DIMENSIONALITY REDUCTION METHODS

Tomáš Čelko, Martin Bakoš

PRESENTATION STRUCTURE

- Datasets
  - Flowers
  - Alzheimer
  - Fruits
- Feature extraction
- Reduction methods
- DR quality evaluation
  - kNN
  - Coranking matrix
  - NX curves
- Visualizations

DATASETS - FLOWERS

Description:

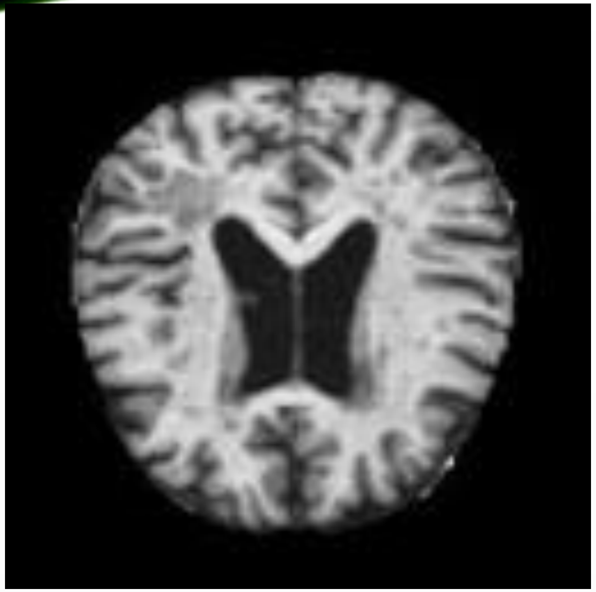
- Noisy

Categories:

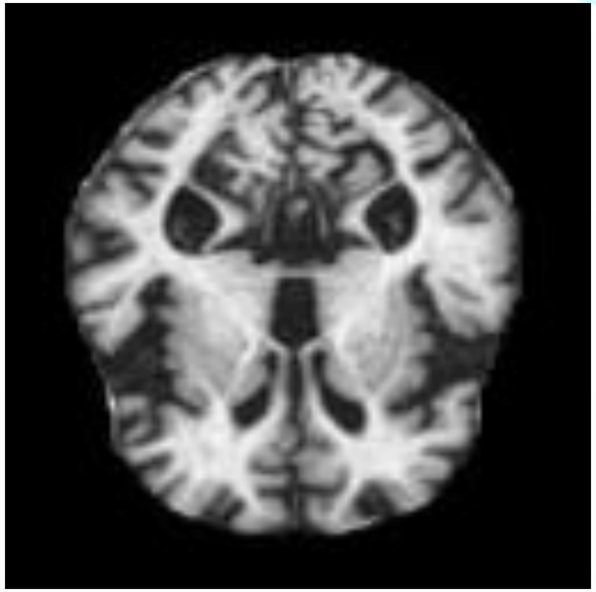
- Rose
- Dadelion
- Sunflower
- Tulip
- Daisy

DATASET - ALZHEIMER

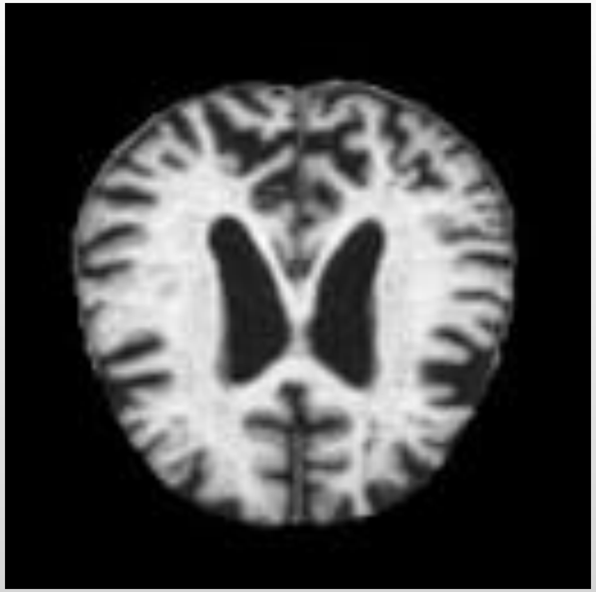
- Non demented (0)
- Very mild demented (1)
- Mild demented (2)
- Moderate demented (3)



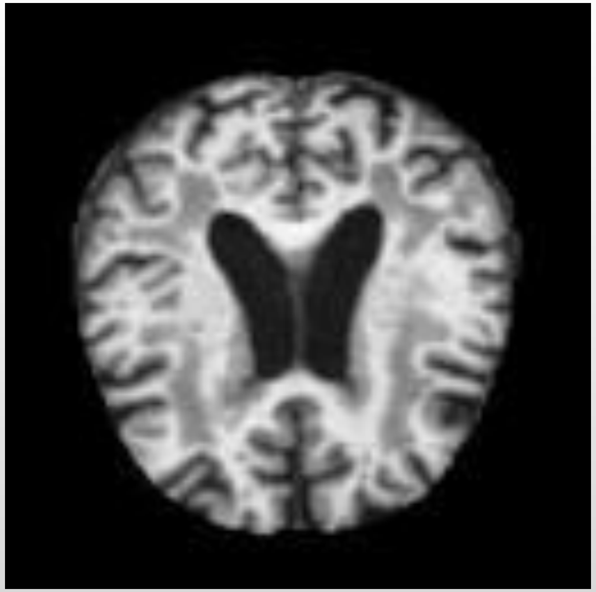
0



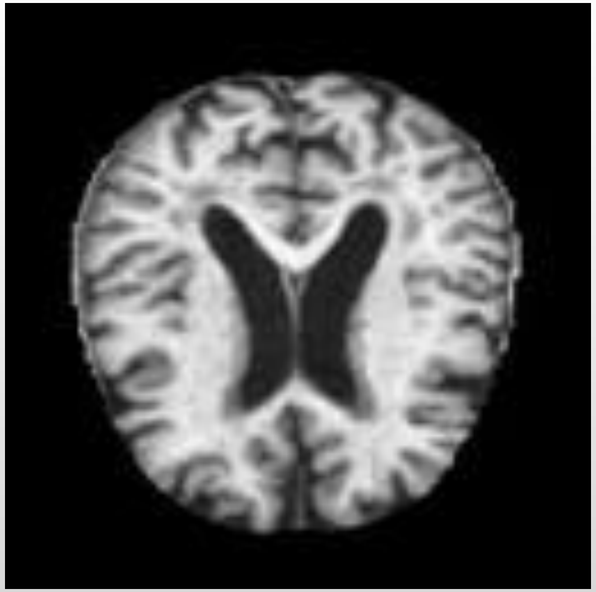
3



3



1



2

DATASET – FRUITS

Description:

- 131 different classes
- Clean images

Classes(examples):

- Apples
- Avocado
- Banana
- Blueberry
- Cucumber
- Grape
- Hazelnut
- Lime
- Onion
- Raspberry
- Tomato
- Watermelon

FEATURE EXTRACTION

- Feature extraction for images:
  - Trivial = Pixel features
  - Non – trivial = computed features
    - Manual (statictics – means, variances of colors for the regions of image)
      - Interpretable
    - Artificially extracted (NN inner representation)
      - Good representation

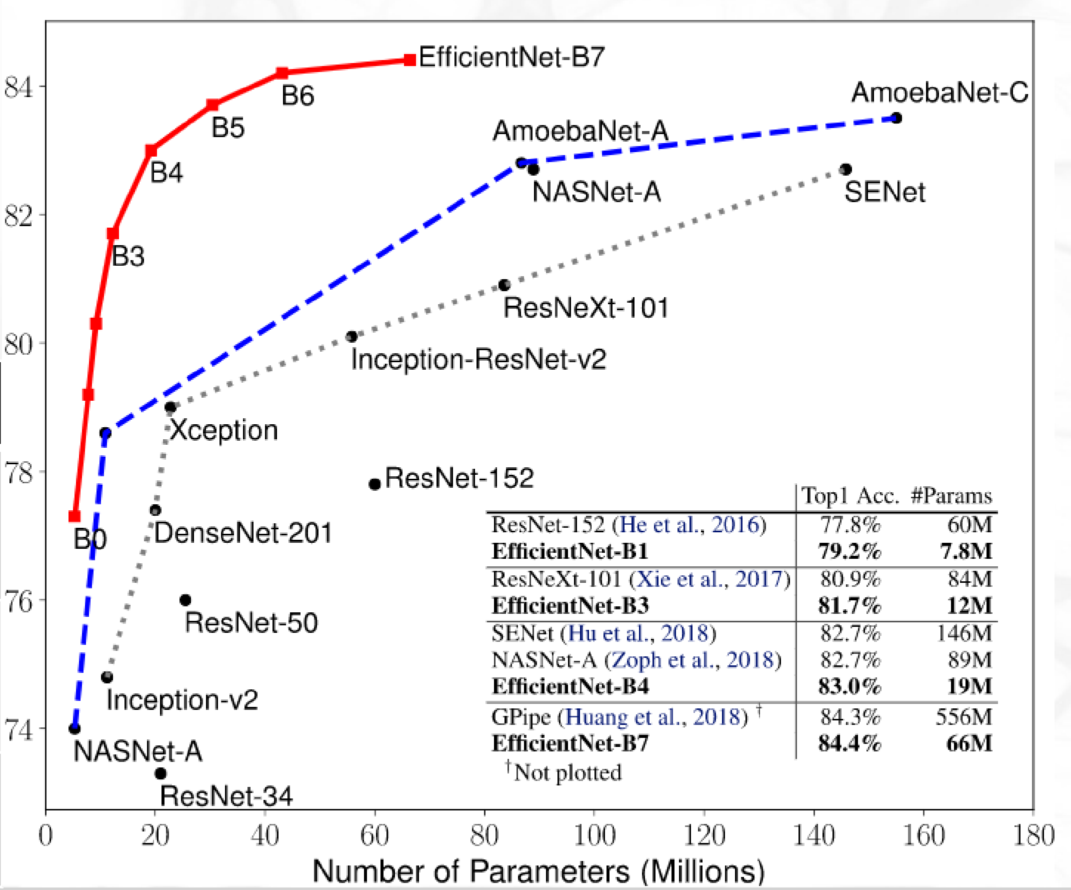
Our approach

EFFICIENT NET

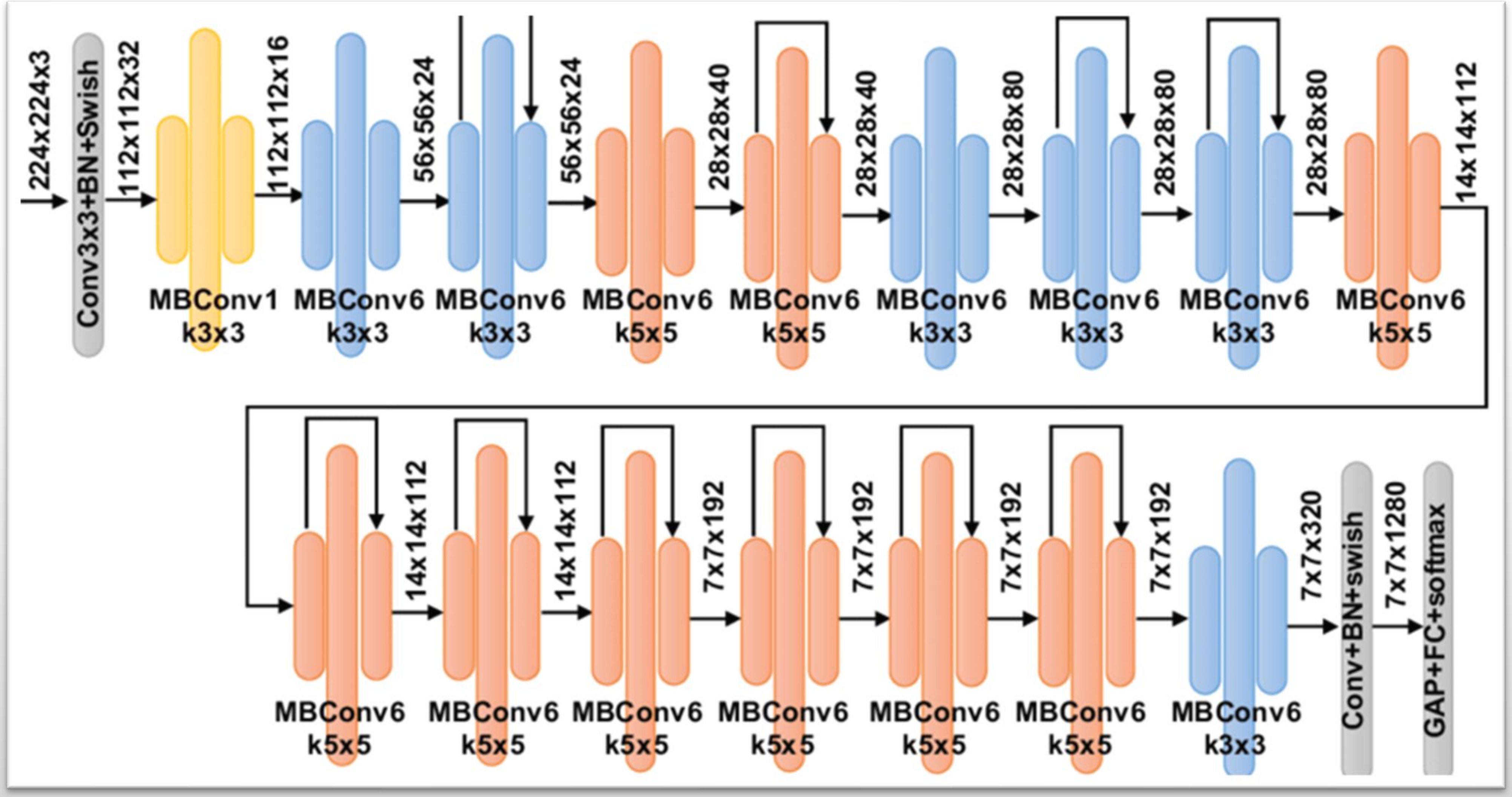
- Computer generated architecture
- Order of magnitude better than human architectures
- Pretrained on ImageNet dataset

Used Efficient net architectures:

Name	Weights	Imagenet Acc	#Parameters
EfficientNetB0	noisy-student	78.8%	5.3M
EfficientNetB6	noisy-student	86.4%	43.3M
EfficientNetV2M	imagenet	85.3%	54.4M



EFFICIENT NET V. B0



EFFICIENT NET TRAINING

- Fine tuning the network on GPU

Parameters:

- Batch size => 5~30
- Epochs =>15
- Regularization =>
  - L2
  - Dropout
  - Label smoothing

Reached accuracies on development set:

• Flowers dataset	93.4%
• Alzheimer dataset	96.3%
• Fruits dataset	99.7%

REDUCING DIMENSION TO 2D

PCA = Principal Component Analysis

- Returns new basis for the data, where each component tries to maximize the variance in its direction
- Eigenvectors of the covariance matrix
- Preserving relative distances

- **t-SNE = t – distributed Stochastic Neighbor Embedding**
  - "The similarity of datapoint x_j to datapoint x_i is the conditional probability, p_(j|i), that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a t-distribution centered at x_j"
  - Variance – set by perplexity parameter
  - Minimize KL - divergence
  - Preserving local neighborhood

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

REDUCING DIMENSION TO 2D

UMAP = Uniform Manifold Approximation and Projection

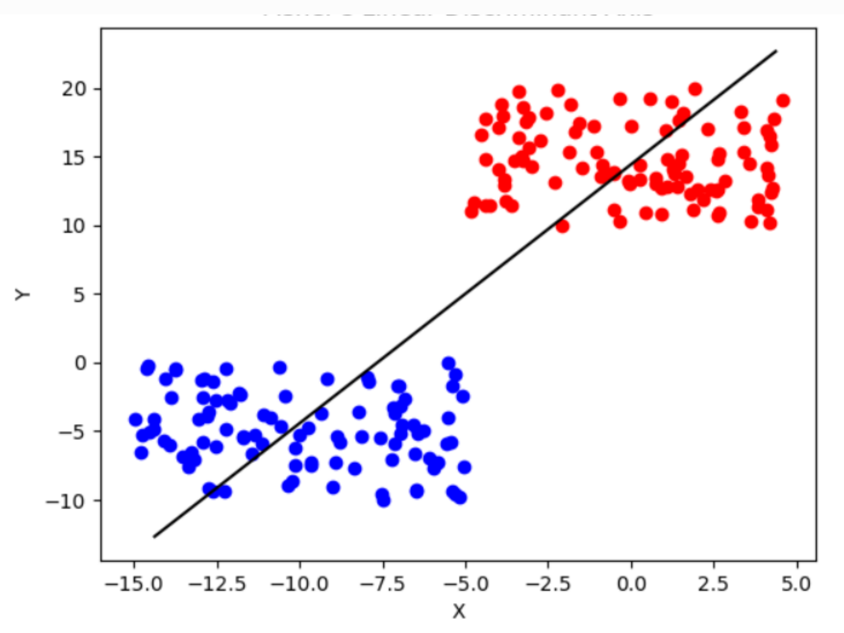
- Inspired by t-SNE
- Non-parametric
- Uses distribution *q*
- Minimizes cross entropy $H(p,q) = -\sum_x p(x) \log q(x)$

$$q_{ij}^{UMAP} = \left(1 + a\left\|z_i - z_j\right\|^{2b}\right)^{-1}$$

LDA – Linear Discriminant Analysis

- Estimate class means μ from training set and class covariances Σ
- Then estimate the discriminator *w*
- Project the data on *w*

$$\vec{w} \propto (\Sigma_0 + \Sigma_1)^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$$



Other feature selection techniques - returning subset of features

- Poor results (there are probably no 2 dominat attributes that would expain all 1280 features)

EVALUATING THE QUALITY OF DIMENSIONALITY REDUCTION

Difficult, many different metrics:

