# Data Mining CS-535

Project 1

Sarthak Zende B01035919

## Questions:

1. You are asked to implement two specific reduction techniques: PCA and DCT, and apply them, respectively, to the three datasets. Report the resulting dimensionalities for all the scenarios (40 pts.).

Answer :

### Implementation of PCA :

| | Dataset | PCA Dimensions | Original Dimensions | Reconstruction Error |
|---|---|---|---|---|
| 0 | Table1.txt | 17 | 62 | 0.0454 |
| 1 | Table2.txt | 25 | 62 | 0.0489 |
| 2 | Table3.txt | 38 | 62 | 0.0663 |

PCA was implemented by first normalizing the datasets so that each attribute contributed equally. The eigenvalues and eigenvectors were extracted by computing the covariance matrix of the standardized data. These were ranked according to the magnitude of the eigenvalues, which represent the variance that each component captures. A cumulative variance ratio was determined, and a dynamic threshold was set to keep components that accounted for 95% of the total variance, allowing principal components to be dynamically selected based on the intrinsic dimensionality of each dataset. This technique ensured maximal information retention while minimizing dimensionality.

The original and PCA-reconstructed images are shown below to demonstrate the successful dimensionality reduction. We are able to recreate the image while keeping most of the information with lesser dimensions.

**Table1:**

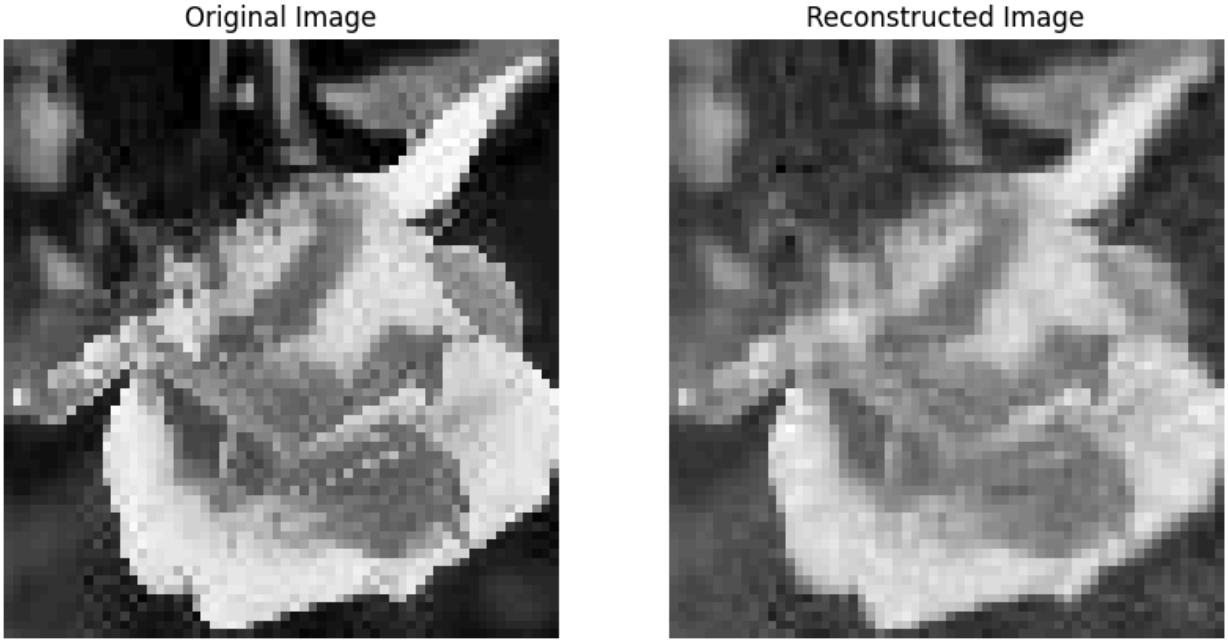



**Table2:**

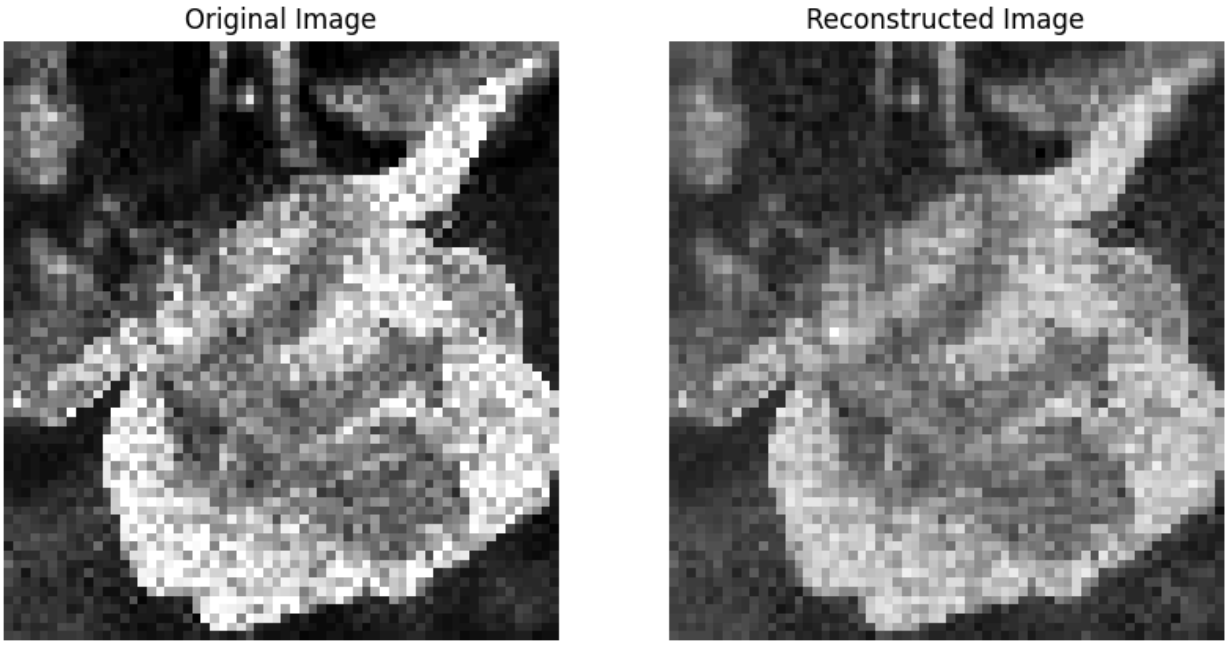
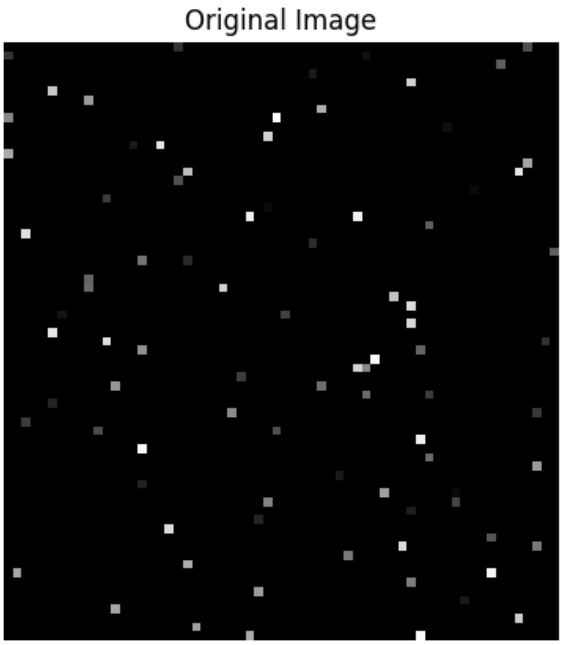
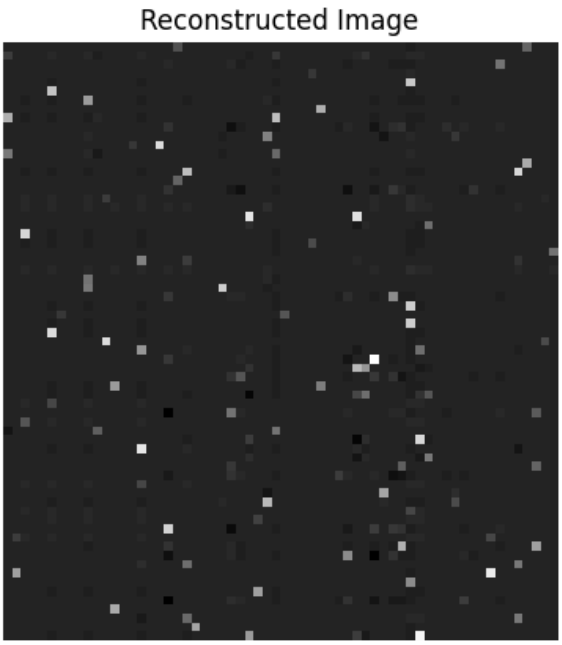



**Table3:**




## Implementation of DCT :

| | Dataset | DCT Dimensions | Original Dimensions | Reconstruction Error |
|---|---|---|---|---|
| 0 | Table1.txt | 17 | 62 | 0.0044 |
| 1 | Table2.txt | 20 | 62 | 0.0064 |
| 2 | Table3.txt | 47 | 62 | 0.0002 |

In the DCT approach, I used a 1D discrete cosine transform along the datasets' attribute axis. The significant coefficients were kept by counting the number of coefficients that combined for 95% of the total energy of the transform coefficients. This method is intended at energy compression, in which the majority of the signal information is compressed in the fewest available coefficients. The less significant coefficients were wiped out, leaving only the most informative parts of the data for reconstruction.

The original and DCT-reconstructed images are shown below to demonstrate the successful dimensionality reduction. We are able to recreate the image while keeping most of the information with lesser dimensions.

**Table1:**

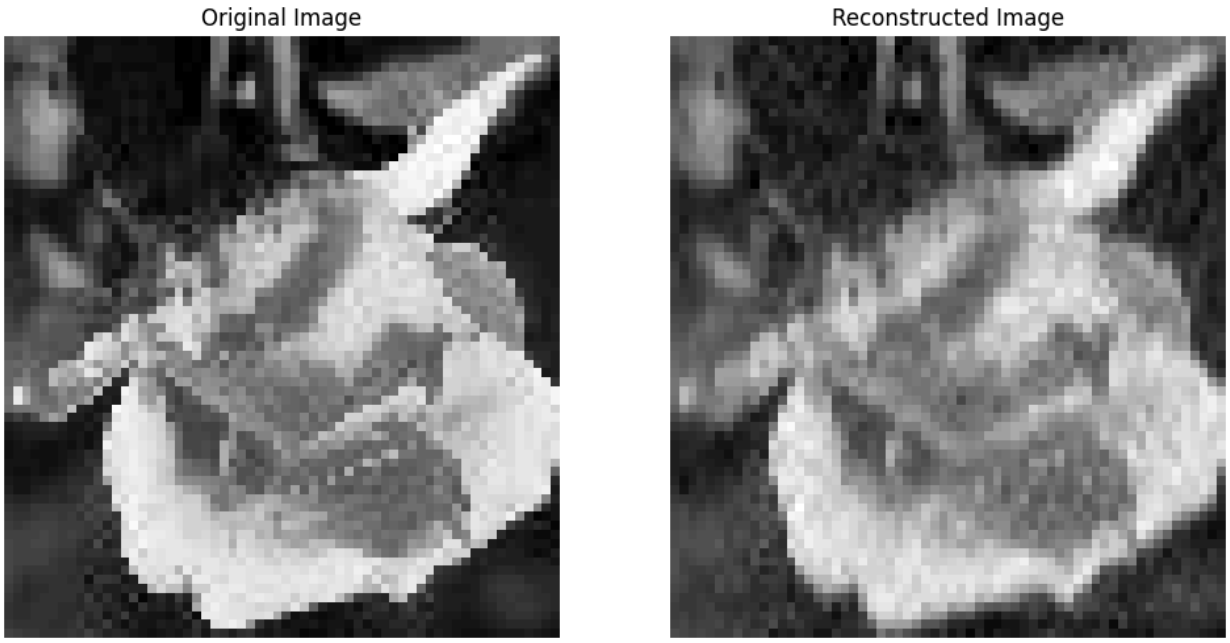



**Table2:**

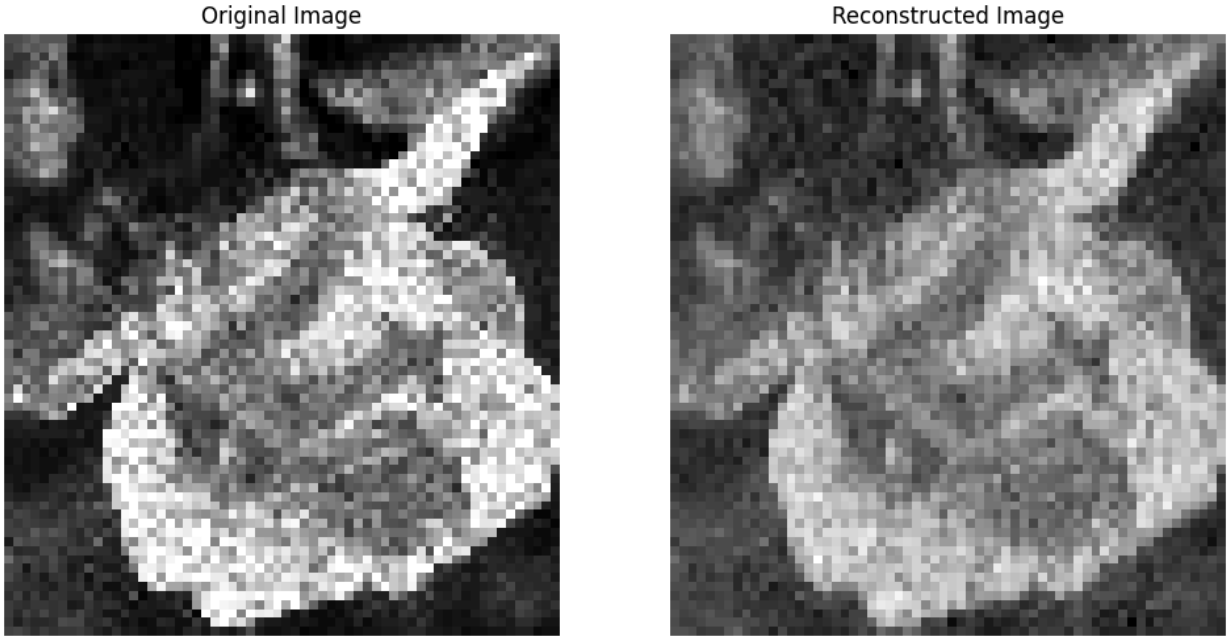
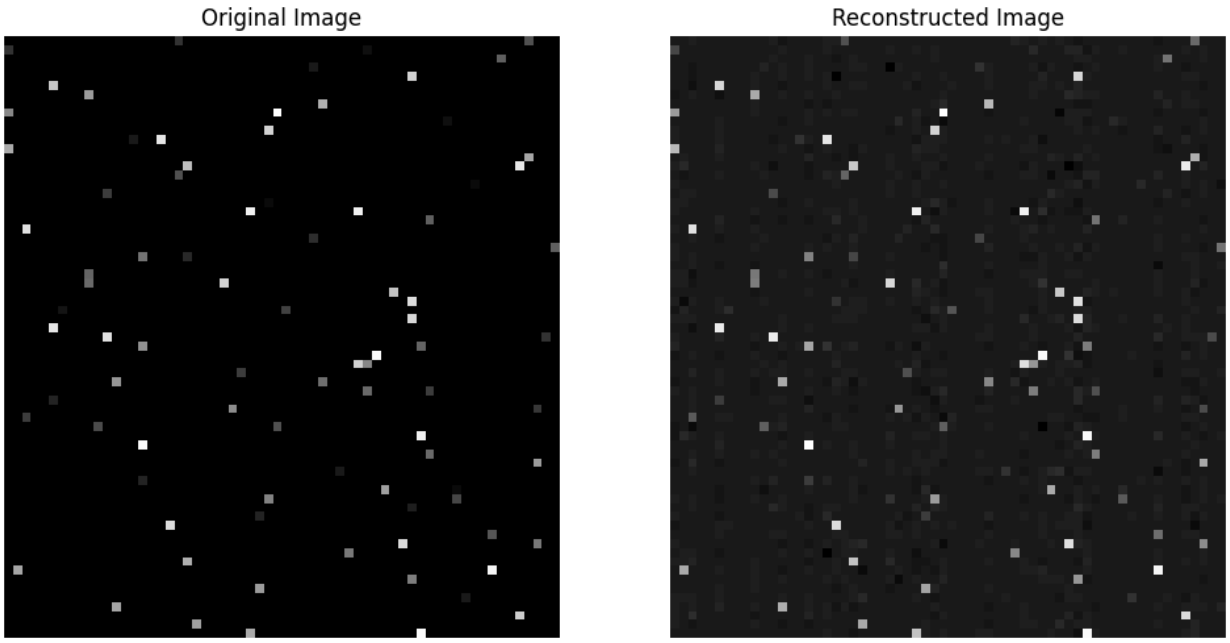


**Table3:**




2. Surf the literature to identify one more dimensionality reduction method and either implement it or directly use the code from the Internet to apply to the three datasets and report the resulting dimensionalities (30 pts.).

Answer:

## Implementation of SVD

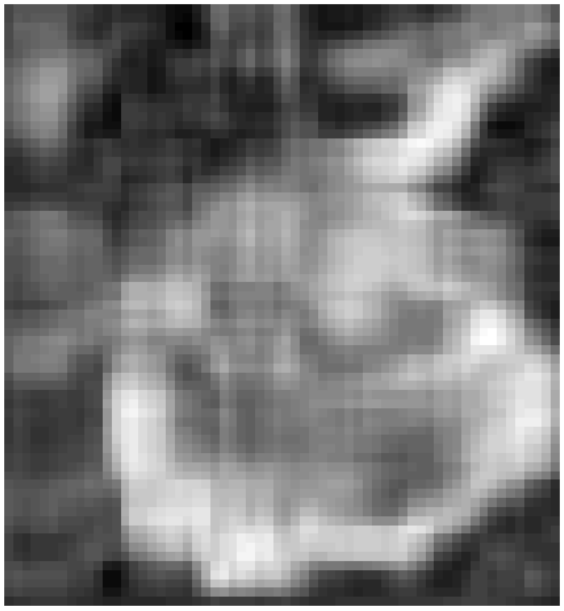
| | Dataset | SVD Dimensions | Original Dimensions | Reconstruction Error |
|---|---|---|---|---|
| 0 | Table1.txt | 6 | 62 | 0.0128 |
| 1 | