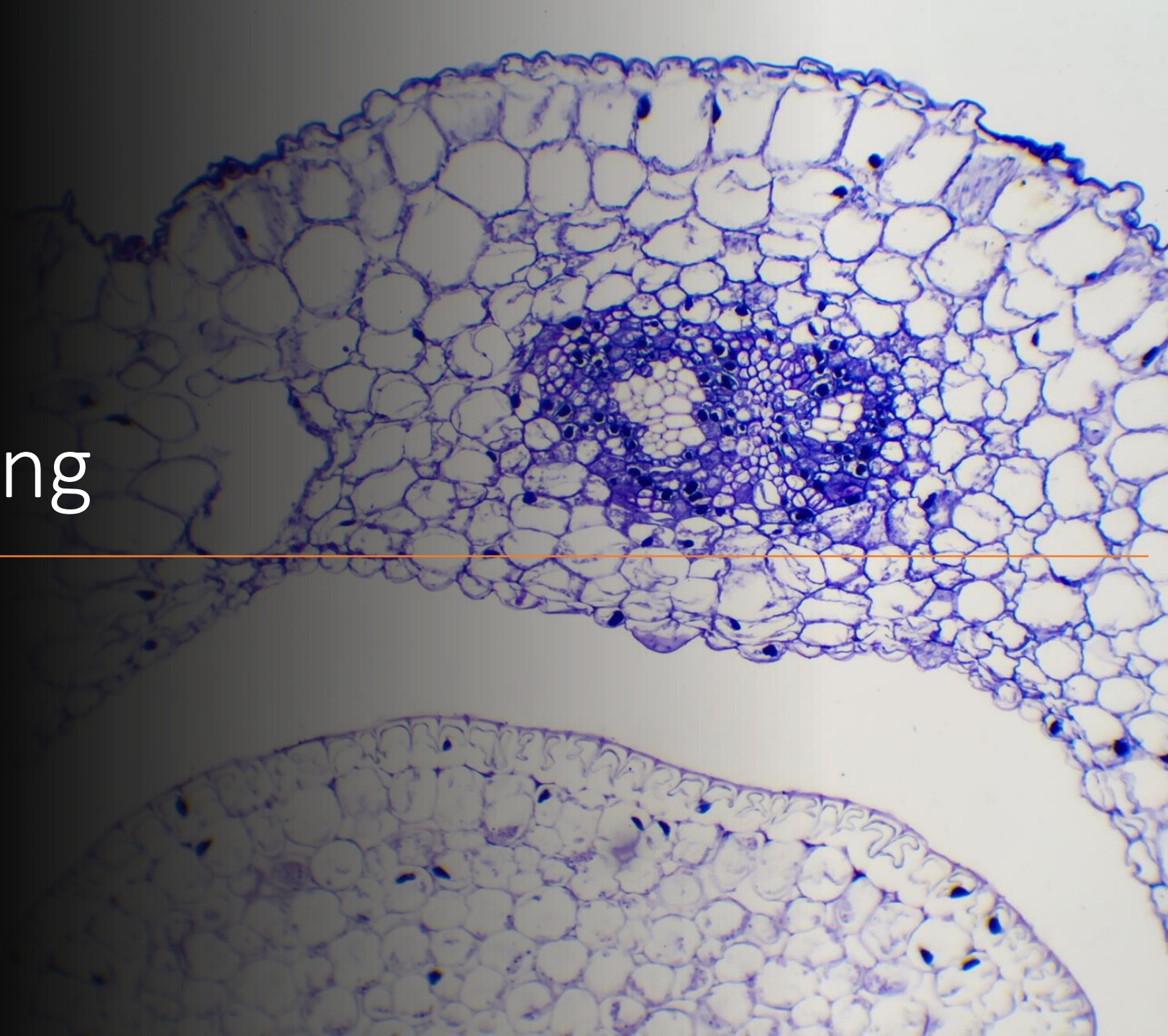
Skin Diseases Image Clustering

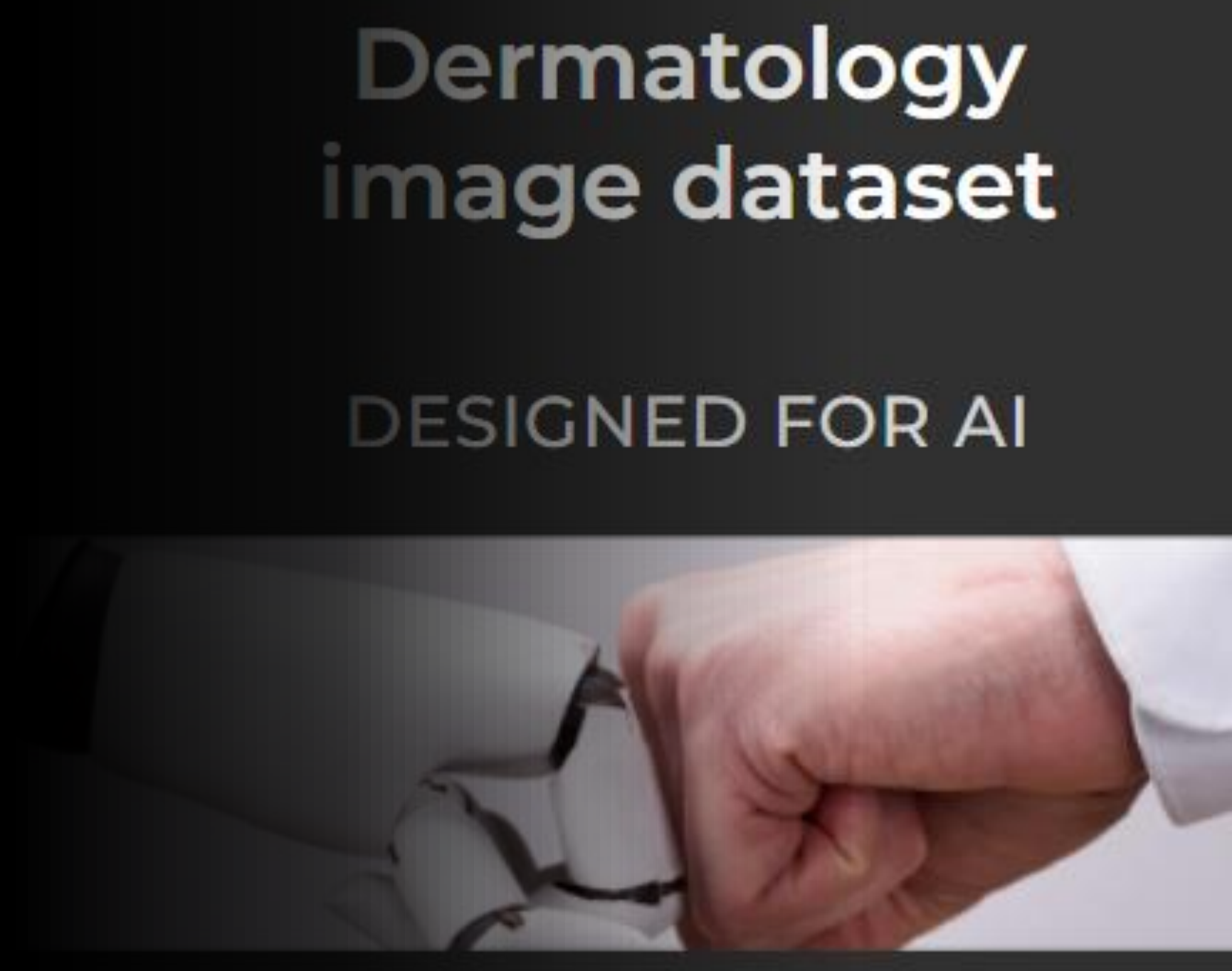


Igor Kołodziej | Kamil Eliaszuk

Data and Business case

- **Dataset**: DERMNET – online dermatology source, 19 500 images (15 500 in training set), 23 types of skin diseases, RGB, JPEG format
- **Business case**: To divide a large, messy dataset into groups that correspond to different parts of the body.

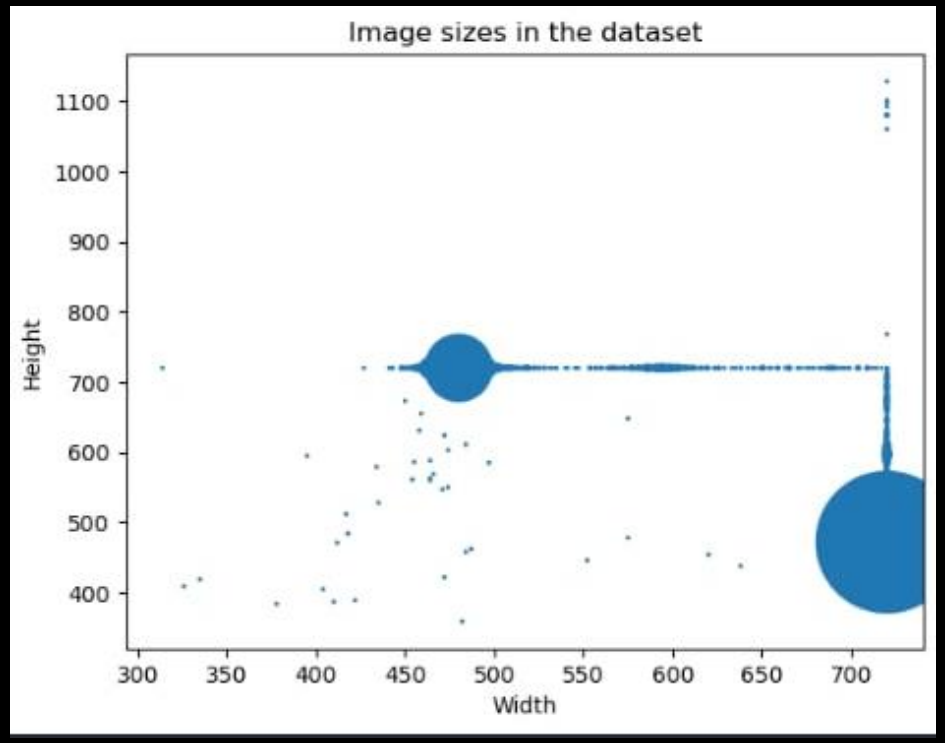
This would allow further research or model creation on our clusters



Meet out data

Sample scaled images

Meet out data



Various image sizes

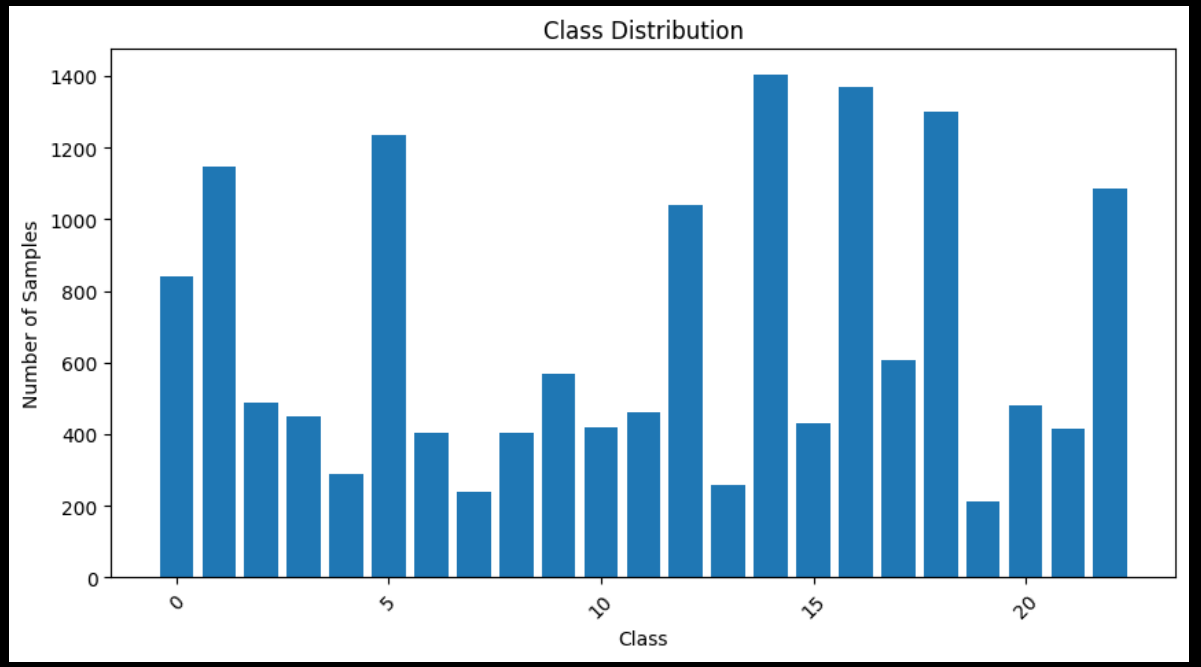
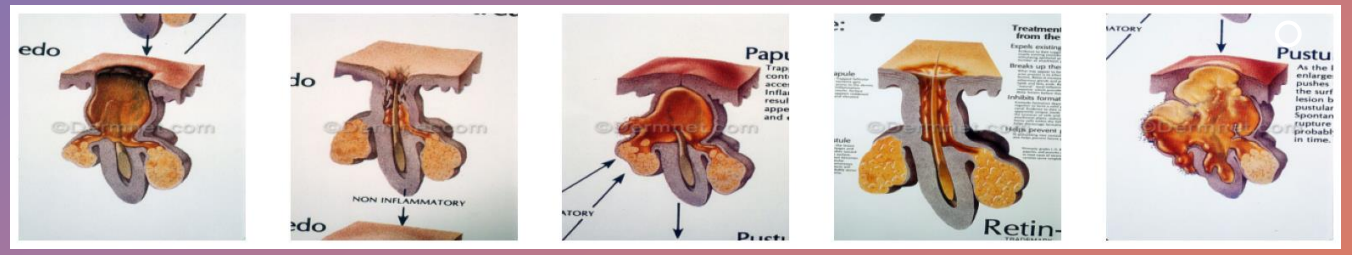
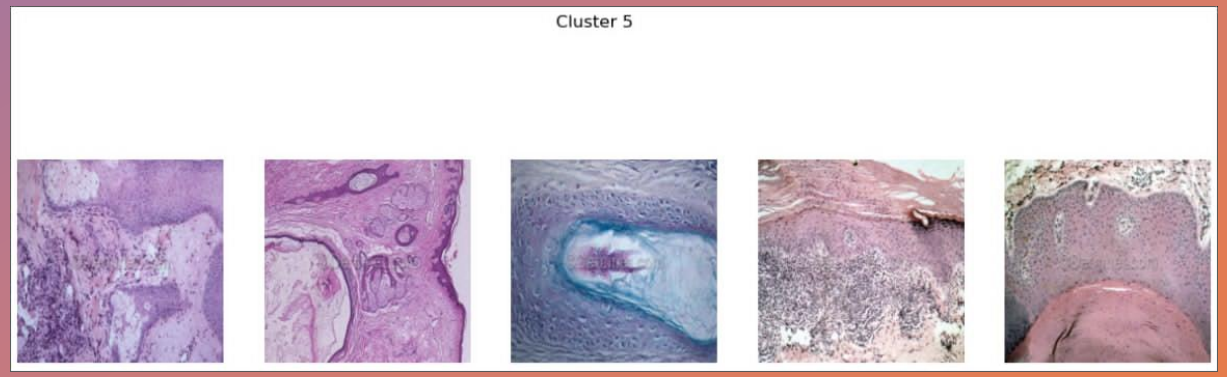


Image labels (unused later in clustering)

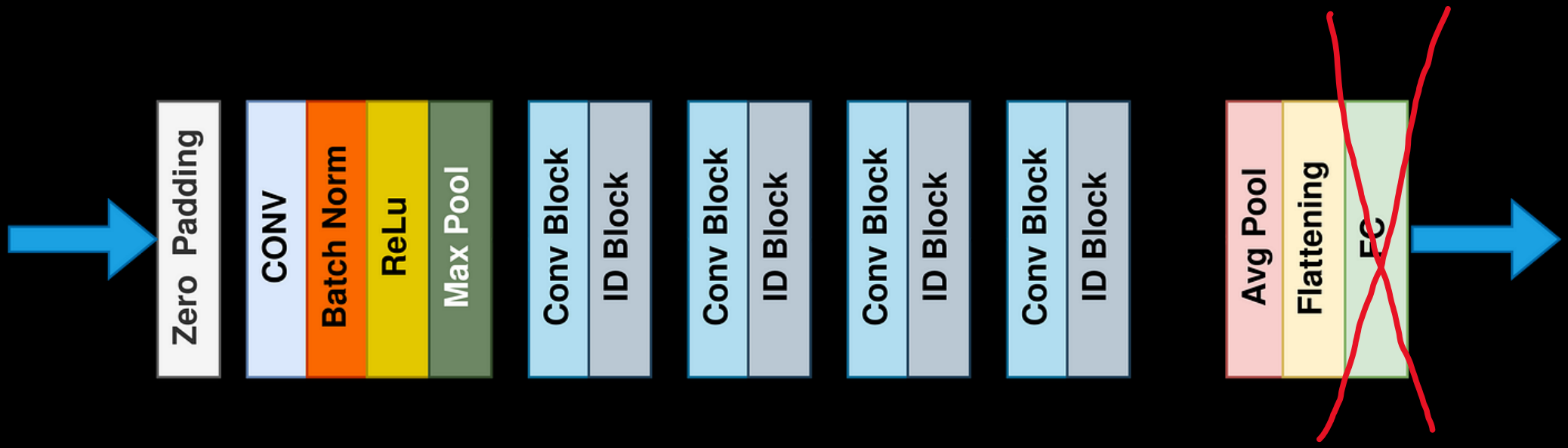
Weird data and duplicates





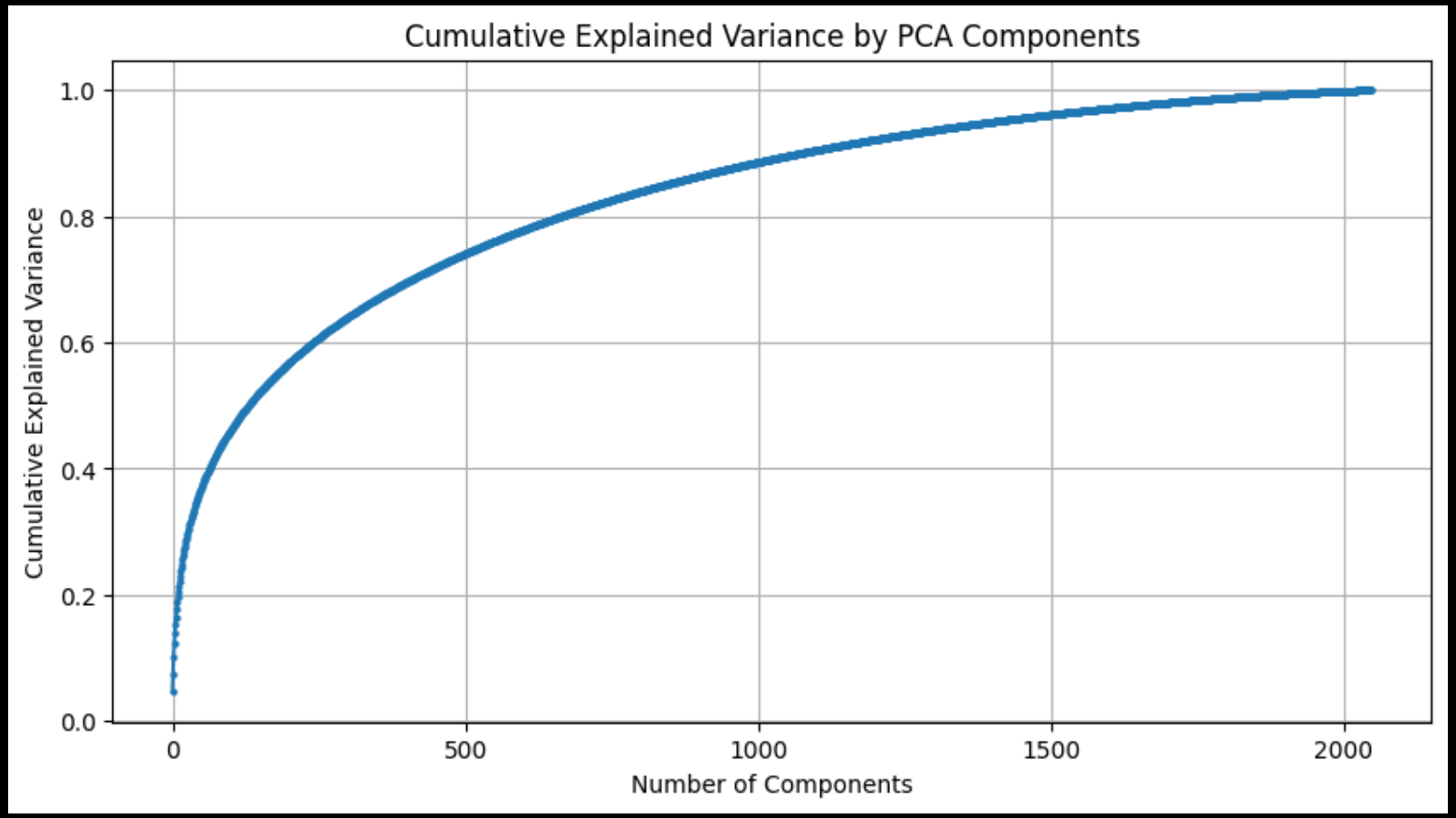
Data preprocessing and feature extraction

1st approach:
ResNet-50 CNN architecture without the last layer and pre-trained weights on ImageNet

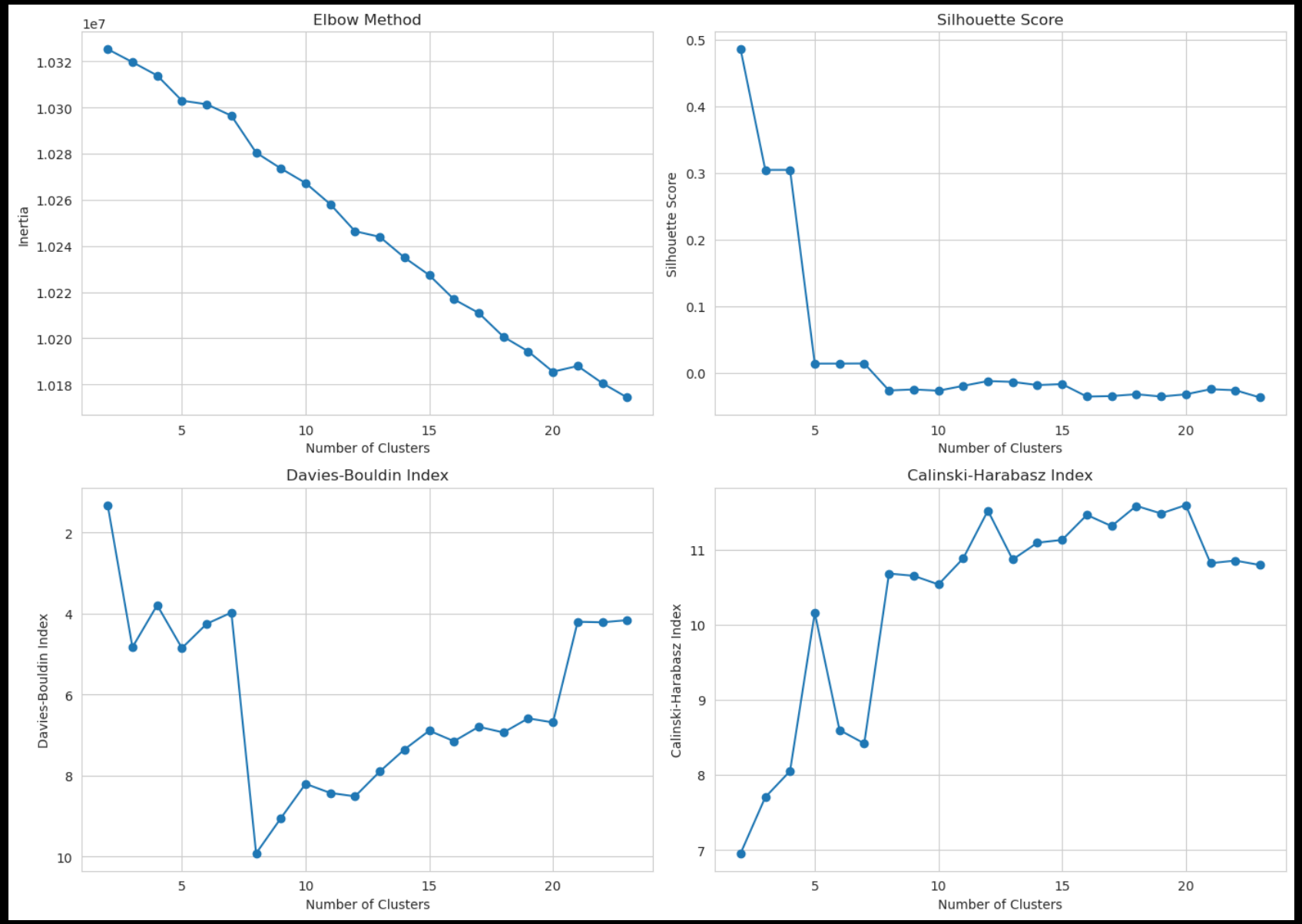


Data preprocessing and feature extraction

1st approach:
2048 features from ResNet -> Standarization -> dimensionality reduction with PCA -> Standarization