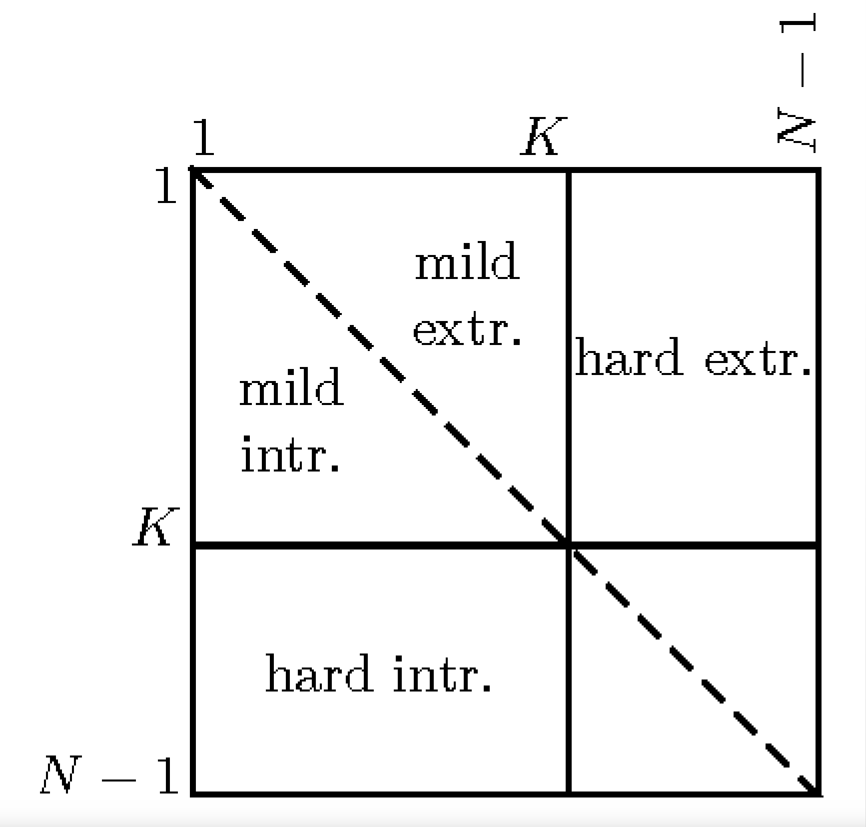
Year	Name of the measure	Criterion
1962	Sheppard Diagram (SD)	Global
1964	Kruskal Stress Measure (S)	Global
1969	Sammon Stress (S_S)	Global
1988	Spearman's Rho (S_R)	Local
1992	Topological Product (T_{Pr})	Local
1997	Topological Function (T_F)	Local
2000	Residual Variance (R_V)	Global
2000	König's Measure (K_M)	Local
2001	Trustworthiness & Continuity (T&C)	Local
2003	Classification error rate	classification error
2006	Local Continuity Meta-Criterion (Q_k)	Local
2006	Agreement Rate (A_R)/Corrected Agreement Rate (CA_R)	Local
2007	Mean Relative Rank Errors (MRRE)	Local
2009	Procrustes Measure (P_M)/Modified Procrustes Measure (P_{MC})	Local
2009	Co-ranking Matrix (Q)	Local
2011	Global Measure (Q_Y)	Local and global
2011	The Relative Error (R_E)	Global
2012	Normalization independent embedding quality assessment (NIEQA)	Local/global/local&global

5-NN CLASSIFICATION

Classification of the same data by means of **k-nearest neighbor**

KNN errors (%)	Alzheimer dataset	Flower dataset	Fruit dataset
PCA	0.80	9.20	24.12
t-SNE	0.70	2.50	0
UMAP	0.80	2.80	0.02
LDA	0.00	5.30	9.39
Variance thresholding	0.70	2.00	0

CORANKING MATRIX



- Original space:
- $p_{ij} = \left|\{k|d_{ik} < d_{ij}\}\right|$
- Projected space:
- $r_{ij} = \left|\{k|e_{ik} < e_{ij}\}\right|$

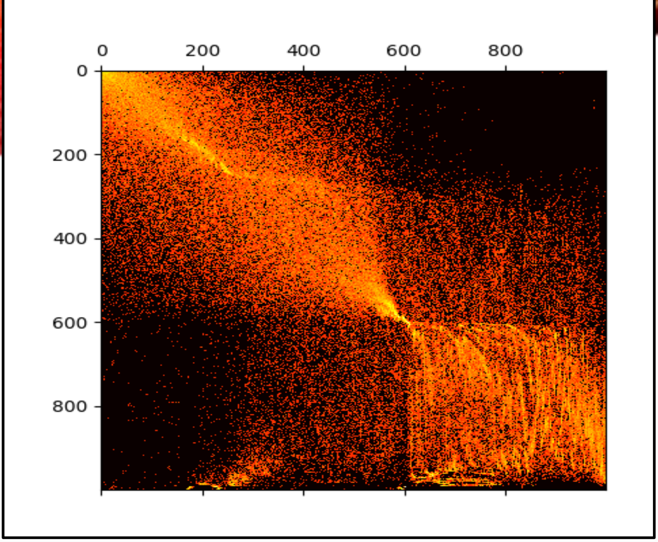
- "number of closer elements to *i* than the distance of *j* to *i*"

Coranking matrix Q:

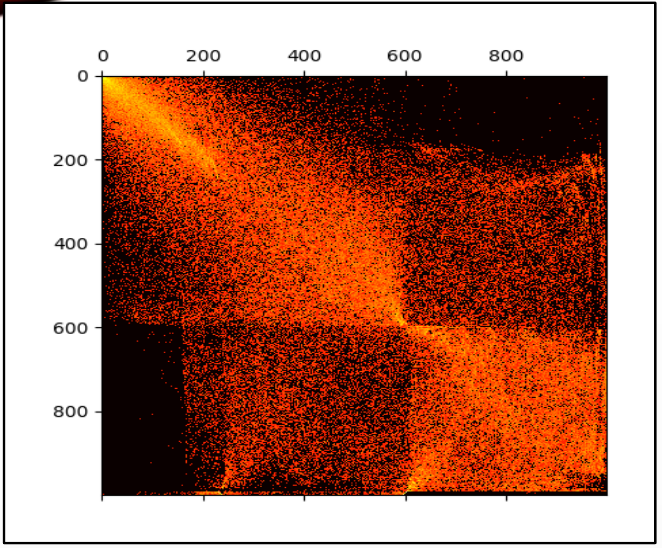
- $q_{ij} = |\{(k,l)|p_{kl} = i \text{ and } r_{kl} = j\}|$
- "number of neighborhoods of size *i* in original space and size *j* in projected space"

VISUALIZING CORANKING MATRIX

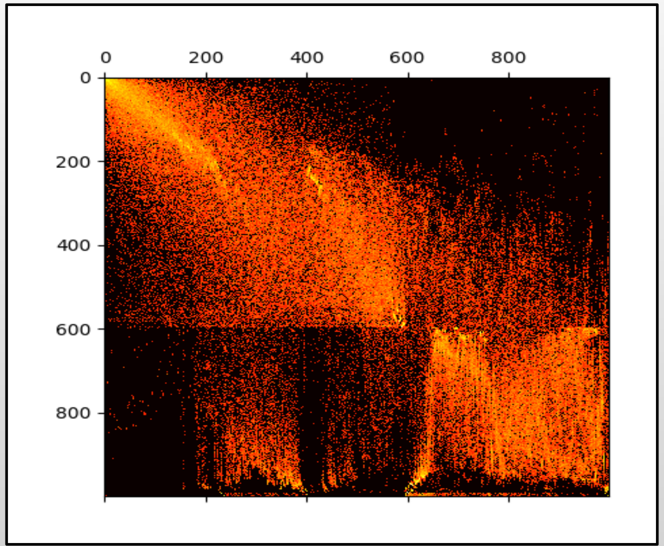
PCA Alzheimer



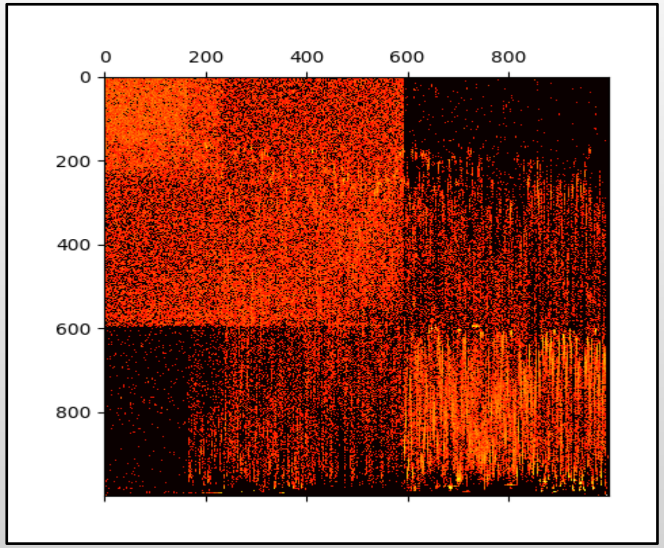
t-SNE Alzheimer



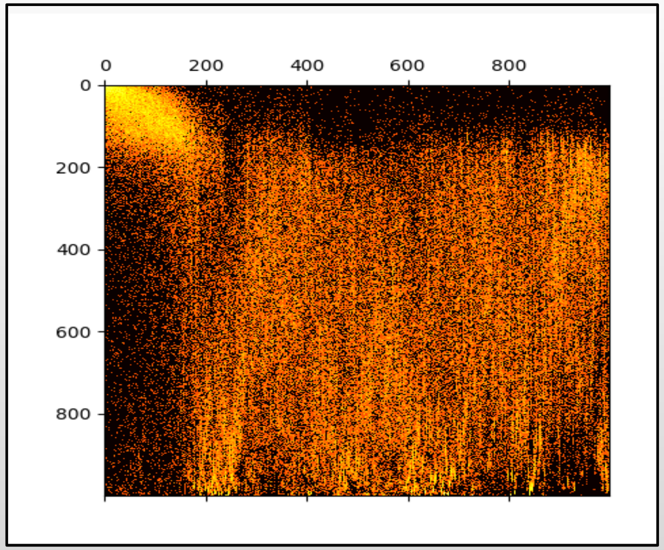
Umap Alzheimer



LDA Alzheimer



Umap Flowers



CORANKING MATRIX EMBEDDINGS

- Q_{NX} curve
  - Q_{NX}(K) ∈ $[0,1]$ (1 means ideal embedding)

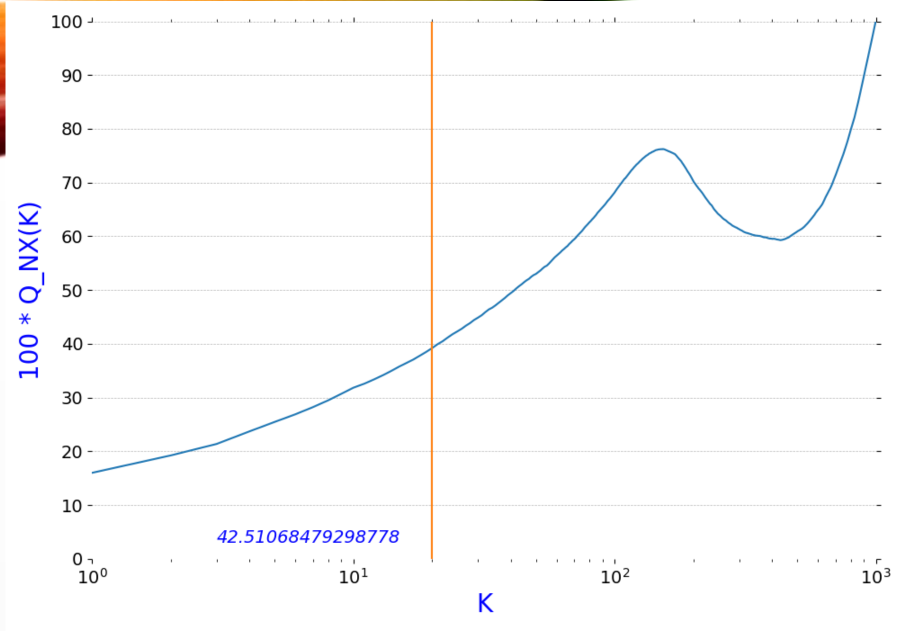
$$Q_{NX}(K) = \frac{1}{Kn} \sum_{k=1}^{K} \sum_{l=1}^{K} Q_{kl}$$

- B_{NX} curve
  - B_{NX}(K) ∈ $[-1,1]$ (1 means extreme intrusion, -1 means extreme extrusion)
  - Subtracts elements outside of a diagonal

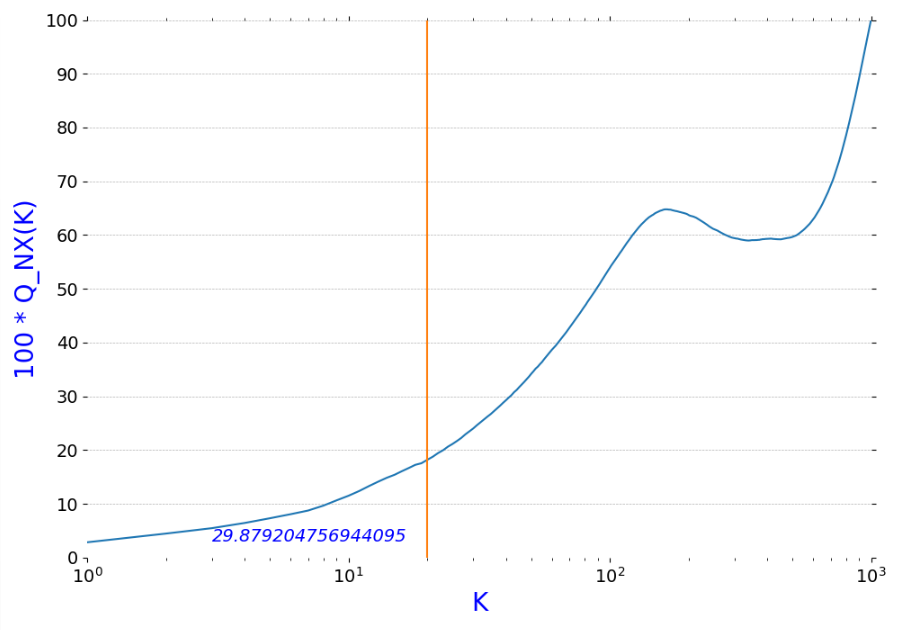
- R_{NX} curve
  - R_{NX}(K) ∈ $[0,1]$ (1 means ideal embedding)
  - Relative improvement of embedding against random embedding

Q_{NX}

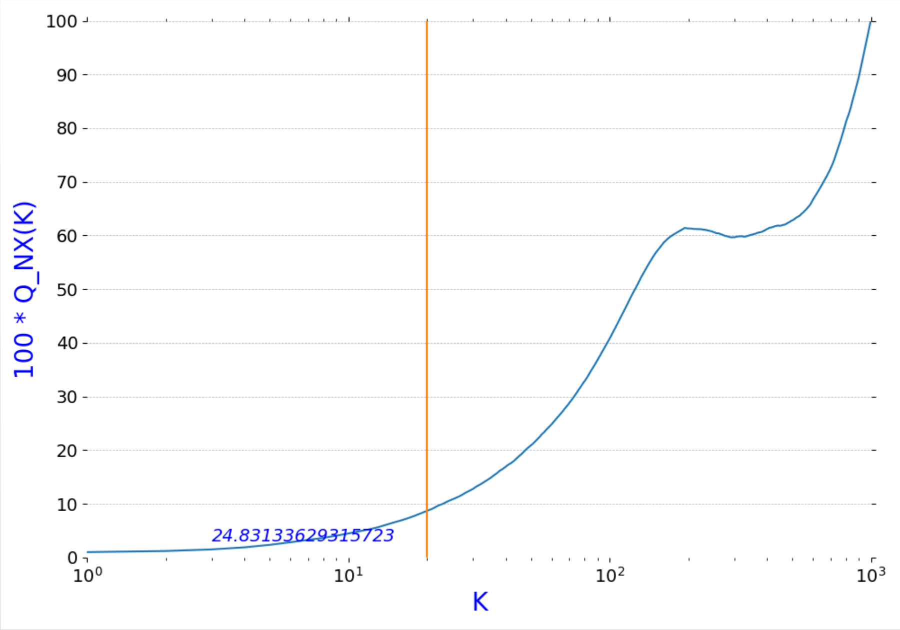
UMAP



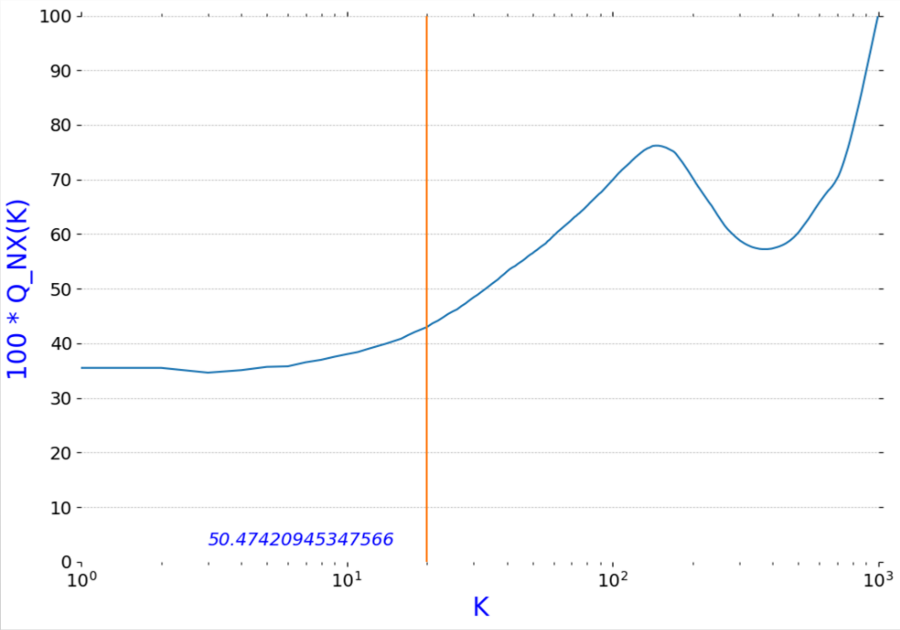
PCA



LDA

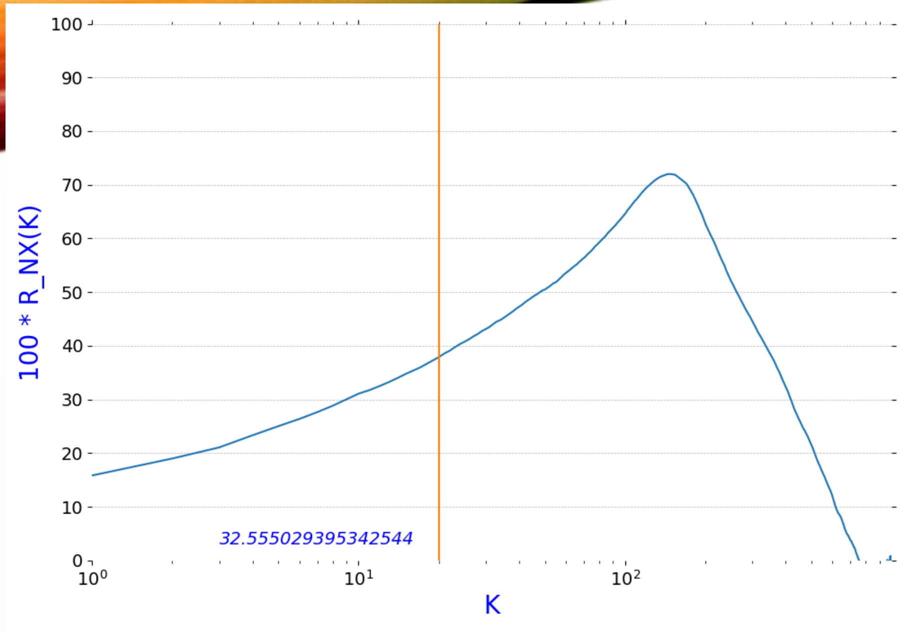


t-SNE

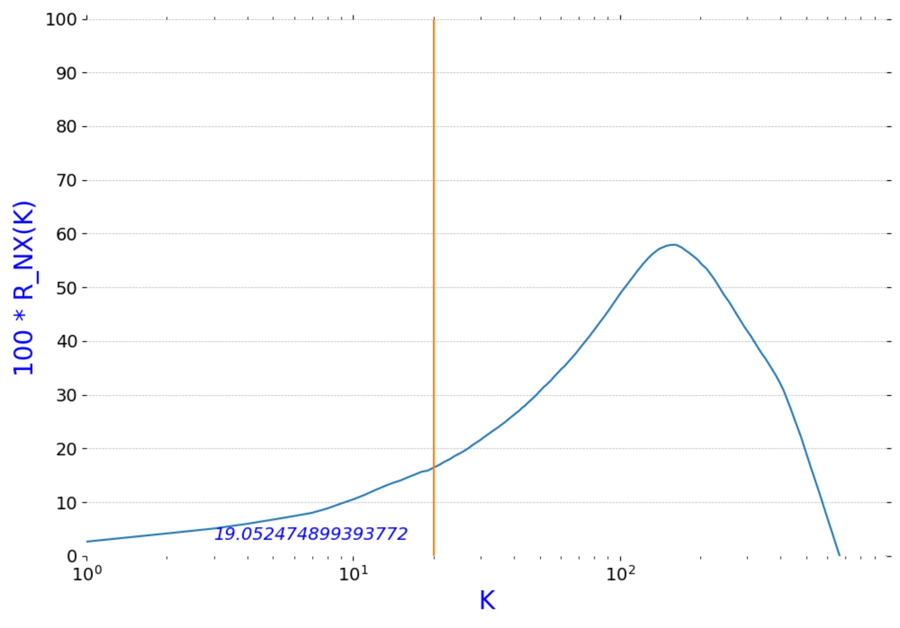


R_{NX}

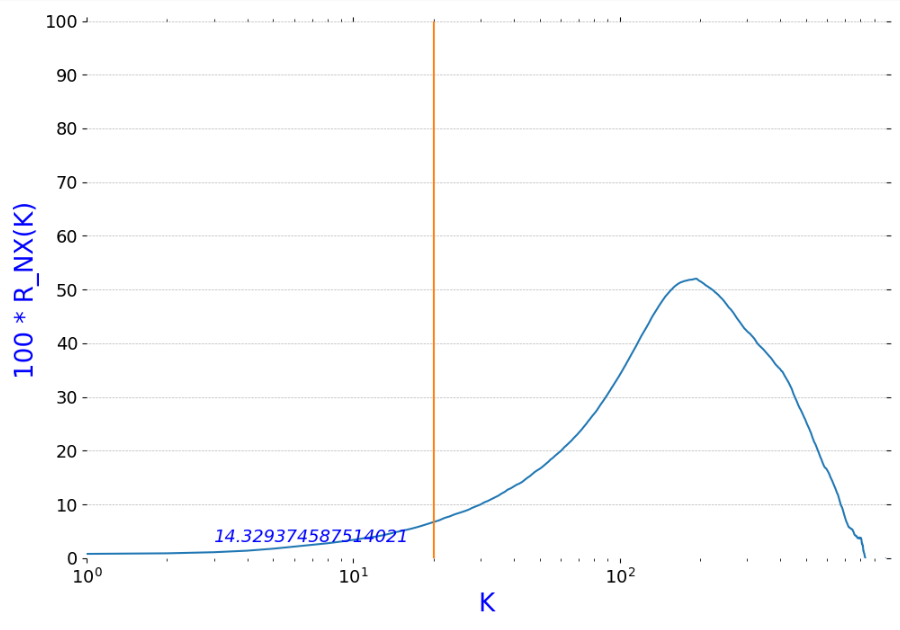
UMAP



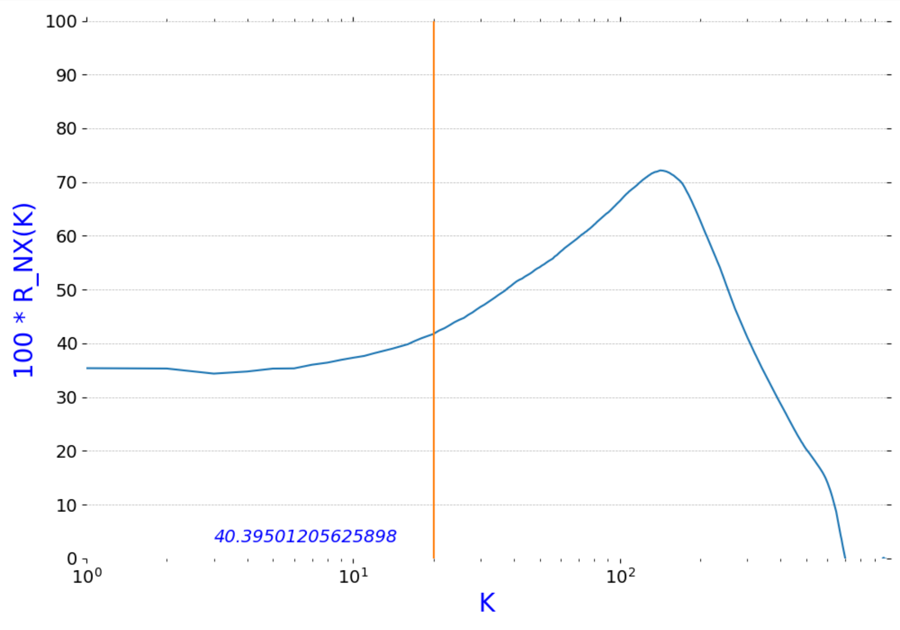
PCA



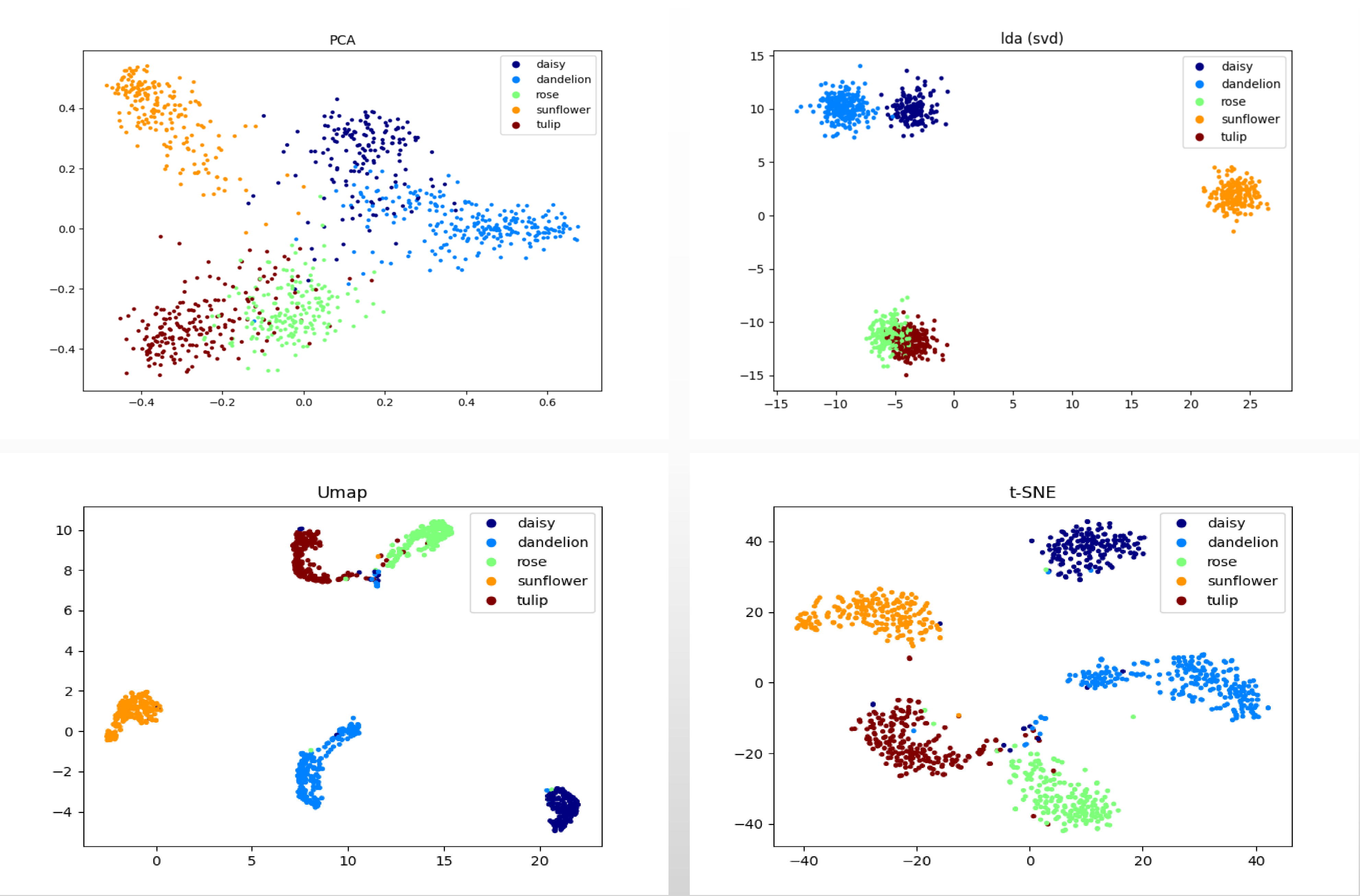
LDA



t-SNE

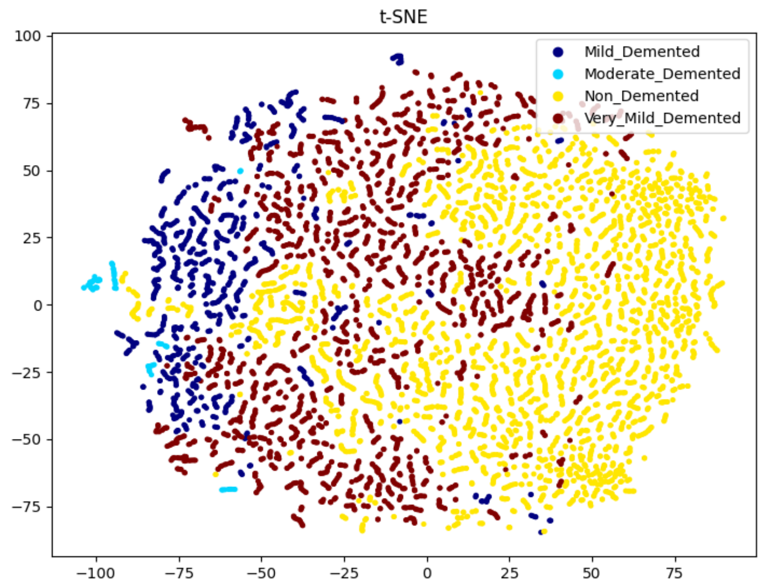


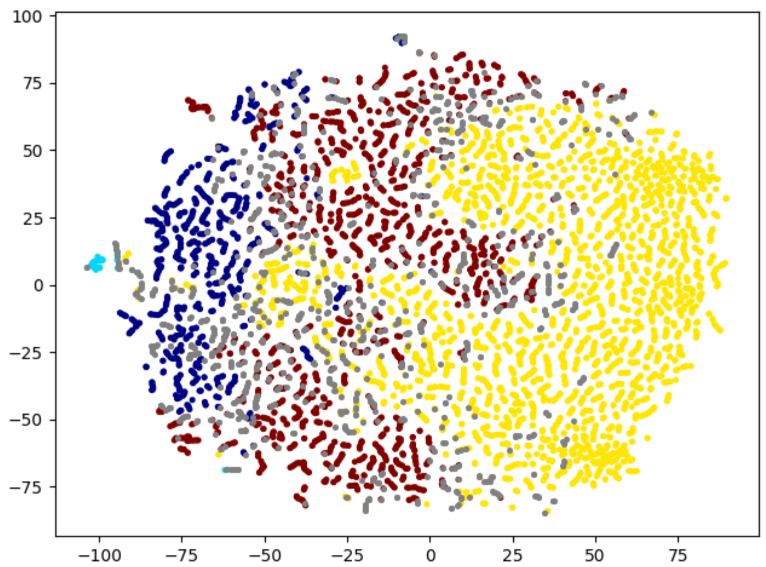
VISUALIZATIONS - FLOWERS

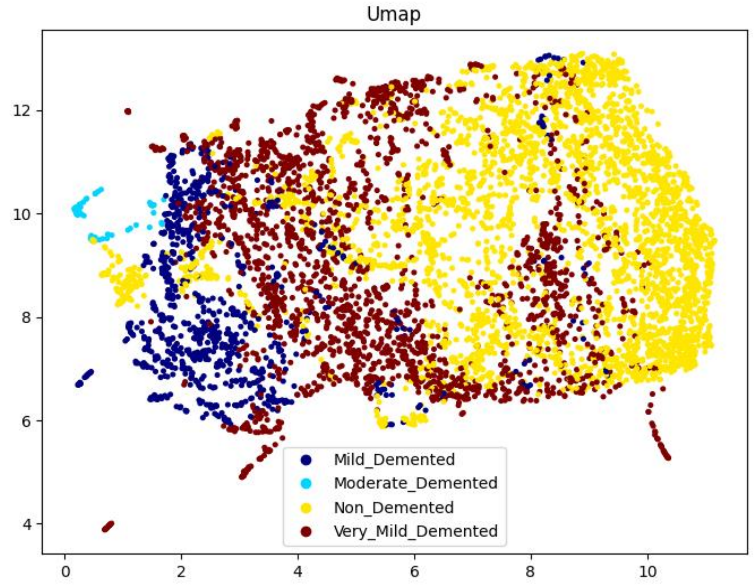

PCA
daisy
dandelion
rose
sunflower
tulip
lda (svd)
daisy
dandelion
rose
sunflower
tulip
Umap
daisy
dandelion
rose
sunflower
tulip
t-SNE
daisy
dandelion
rose
sunflower
tulip

