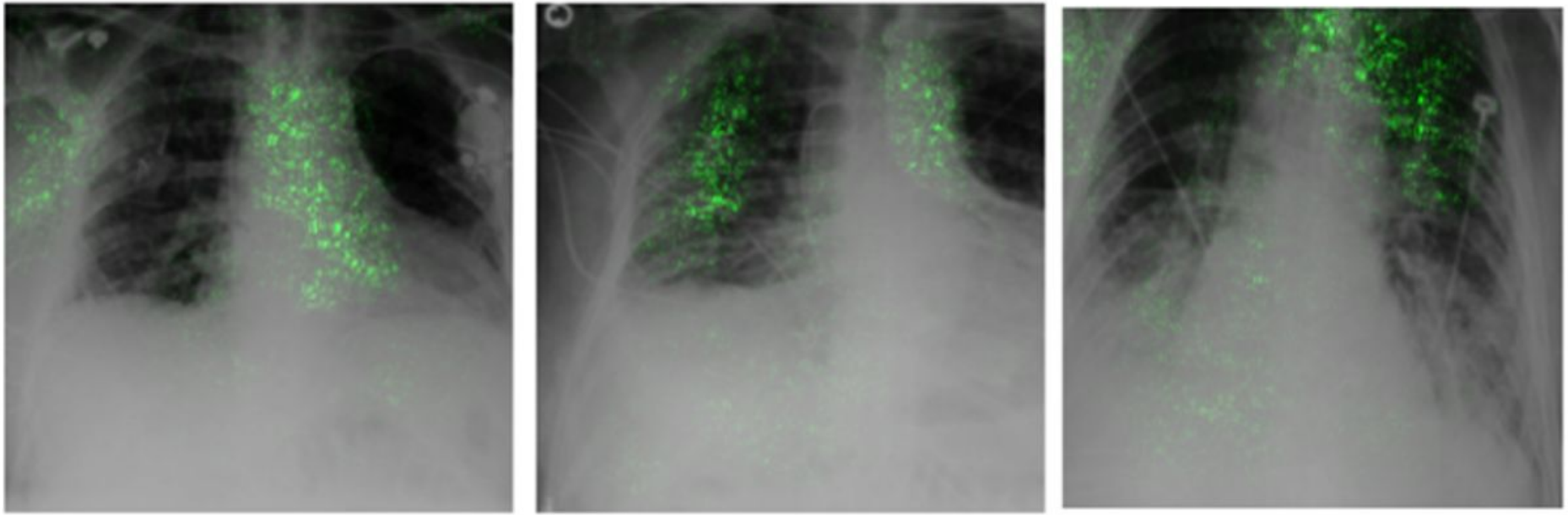
Gradienty dla różnych wag przenikania obrazów

Rodzaje różnic w wyjaśnieniach

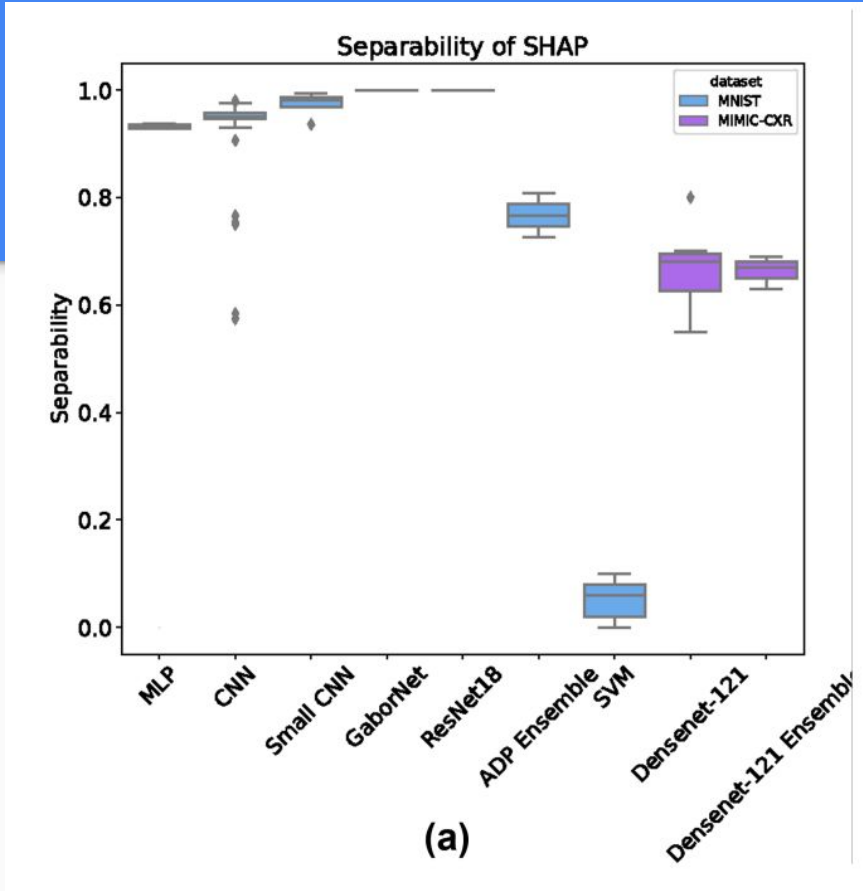
Różnice wyjaśnień:

- Różnice w wyjaśnianiu elementów istotnych dla klasyfikacji
- Różnice w wyjaśnianiu szumów

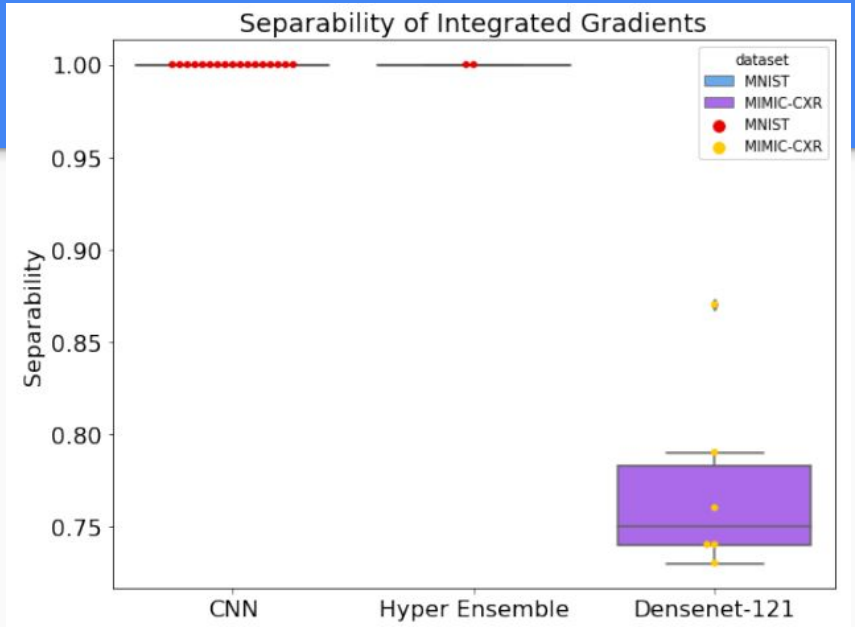
Wyjaśnianie klasyfikacji modelu metodą Shap



Trzy losowe zdjęcia ze zbioru MIMIC-CXR-JPG z wyjaśnieniem modelu Densenet121


Separability of SHAP
dataset
MNIST
MIMIC-CXR
Separability
1.0
0.8
0.6
0.4
0.2
0.0
MLP
CNN
Small CNN
GaborNet
ResNet18
ADP Ensemble
SVM
Densenet-121
Densenet-121 Ensembl
(a)

MNIST


Separability of Integrated Gradients
dataset
MNIST
MIMIC-CXR
MNIST
MIMIC-CXR
Separability
1.00
0.95
0.90
0.85
0.80
0.75
CNN
Hyper Ensemble
Densenet-121

MNIST

Model Architecture	Dataset	α	Consistency				Accuracy
			Overall	Shuffle	Random Seed	Dropout	
MLP	MNIST	6	0.0668	0.062	0.066	0.0687	98.125 ± 0.9270
SVM	MNIST	10	0.9444	0.96	0.94	n/a	94.0556 ± 0.6213
Small-CNN	MNIST	6	0.0252	0.018	0.06	0.034	98.3486 ± 0.0360
GaborNet	MNIST	12	0	0	0	0	95.038 ± 0.2824
ResNet18	MNIST	10	0	0	0	n/a	99.425 ± 0.0626
ADP Ensemble	MNIST	6	0.2193	0.192	0.233	n/a	99.083 ± 0.2514
CNN	MNIST	12	0.0652	0.052	0.0564	0.0914	98.9976 ± 0.5756
Densenet-121	MIMIC-CXR	6	0.3329	n/a	0.3329	n/a	75.6723 ± 1.1379
Densenet-121 Ensemble	MIMIC-CXR	4	0.3367	n/a	0.3667	n/a	80.8 ± 0.7483
CNN (IG)	MNIST	12	0	0	0	0	98.9976 ± 0.5756
Hyperensemble (IG)	MNIST	2	0	n/a	0	n/a	99.32 ± 0.0082
Densenet-121 (IG)	MIMIC-CXR	6	0.168	0.115	0.2033	n/a	75.6723 ± 1.1379

Wnioski

- Sieci neuronowe są bardzo podatne na zmiany wyjaśnialności
- Nawet niepozorna zmiana kolejności danych, dropout czy inicjalizacja wag powoduje duże zmiany w wyjaśnialności modeli
- Problematyka leży w stochastycznej DNN
- Sieć SVM prawdopodobnie dzięki jądru RBF wykazuje dużą stałość wyjaśnialności