Analysis Report Group Project

Group 6

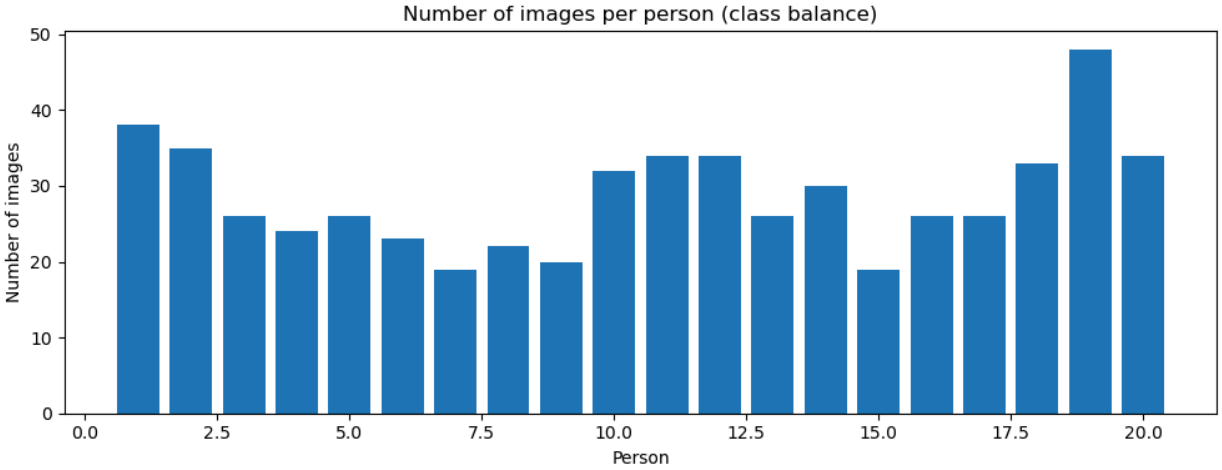
1. Data Preparation

The analysis starts by loading the UMIST Face Database (umist_cropped.mat) and exploring its structure. This dataset contains 575 grayscale images of 20 different individuals, where each image has a resolution of 112 × 92 pixels (shape (112, 92)). Each image is flattened into a 10 304-dimensional vector to facilitate matrix operations and machine-learning workflows. Labels identifying each person (from 0 to 19) are attached to these feature vectors, forming a Pandas DataFrame.

Class Balance Overview:

- Total images: **575**
- Total persons: **20**
- Average images per person: **28.75**
- Standard deviation: **≈ 7.12**
- Minimum per person: **19**, Maximum: **48**

This image summarizes how many images each person contributes:

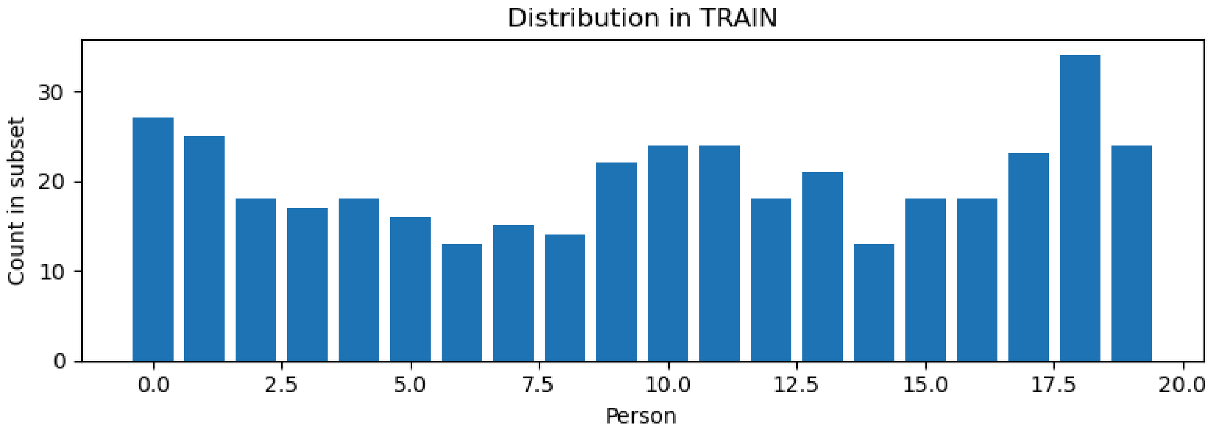


2. Data Splitting

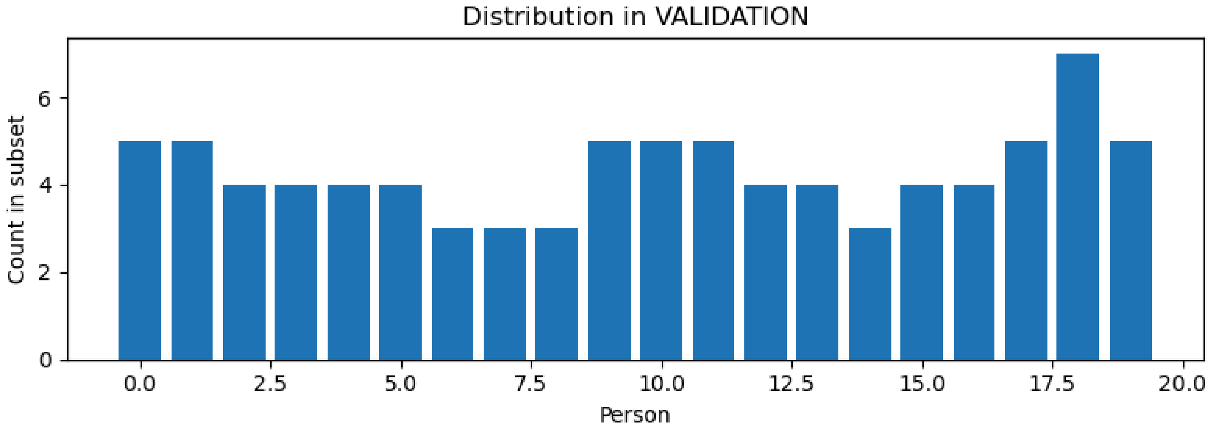
To evaluate machine-learning models properly, the dataset is divided into training, validation, and test sets. A stratified split is performed with proportions 70 % (train), 15 % (validation), and 15 % (test). Stratified sampling ensures that the class distribution within each subset reflects the overall dataset, preventing bias toward classes with more images and enabling fair evaluation.

After splitting:

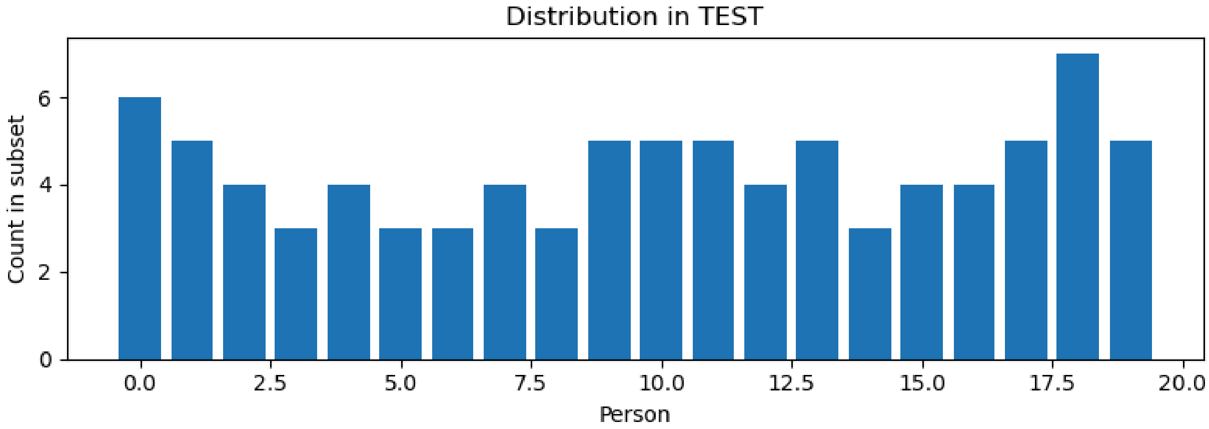
- **Training set**: 402 images



- **Validation set**: 86 images



- **Test set**: 87 images



Each split maintains similar class proportions to the full dataset.

Normalization:

Images are standardized using StandardScaler. The scaler is fitted **only on the training data** and then applied to the validation and test sets to prevent data leakage. Normalization centers each feature (pixel) at zero and scales to unit variance, improving convergence and ensuring that no feature dominates due to its scale.

3. Dimensionality Reduction

High-dimensional image data (10 304 features) is challenging for modelling. Two techniques are explored: Principal Component Analysis (PCA) and an Autoencoder. Both aim to compress the images into a lower-dimensional representation while retaining important information.

3.1 Principal Component Analysis (PCA)

PCA is a linear method that projects data onto orthogonal axes (principal components) ranked by variance captured. The cumulative explained variance helps determine how many components to keep.

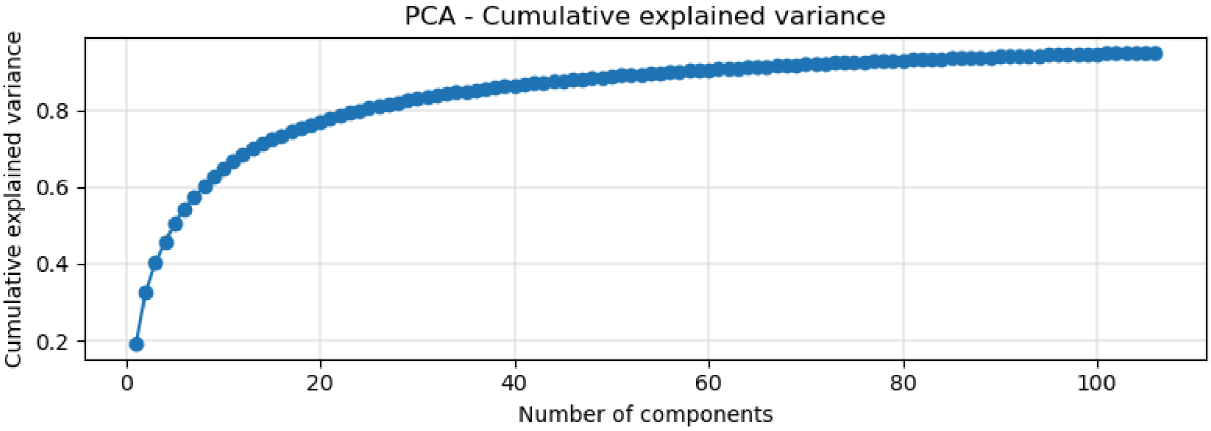
- **Variance threshold**: 95 %
- **Components selected**: **106**
- **Cumulative explained variance**: **95.01 %**

