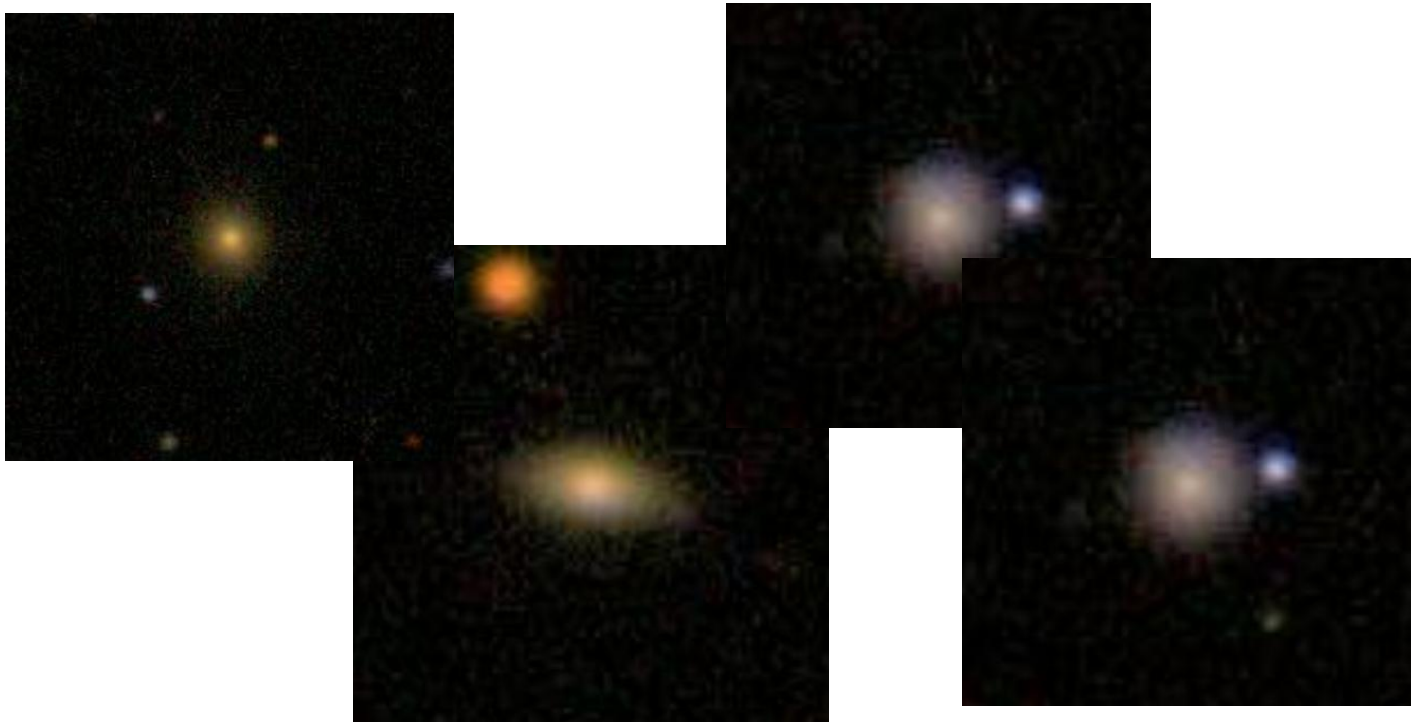
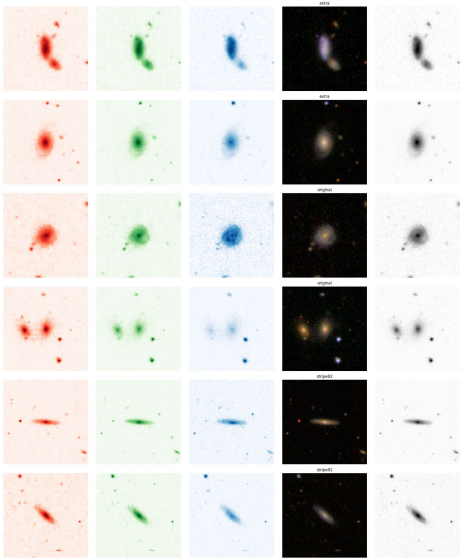
Galaxy Zoo 2 - galaxy clusterization