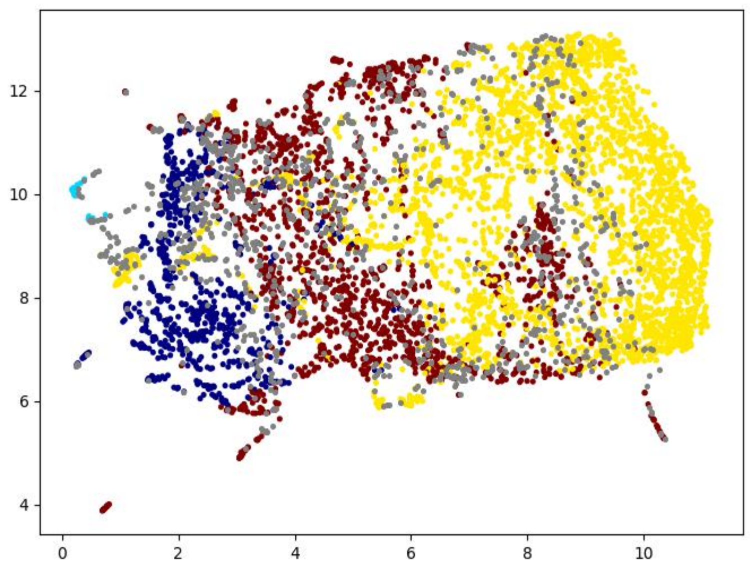
VISUALIZATIONS - ALZHEIMER

Low Accuracy 80%

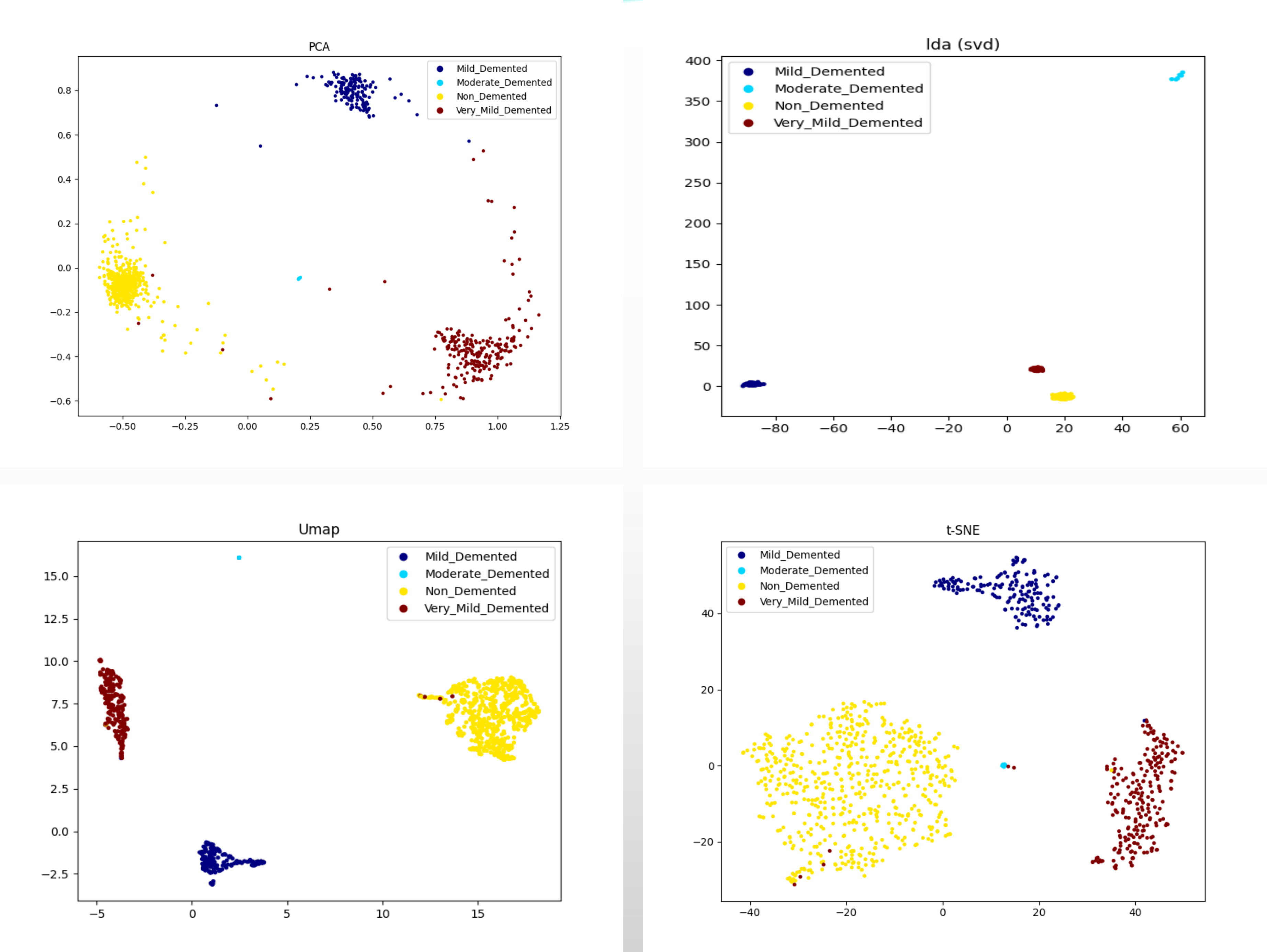




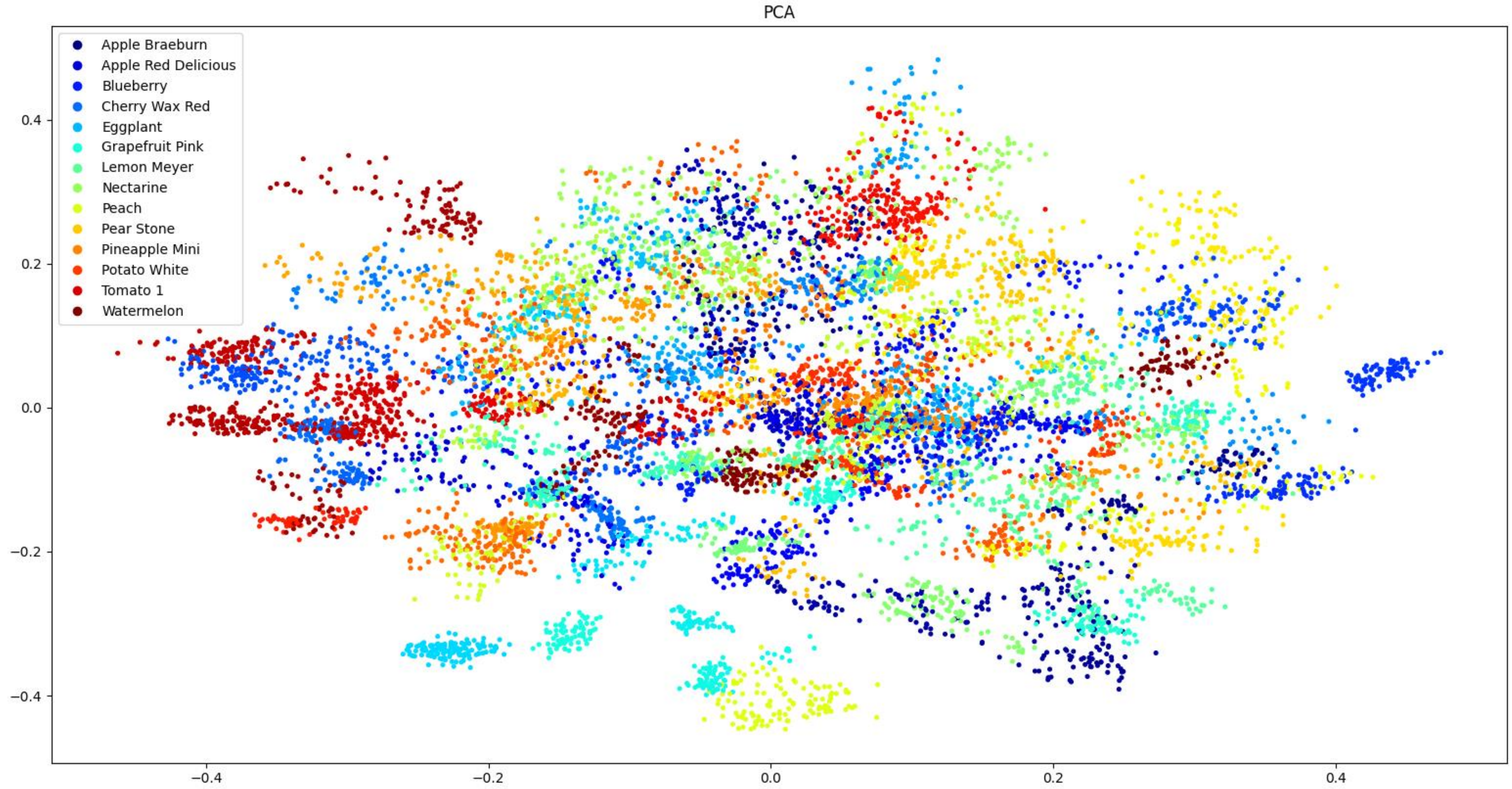


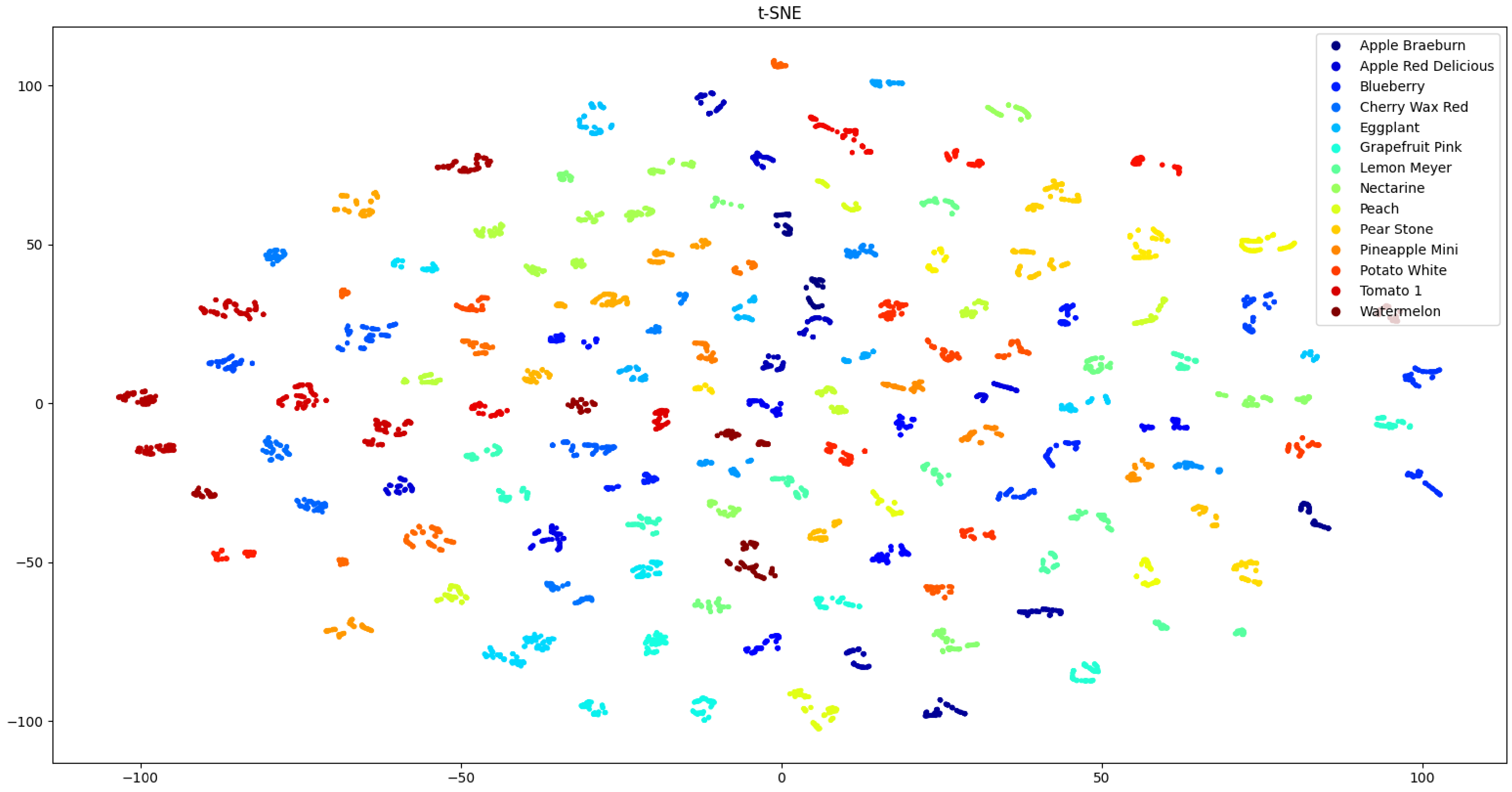


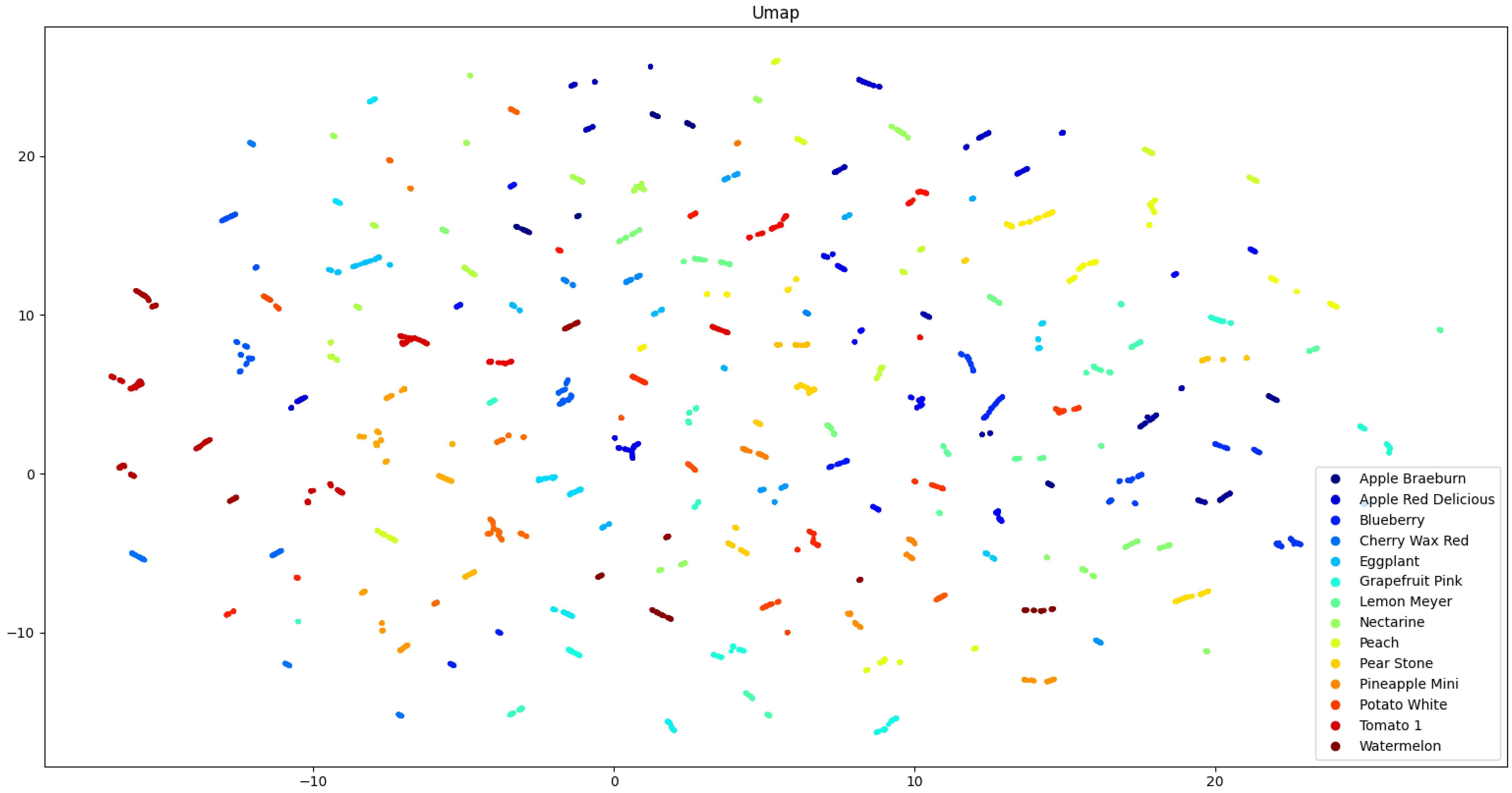
High Accuracy 96%

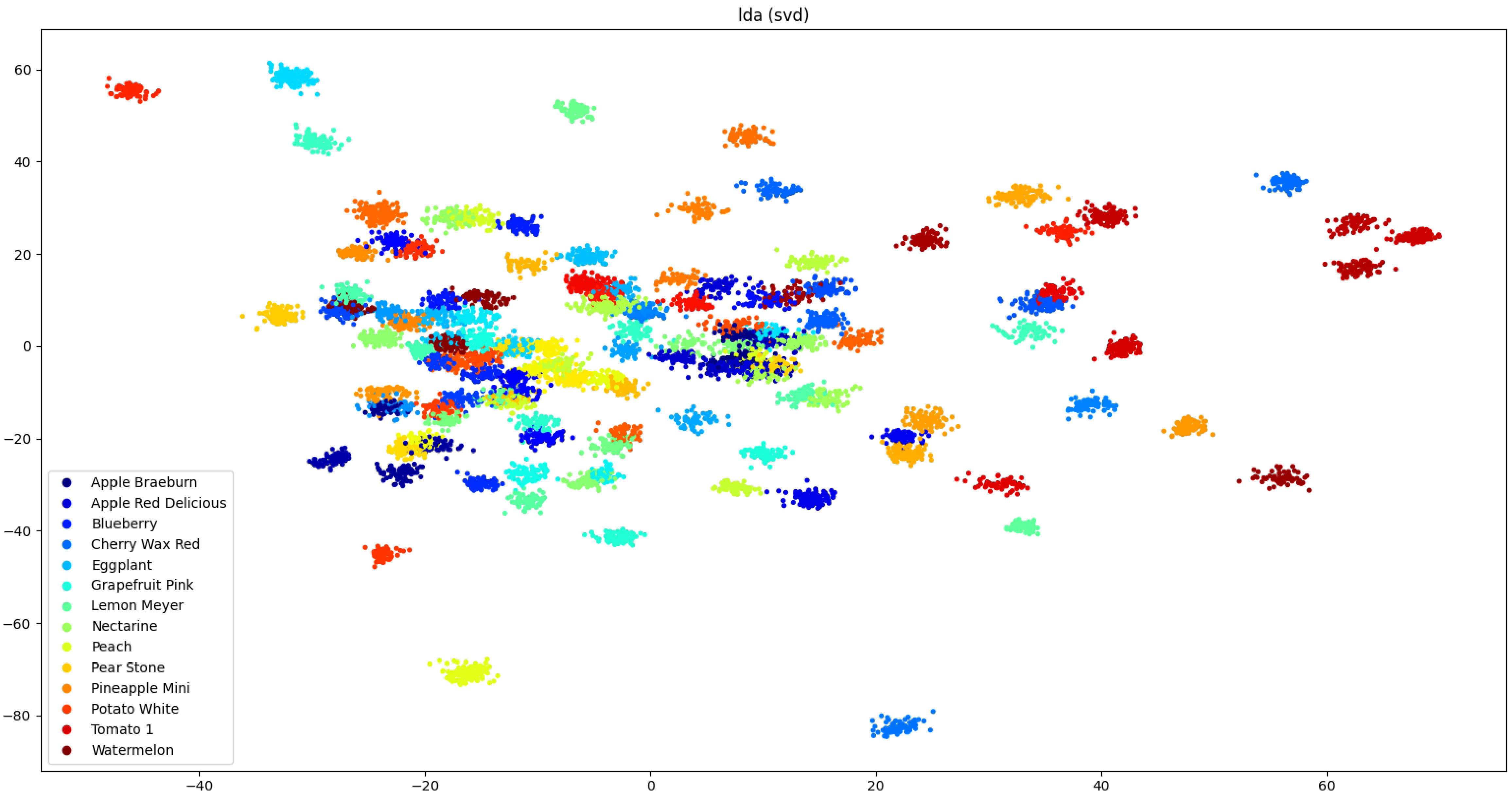


VISUALIZATIONS - FRUITS


PCA
Apple Braeburn
Apple Red Delicious
Blueberry
Cherry Wax Red
Eggplant
Grapefruit Pink
Lemon Meyer
Nectarine
Peach
Pear Stone
Pineapple Mini
Potato White
Tomato 1
Watermelon
0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4


t-SNE
Apple Braeburn
Apple Red Delicious
Blueberry
Cherry Wax Red
Eggplant
Grapefruit Pink
Lemon Meyer
Nectarine
Peach
Pear Stone
Pineapple Mini
Potato White
Tomato 1
Watermelon
100
50
0
−50
−100
−100
−50
0
50
100


Umap
20
10
0
−10
−10
0
10
20
Apple Braeburn
Apple Red Delicious
Blueberry
Cherry Wax Red
Eggplant
Grapefruit Pink
Lemon Meyer
Nectarine
Peach
Pear Stone
Pineapple Mini
Potato White
Tomato 1
Watermelon


lda (svd)
60
40
20
0
−20
−40
−60
−80
−40
−20
0
20
40
60
Apple Braeburn
Apple Red Delicious
Blueberry
Cherry Wax Red
Eggplant
Grapefruit Pink
Lemon Meyer
Nectarine
Peach
Pear Stone
Pineapple Mini
Potato White
Tomato 1
Watermelon

RESOURCES

- https://www.sciencedirect.com/science/article/pii/S0925231209000101
- https://www.sciencedirect.com/science/article/pii/S0167865510001364
- https://www.researchgate.net/figure/The-EffecientNet-B0-general-architecture_fig2_348470984
- https://www.kaggle.com/datasets/moltean/fruits
- https://www.kaggle.com/datasets/alxmamaev/flowers-recognition
- https://www.kaggle.com/datasets/sachinkumar413/alzheimer-mri-dataset

THANK YOU FOR YOUR ATTENTION