Shapes of the reduced datasets:

- Training set: (402, 106)
- Validation set: (86, 106)
- Test set: (87, 106)

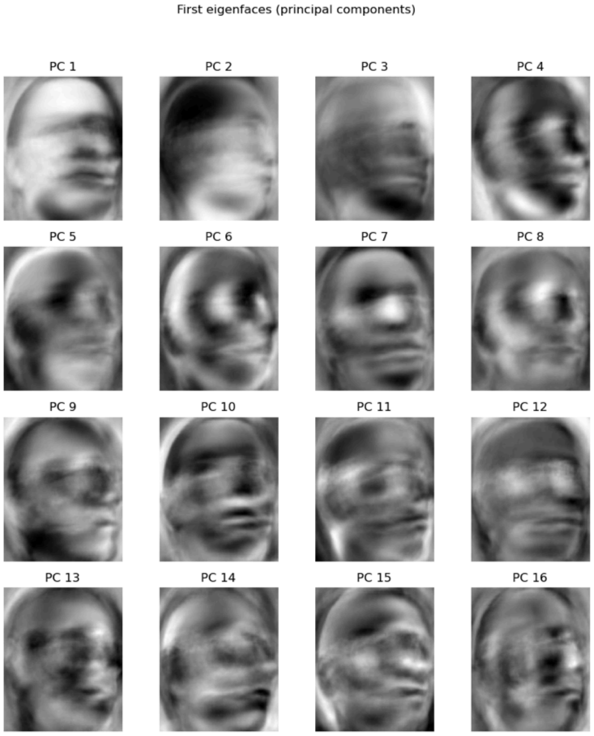
PCA compresses each image from 10 304 features to 106 without significantly losing information. It is a powerful baseline because of its simplicity and efficiency.

Cumulative explained variance plot



- This plot illustrates that the curve flattens after ~100 components, confirming that 95 % of the variance is captured with 106 components.

Eigenfaces



- The “eigenfaces” are the principal components reshaped back into 112 × 92 images. They appear as blurry ghost-like patterns because each captures a global variation across all faces (e.g., lighting, pose)

3.2 Autoencoder

An autoencoder is a neural network that learns a non-linear compressed representation. The network architecture used here:

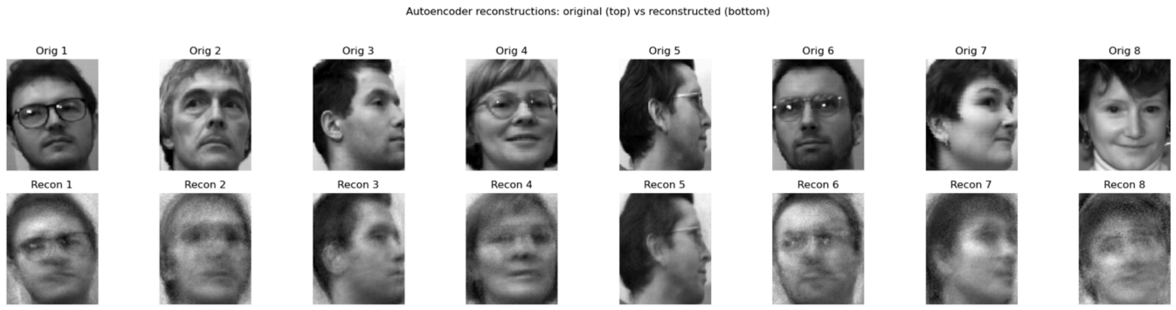
- **Input**: 10 304 features
- **Encoder**: Dense layer with 512 neurons → Dense layer with 64 neurons (latent space)
- **Decoder**: Dense layer with 512 neurons → Dense layer with 10 304 outputs
- **Loss**: Mean Squared Error (reconstruction error)
- **Optimizer**: Adam (learning rate 0.001)
- **Epochs**: 20

Training results:

Metric	Final value
Training MSE	0.1038
Validation MSE	0.1997

The training and validation losses decrease steadily over the epochs, indicating the network is learning to reconstruct faces effectively. The final validation MSE (0.1997) is higher than the training MSE (0.1038), as expected, because the model must generalize to unseen data.

Reconstruction examples:



The reconstructed faces are recognizable but blurrier, indicating that the autoencoder preserves overall facial features while smoothing out fine details. This suggests that the 64-dimensional latent space captures non-linear, high-level information not contained in linear PCA.