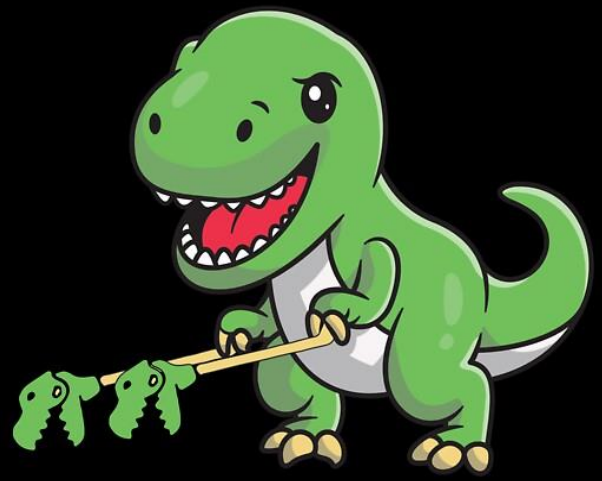


Chosen threshold: 80% variance with 665 components

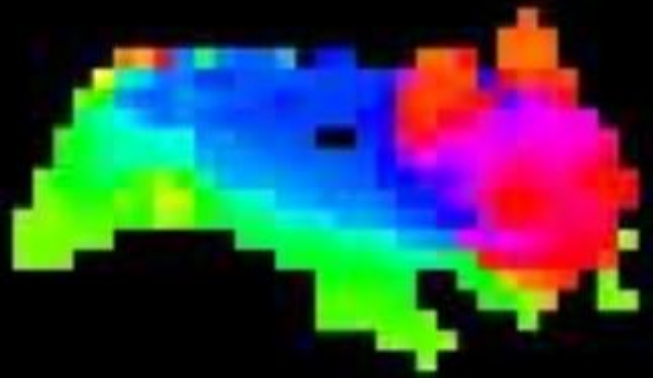


1st approach with Kmeans – bad results

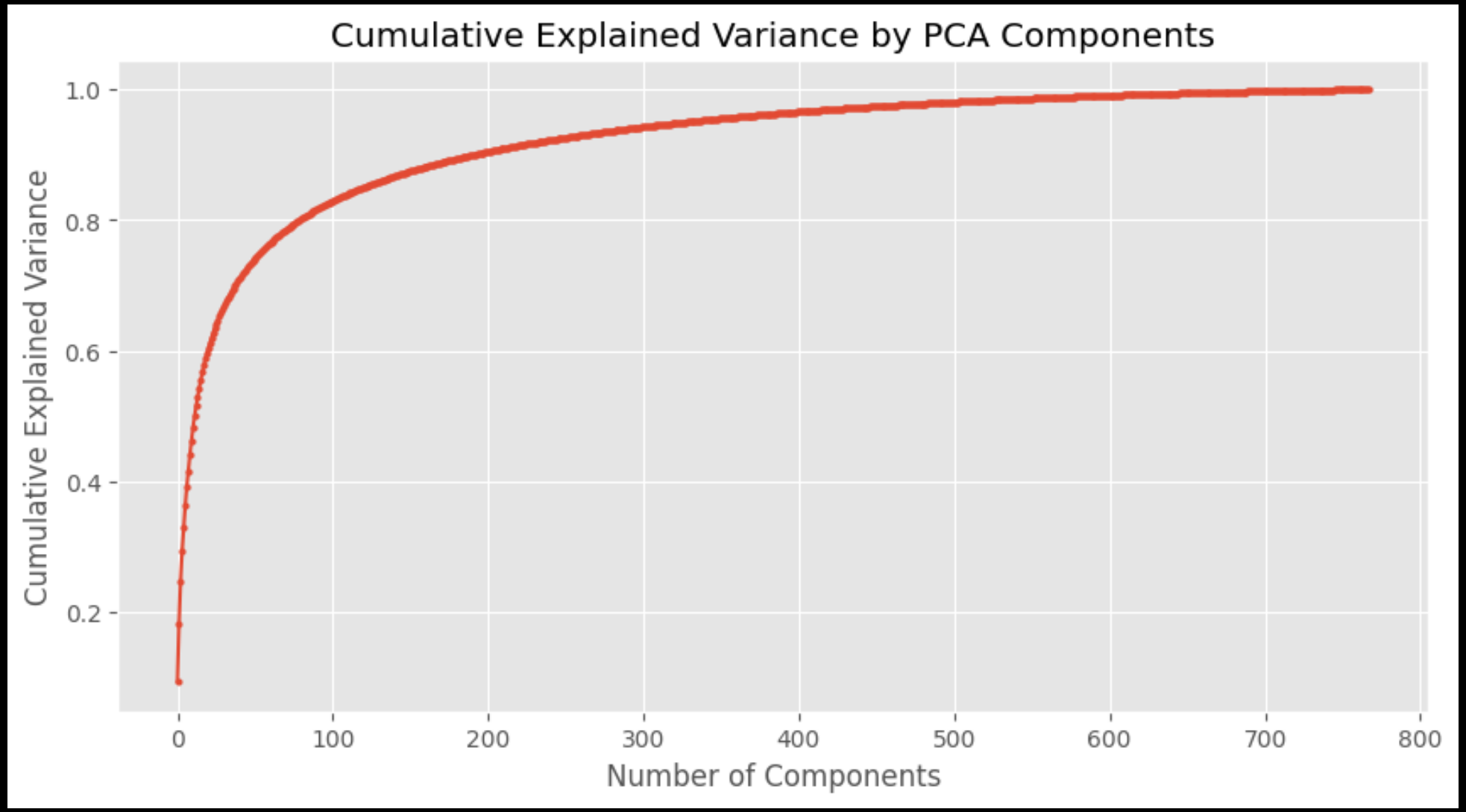
NOW I AM



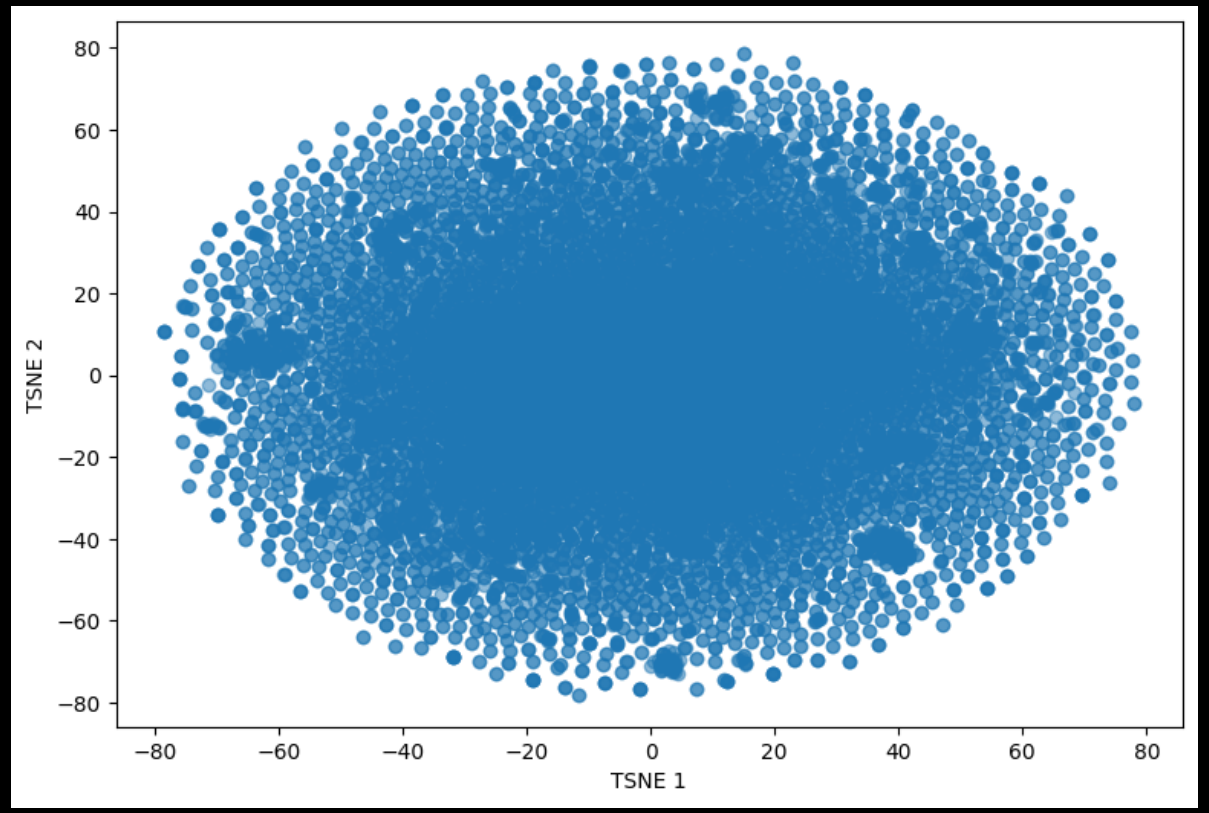
UNSTOPPABLE



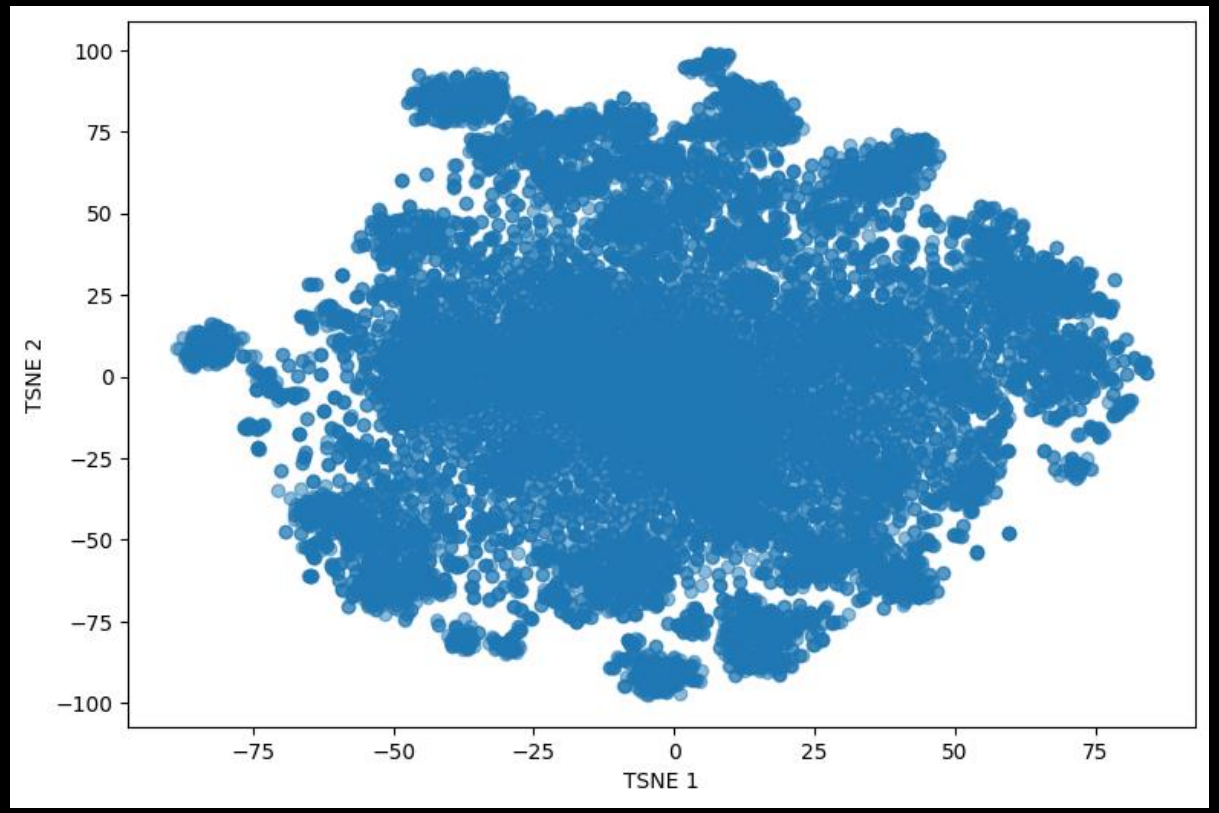
DINOv2 – less features, better results



DINOv2 – less features, better results



T-SNE on ResNet



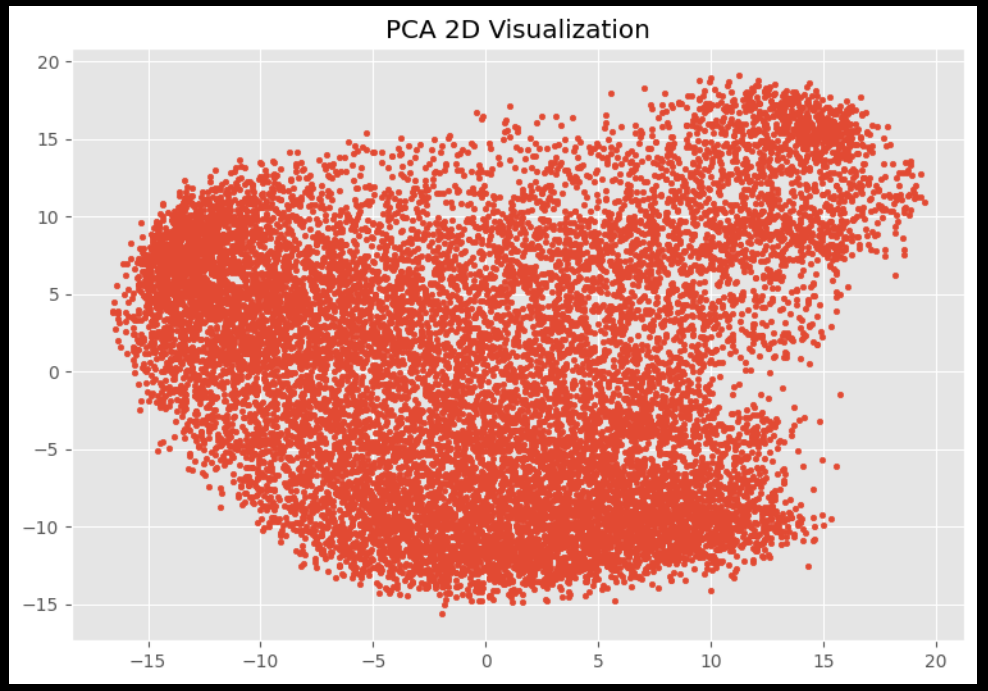
T-SNE on Dino

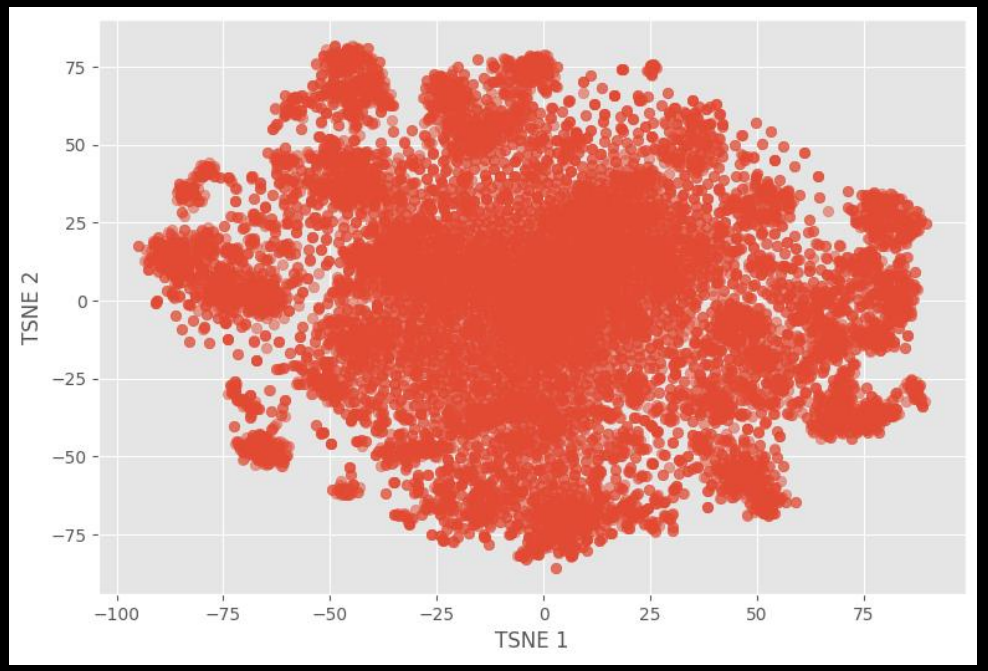
UMAP >> PCA/t-SNE

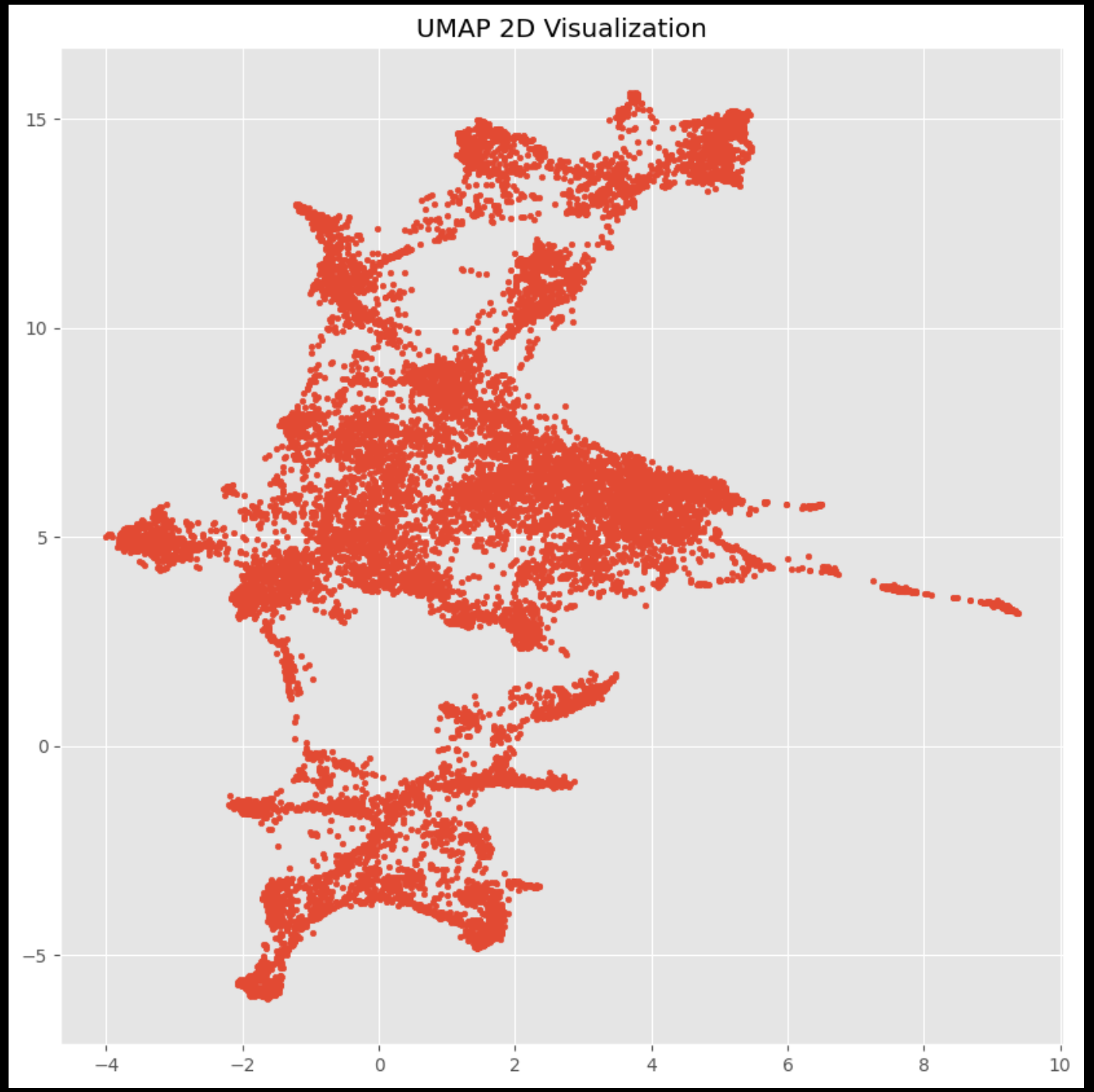
UNIFORM MANIFOLD

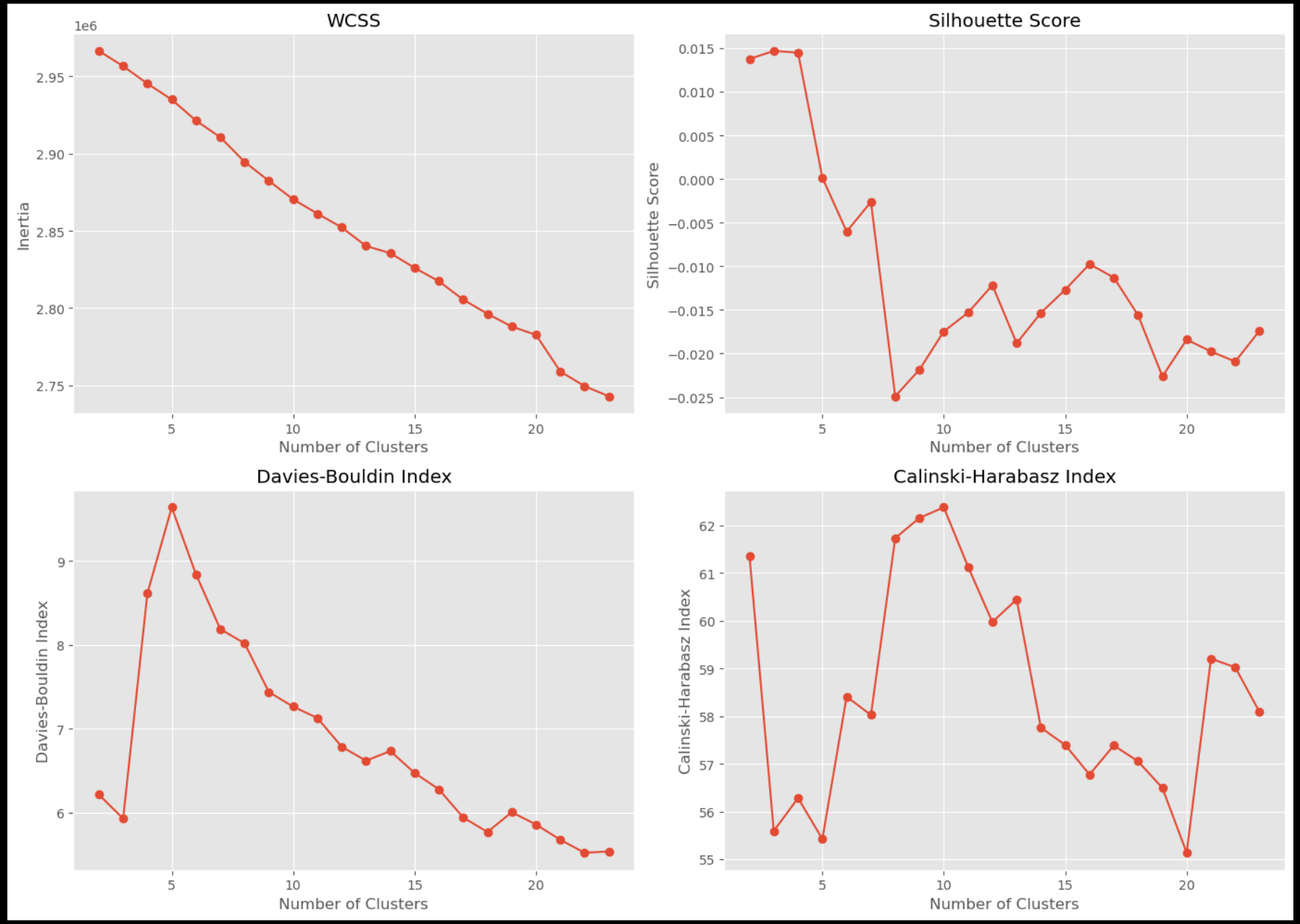
UMAP

APPROXIMATION & PROJECTION

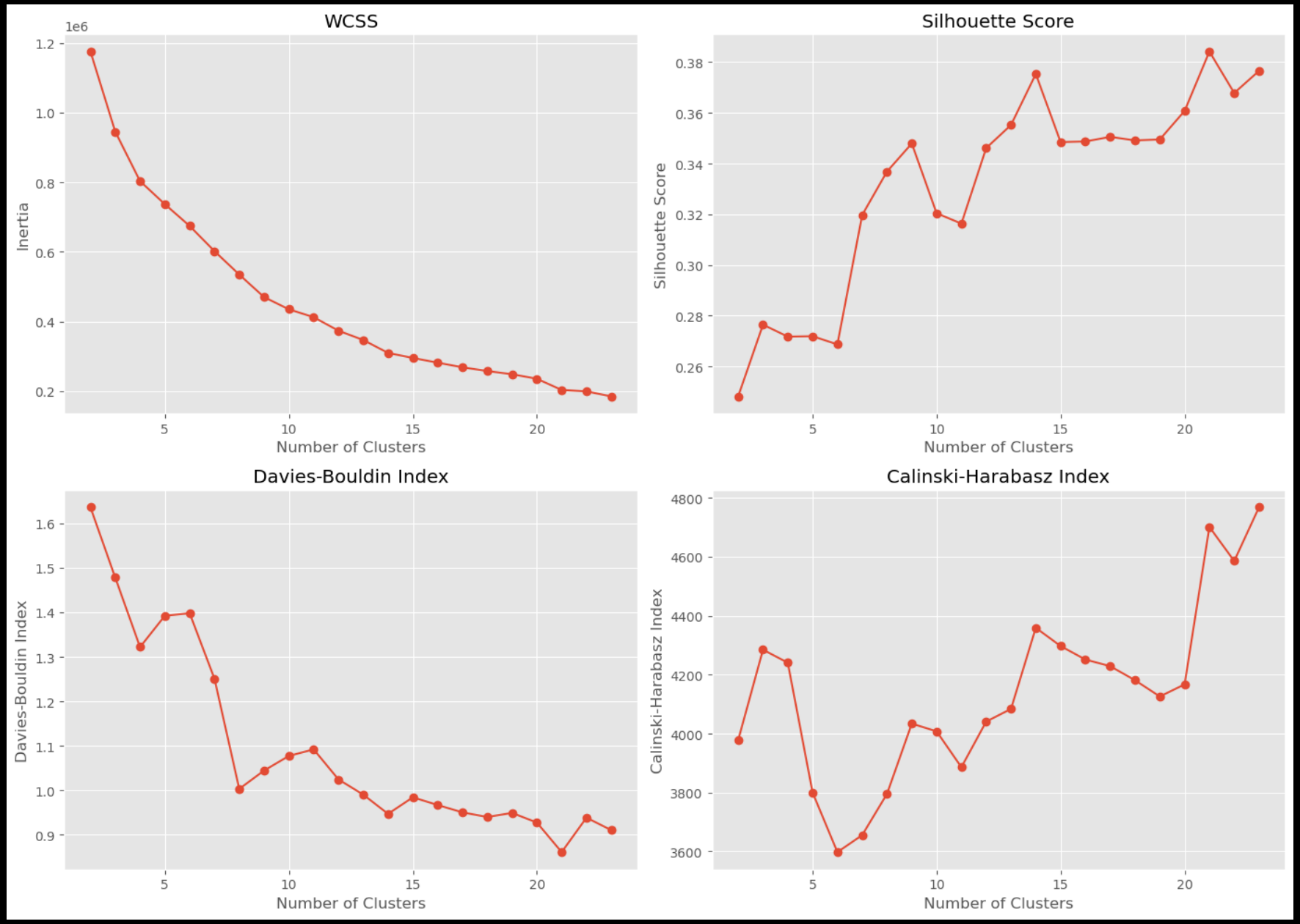

PCA 2D Visualization
20
15
10
5
0
−5
−10
−15
−15
−10
−5
0
5
10
15
20


75
50
25
0
−25
−50
−75
TSNE 2
−100
−75
−50
−25
0
25
50
75
TSNE 1


UMAP 2D Visualization
15
10
5
0
−5
−4
−2
0
2
4
6
8
10



Kmeans on dino and pca

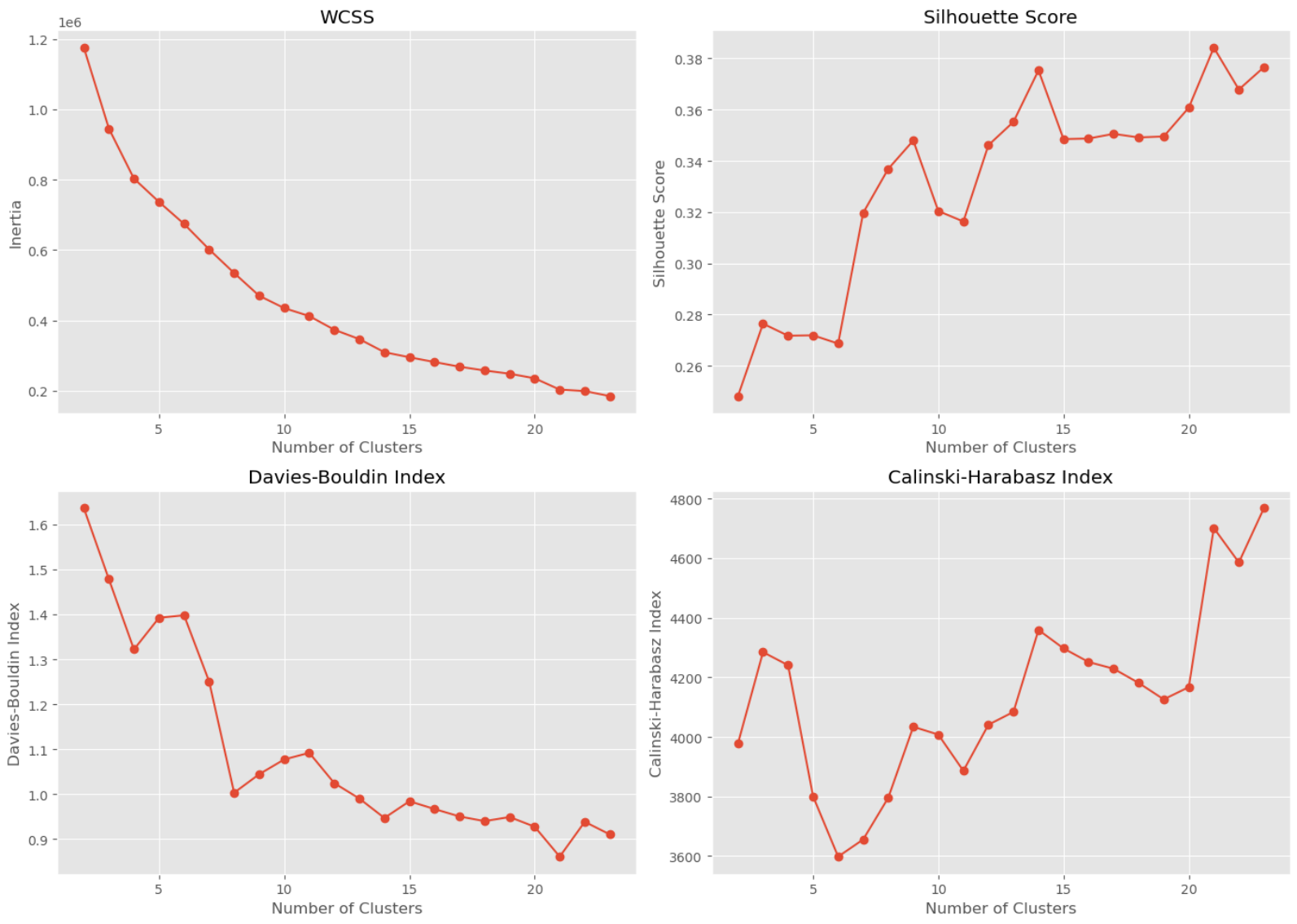


Kmeans on dino and umap

Data preprocessing and feature extraction

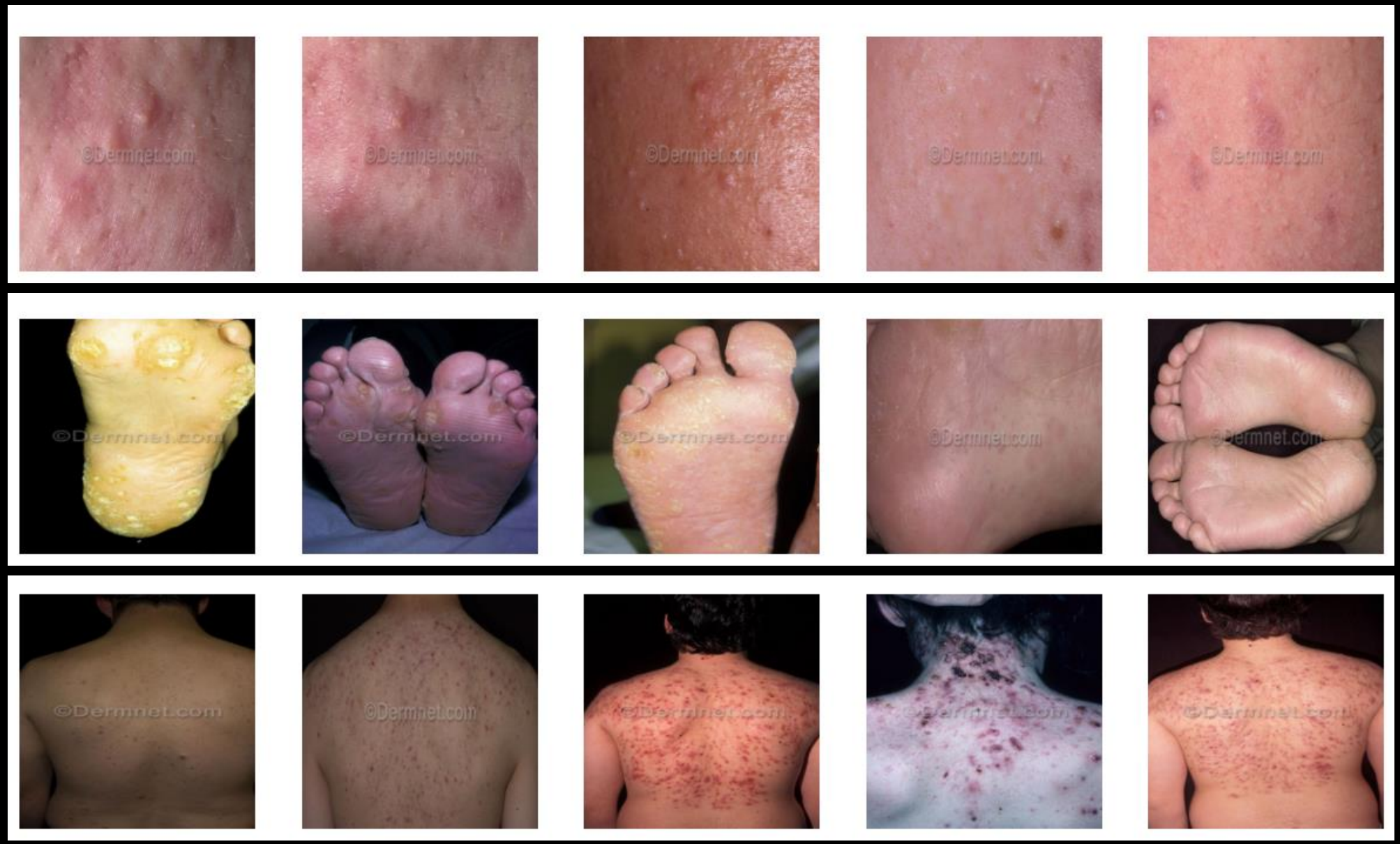
2nd approach:
Feature extraction with Dino -> Standarization -> UMAP with 100 components -> Standarization -> Clustering

Clustering models

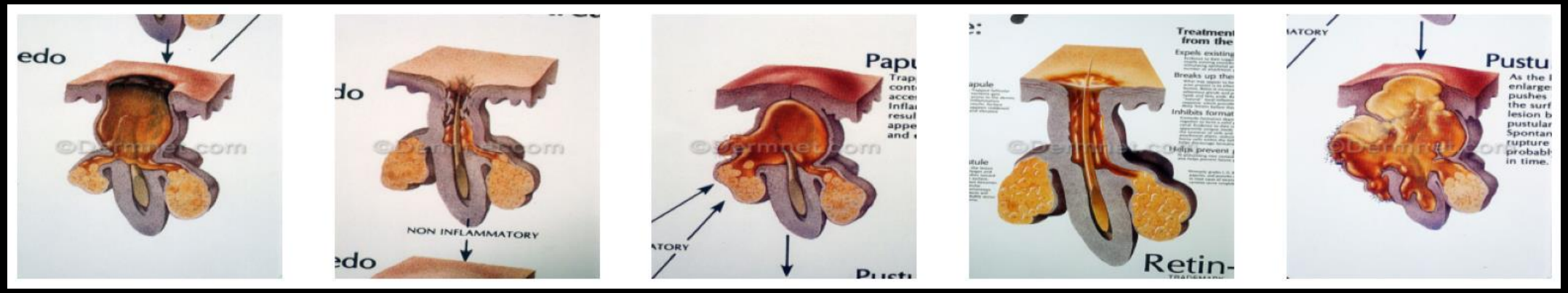


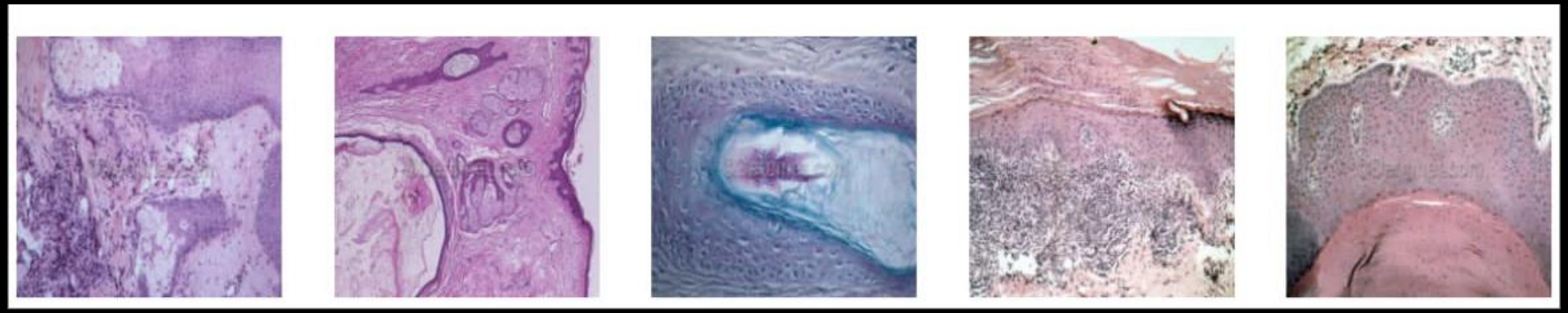
Kmeans – 14 or 21 clusters

Clustering models

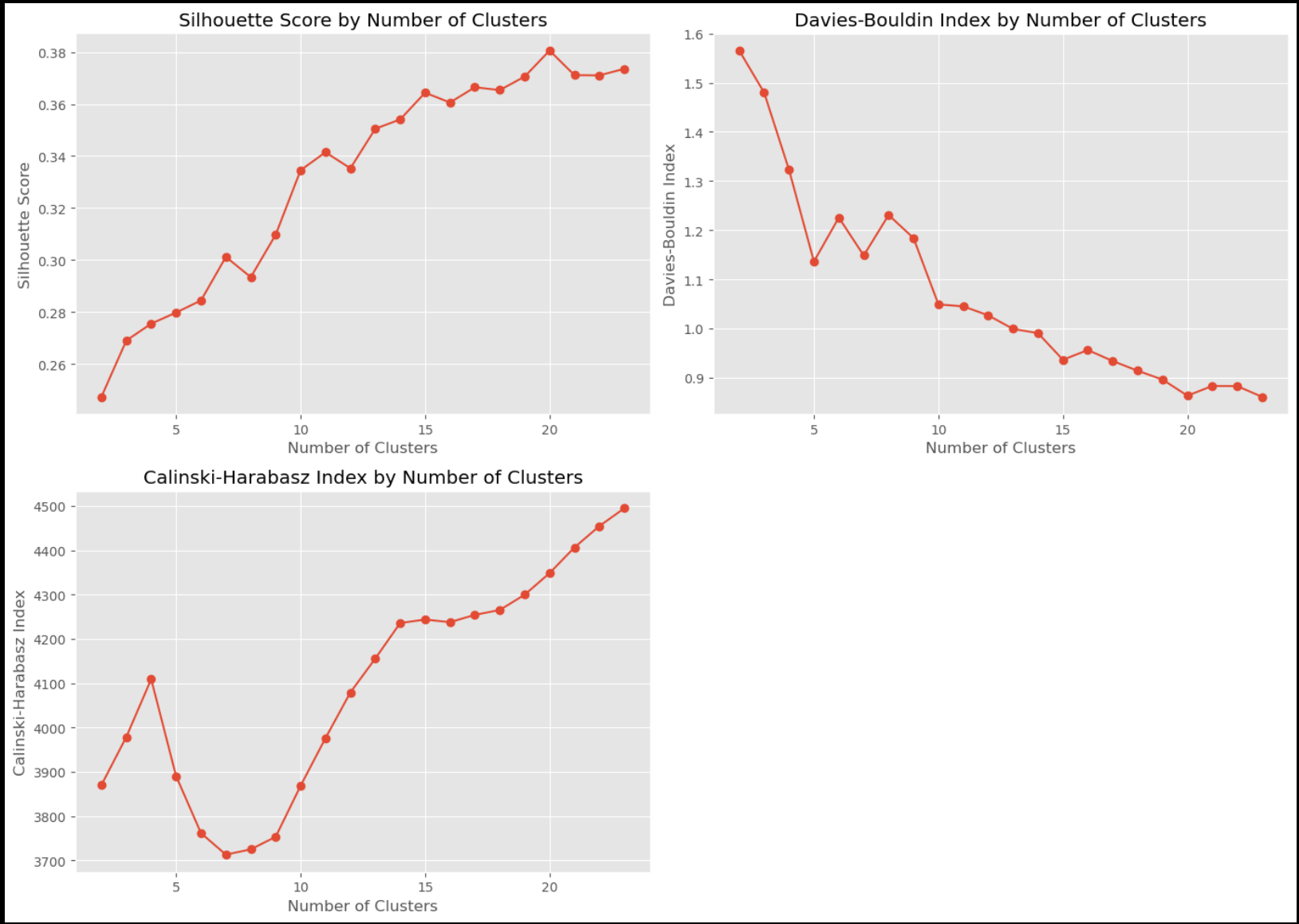


Clustering models





Clustering models



Agglomerative clustering – 15 or 20 clusters, clusters are similar to KMeans

Clustering models

DBSCAN, HDBSCAN, OPTICS – difficult to find good hyperparameters, one big cluster, too may points classified as noise

```
pd.Series(dbscan_umap_labels).value_counts().sort_values(ascending=False)
```

✓ 0.0s

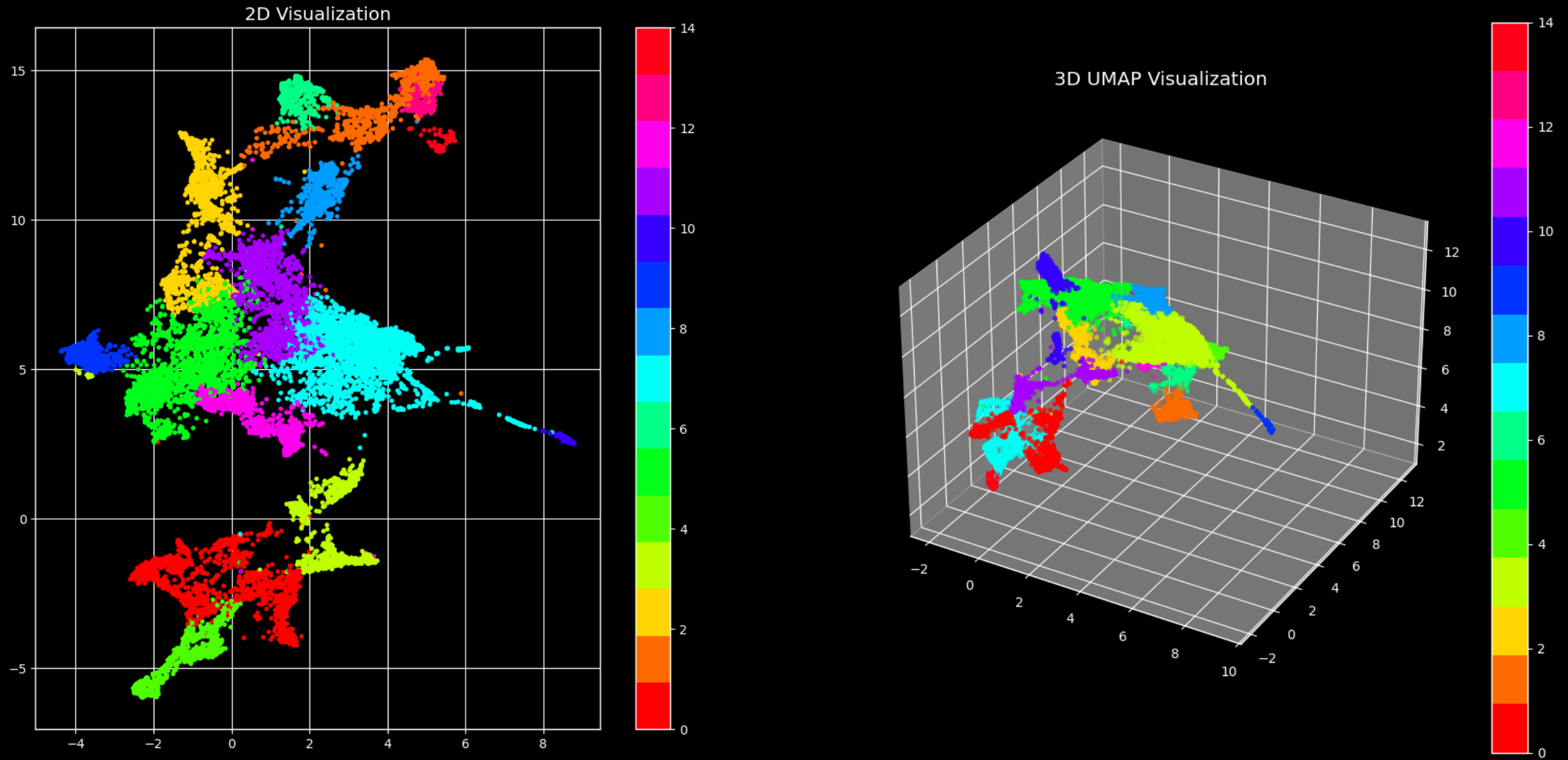
```
 1    8892
-1    2664
 0    1615
 2     564
 3     328
 7     234
 6     231
 4     222
 5     143
Name: count, dtype: int64
```

```
pd.Series(hdbscan_umap_labels).value_counts().sort_values(ascending=False)
```

✓ 0.0s

```
 10    7119
 5     2606
-1     1778
 6      653
 7      590
 9      421
 8      351
 2      315
 4      293
 3      247
 11     213
 0      177
 1      130
Name: count, dtype: int64
```

Final results



Agglomerative clustering { 'linkage': 'ward’, 'metric’: 'euclidean’, 'n_clusters’: 15 }

Final results

In summary our model:

- Extracts meaningful features from images
- Projects them to fewer dimensions to speed up the model/ visualize the structure of the data
- Creates clusters of similar skin images (based mostly on body parts)

It provides a way to clean the data and group it in a relevant way that would help further research or classification tasks.