Autorzy:
Paweł Pozorski
Michał Pytel
Michał Matuszyk

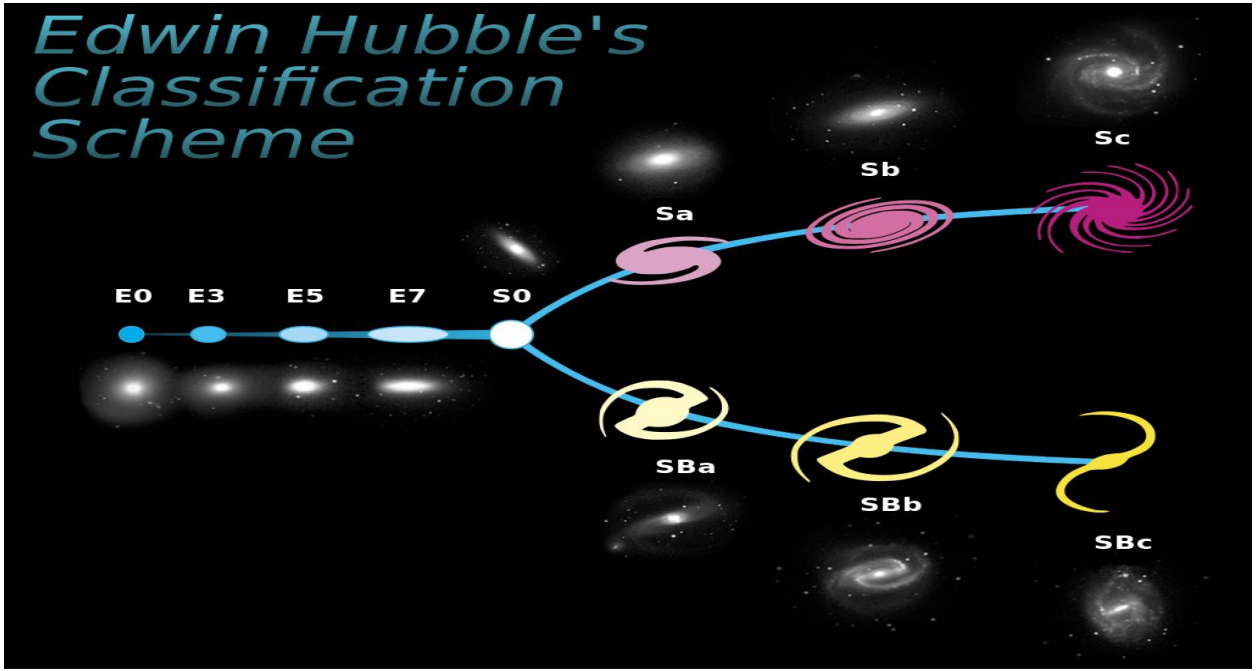
Troszkę o danych

- Źródło: www.kaggle.com/datasets/jaimetrickz/galaxy-zoo-2-images
- Ponad 300k kolorowych zdjęć (355990)
- (424x424x3)





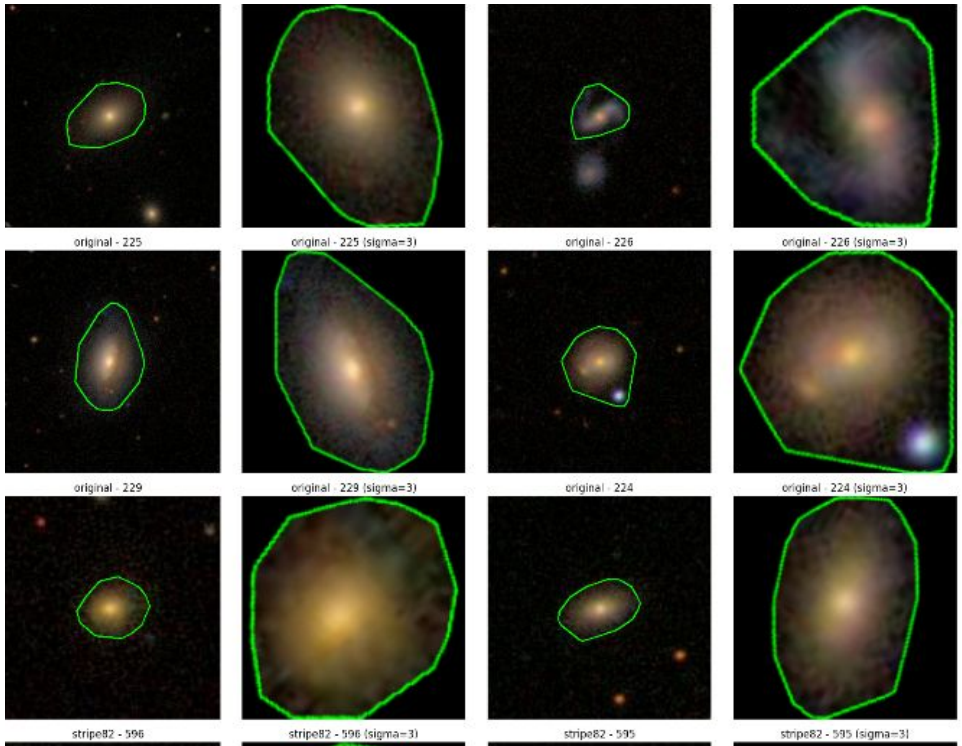
Cel biznesowy



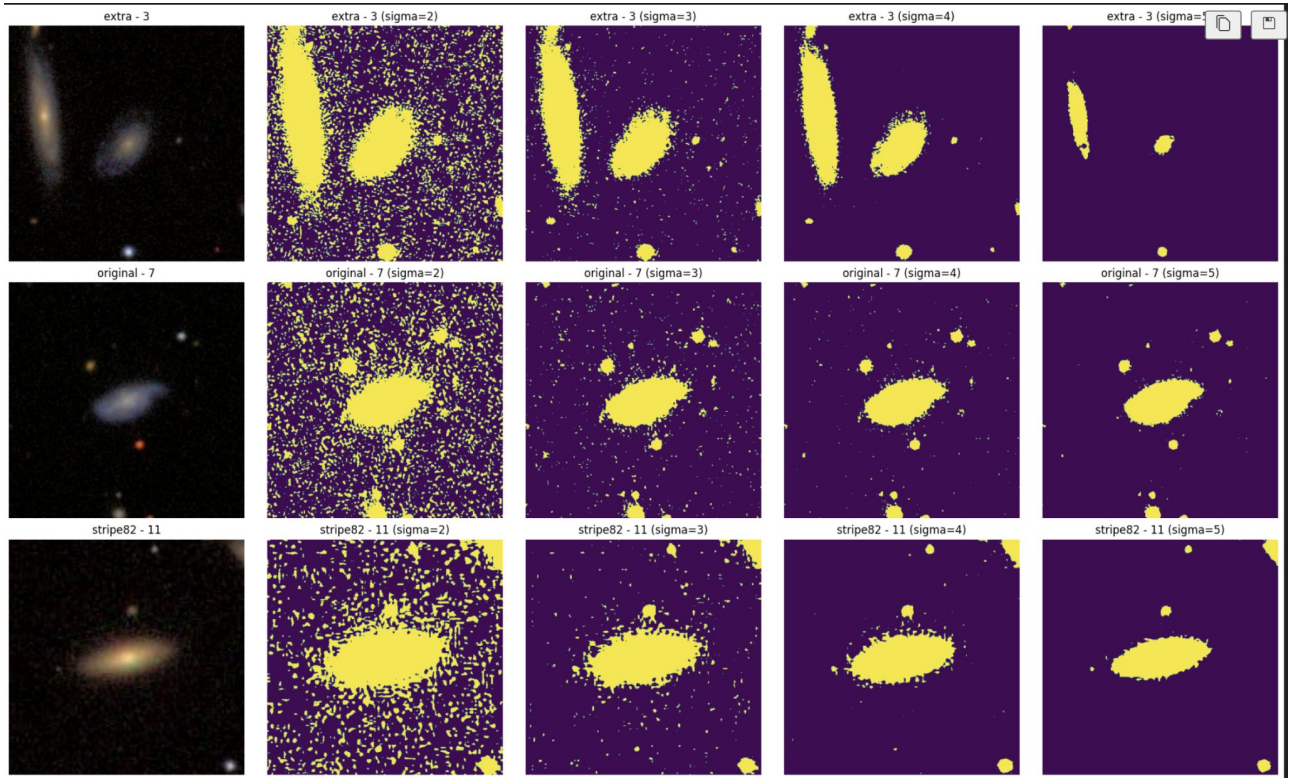
Sprawdzenie czy da się wykazanie zachodzenia sekwencji Hubble’a

Źródło: https://en.wikipedia.org/wiki/Hubble_sequence#/media/File:Hubble_Tuning_Fork_diagram.svg

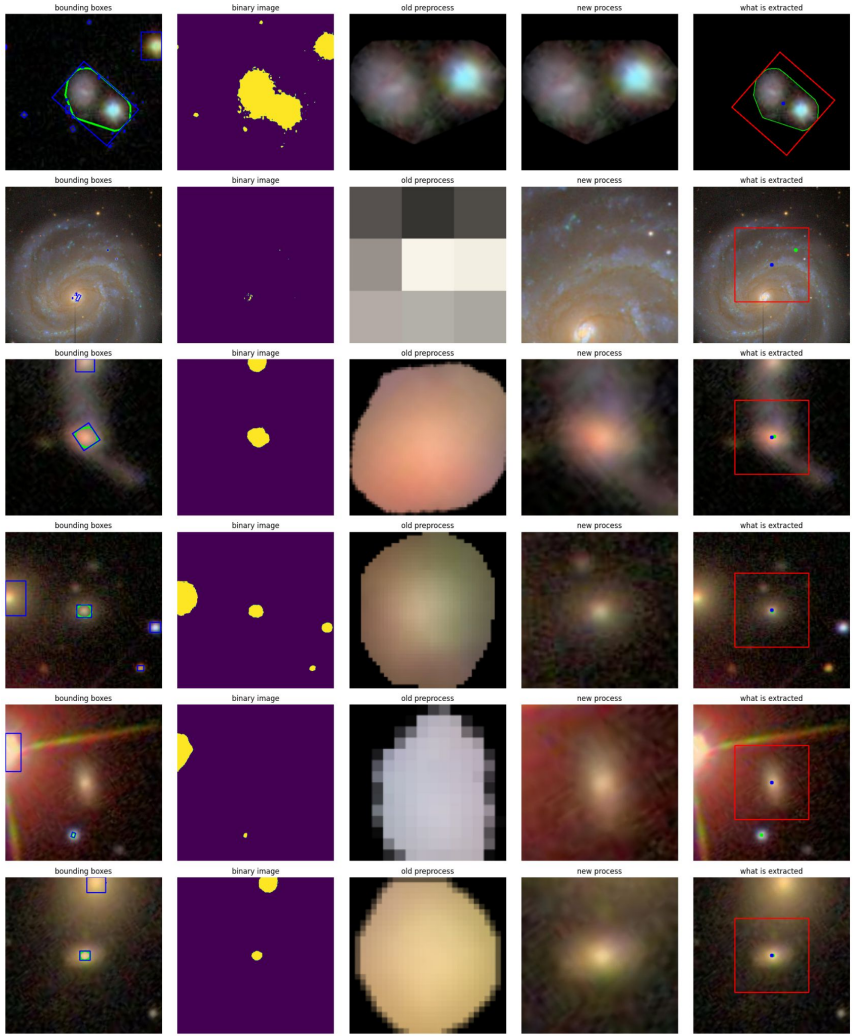
Proces oczyszczania danych



Proces oczyszczania danych



Proces oczyszczania danych



Proces oczyszczania danych

Empirycznie - proces czyszczenia zdjęcia Z0 składa się z kolejnych kroków:

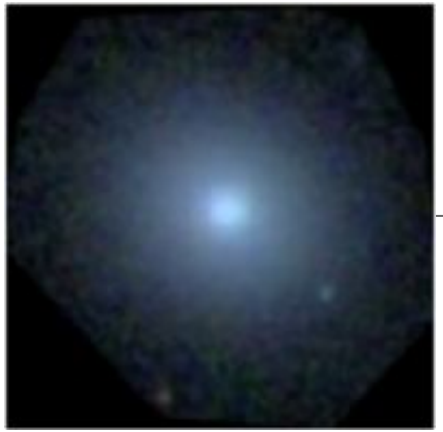
1. Zdjęcie jest rzucane do GrayScale Z1
2. Na podstawie Z1 obliczane jest mean, median i std
3. Z1 jest rzucane do binarnego Z2 po przeczyszczeniu - <= median + (sigma * std) jest ustawiana jako 0. Przyjęta finalnie sigma to 4
4. Na podstawie Z2 szukamy konturów obiektów, z wszystkich znalezionych konturów wybieramy K maksymalizujący **cv2.contourArea(c)-contour_distance_from_center(c) ** 2**
5. K1 służy jako maska dla Z0 - wszystko poza jego wnętrzem jest ustawiane na 0 - dostajemy Z3
6. Z Z3 wycinany jest minimalny prostokąt zawierający non-0 pixels a następnie rozszerzany i skalowany do kwadratu 150*150 z zachowaniem aspektu

! UWAGA - przeprocesowane zdjecia sprawiają wrażenia niebieskich ponieważ cv2 zmnienia warstwę R z B.

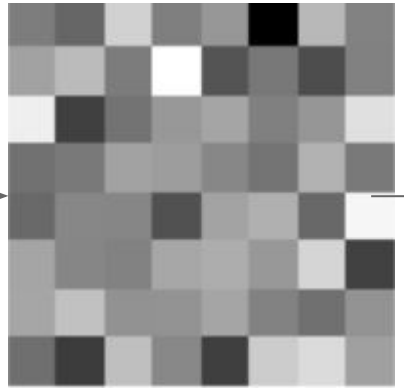
Podejście 1

Sam autoencoder jako compressor i feature extractor -> KMeans

Model



Encoding



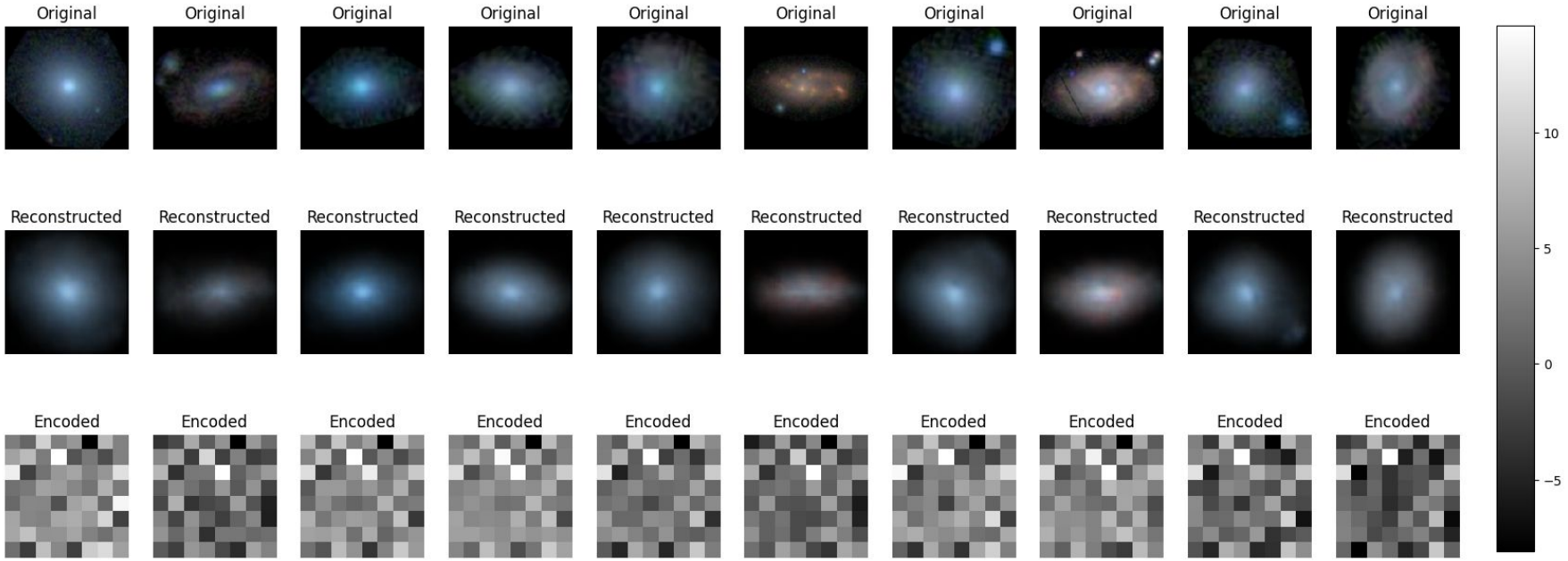
Decoding

Przeczyszczone
zdjęcie 150x150x3

Cechy
Wymiar:

Poglądowo -
czego nasz model
się nauczył?

Efekty



Podejście 2

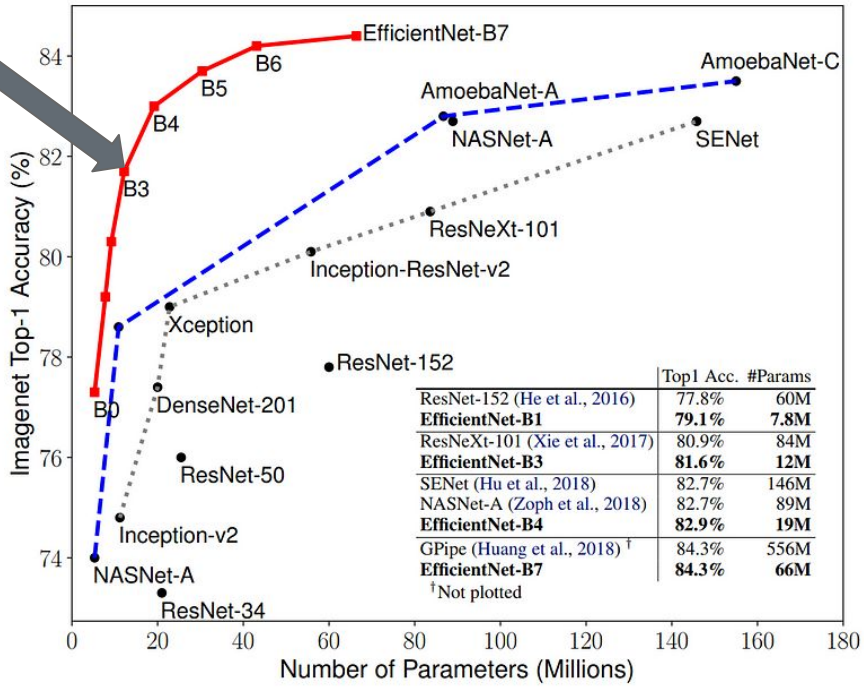
Warstwy konwolucyjne z EfficientNet B3 z ImageNetu -> autoencoder
jako compressor i feature extractor -> PCA -> clustering

EfficientNet zwraca wektor 4x4x1536

Autoencoder wektor 1x256

PCA wektor 1x8 (zachowanie wariancji: 0.99)

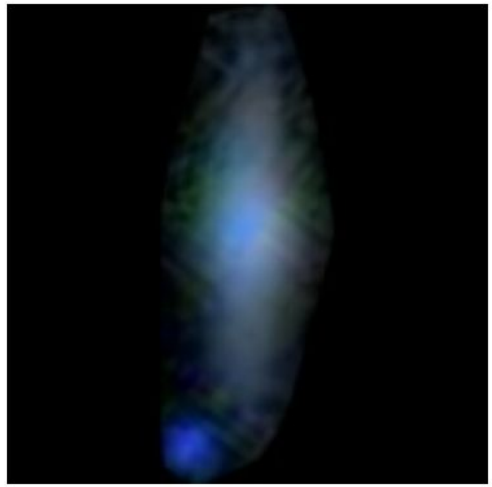
Wybór Feature Extactora

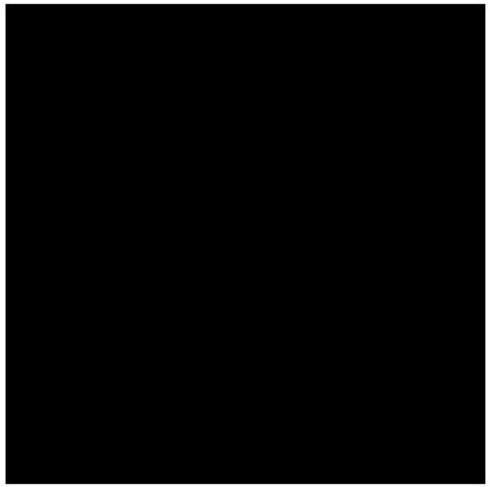


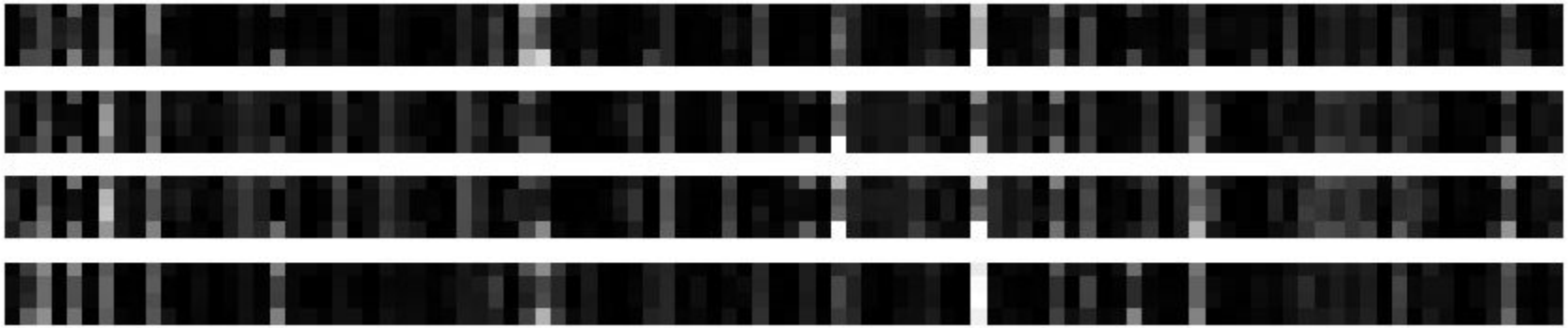
	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.1%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.6%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.9%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.3%**	**66M**

† Not plotted

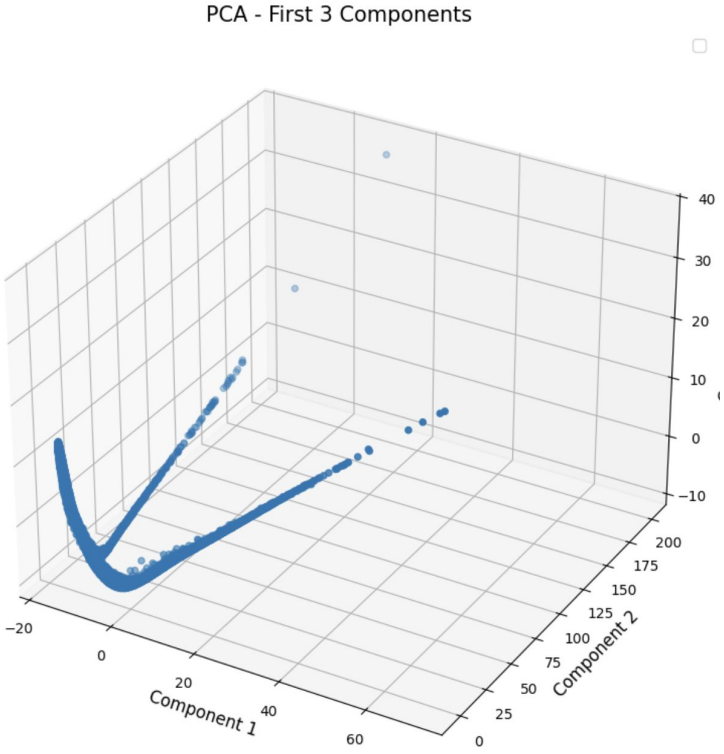
EfficientNet Features



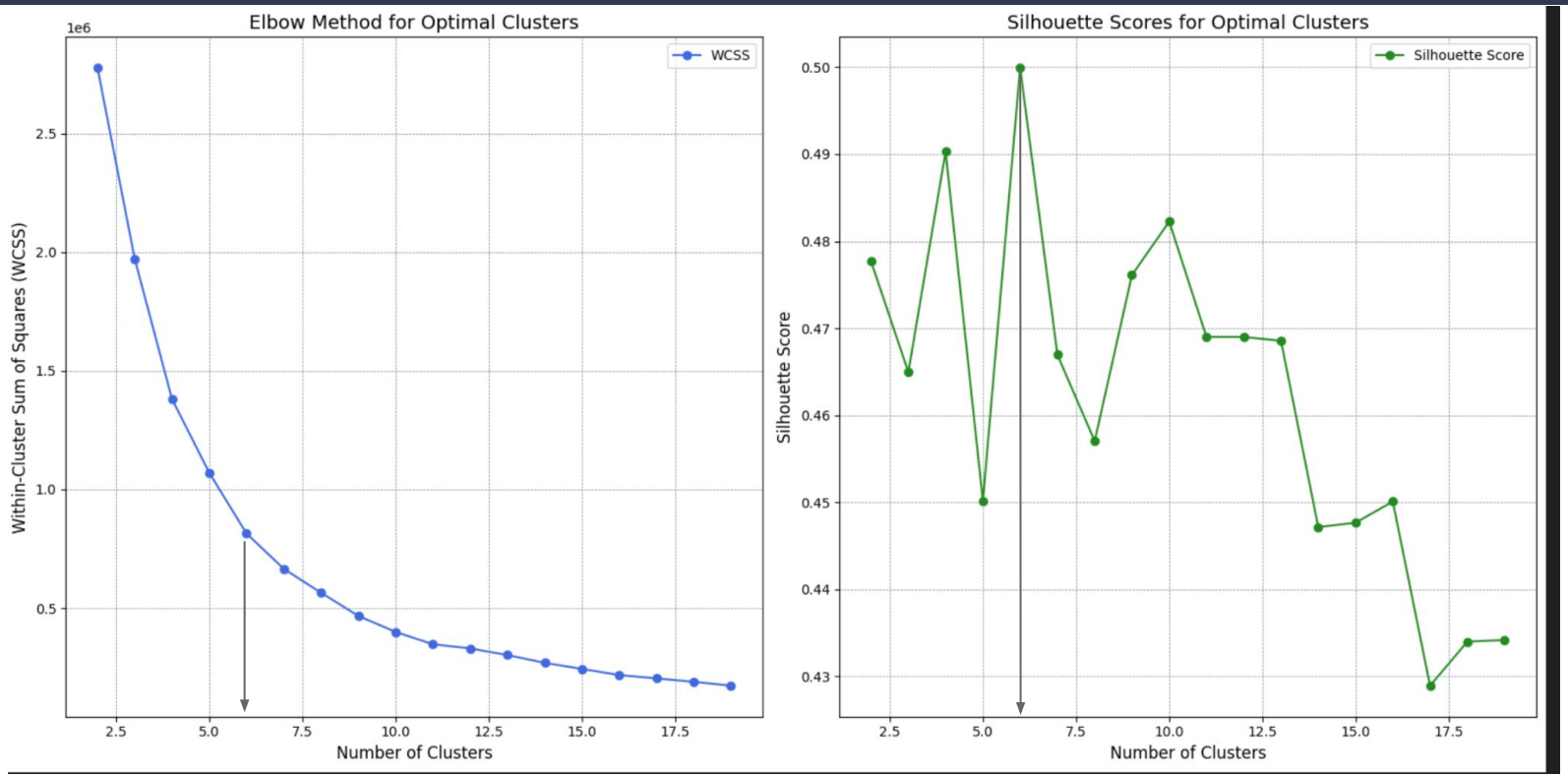




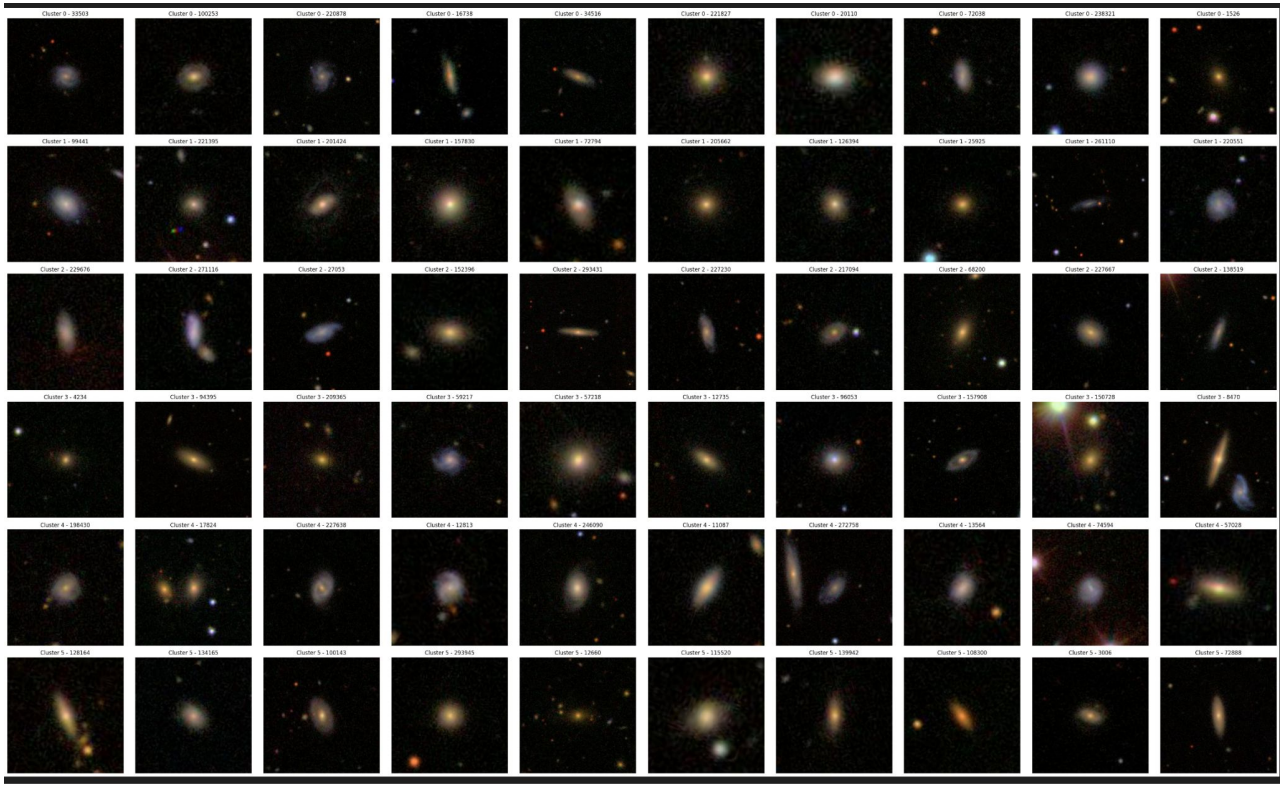
PCA


PCA - First 3 Components
Component 1
Component 2
−20
0
20
40
60
0
25
50
75
100
125
150
175
200
−10
0
10
20
30
40

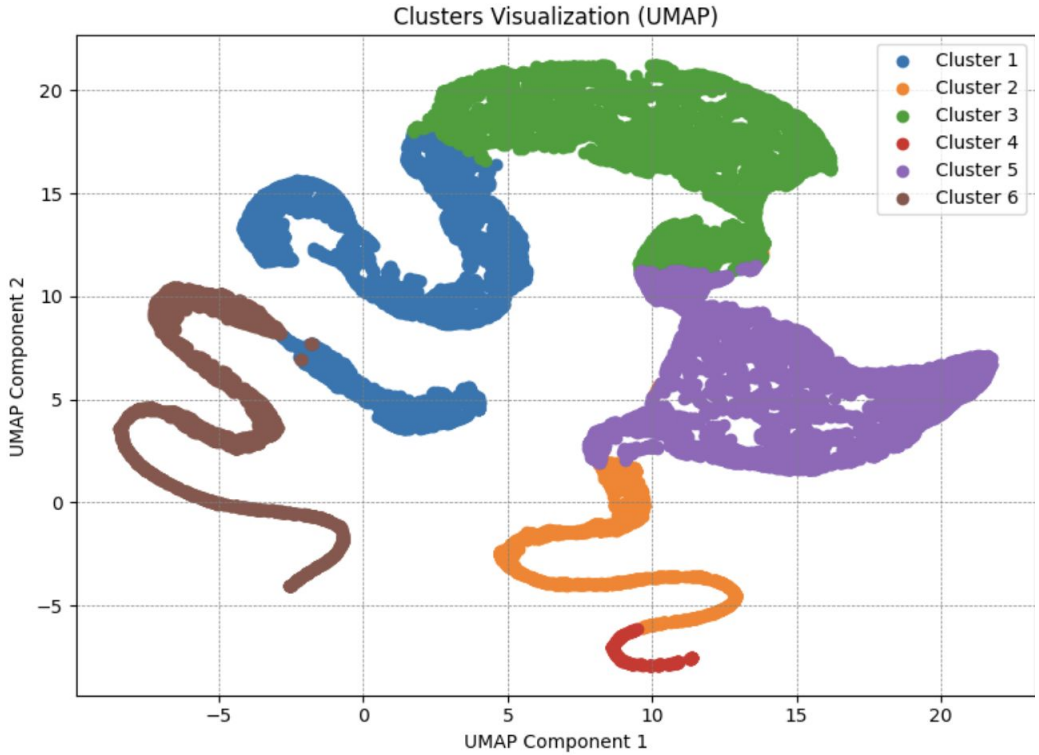
KMeans - wyboór liczby klastrów

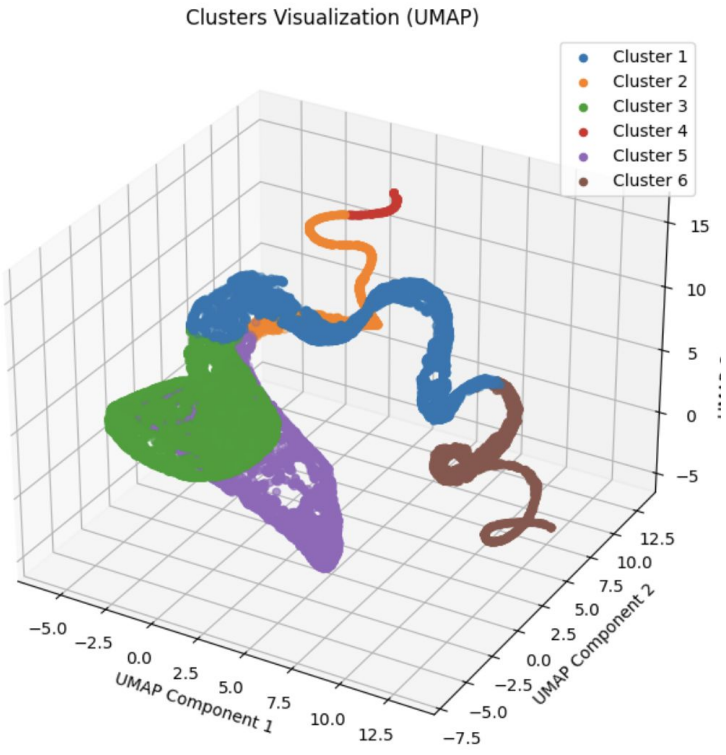


KMeans - Efekty

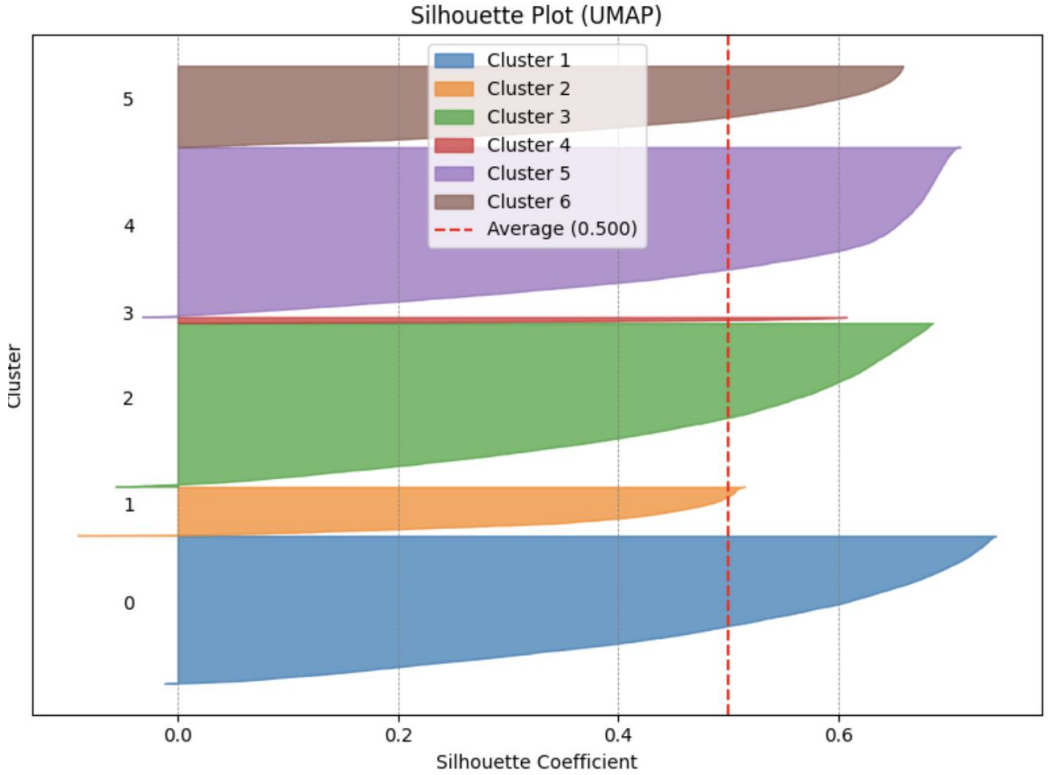


KMeans - Efekty - UMAP 2D, 3D





KMeans - Efekty



Podejście 3

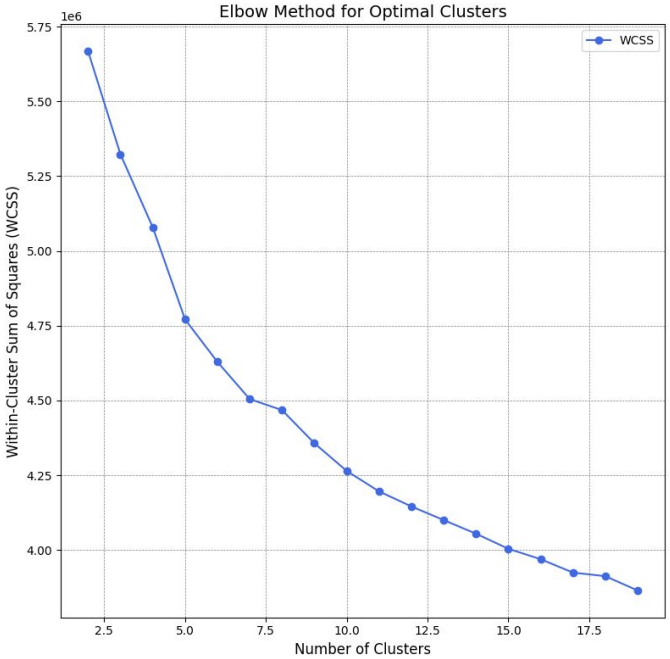
preprcesowanie danych -> feaure extractor - WGG16 -> PCA -> clustering

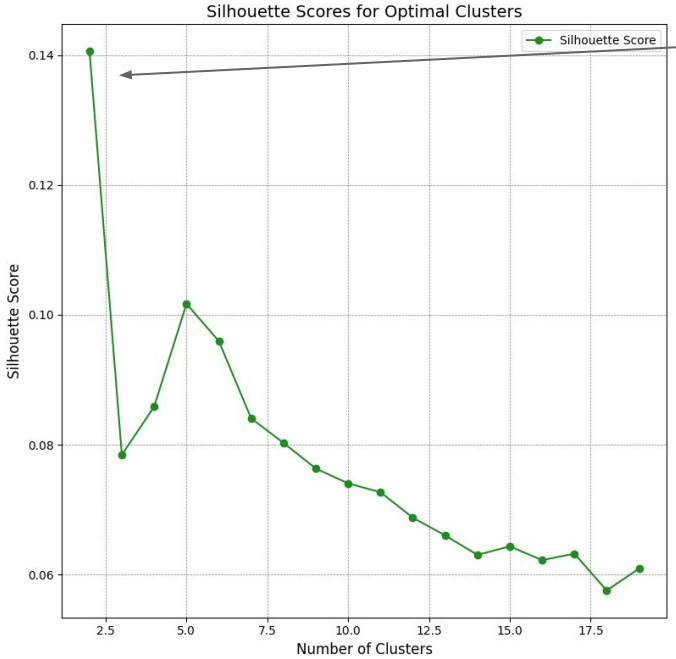
WGG16: zwraca wektor 1x4096

PCA: zwraca 1x1387 (zachowuje 99% wariancji)

Clustering: 5 Klastrów

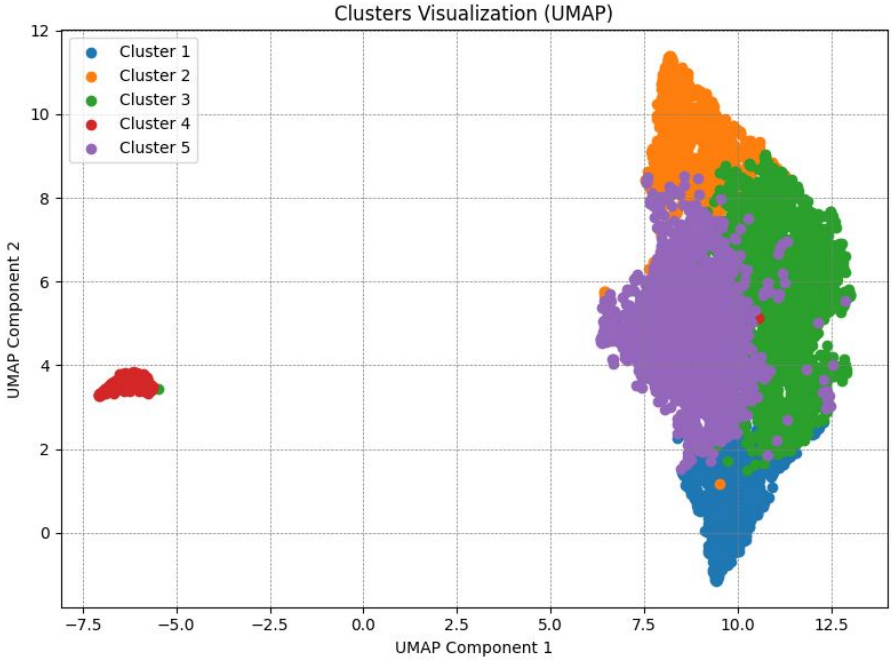
KMeans i silhouette

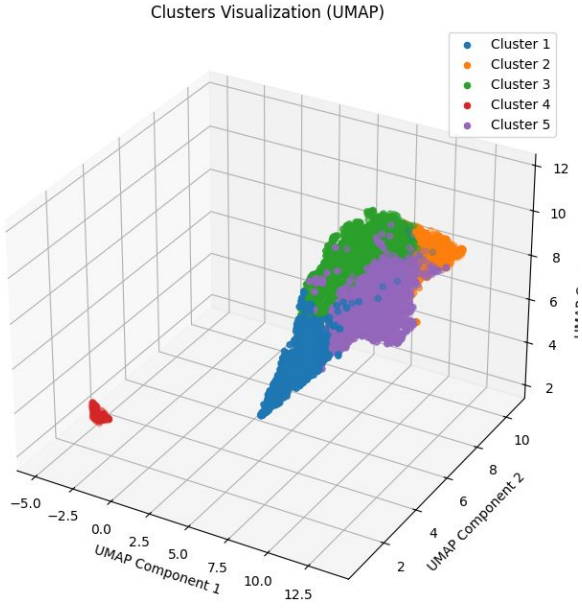




Low score

Wyjaśnialność





Końcowy wybór

- Wybieramy podejście 2
- Najlepszy Silhouette score: 0,52
- Najlepsza interpretowalność
- Indeks Daviesa-Bouldina: 0.7

Dziękujemy za uwagę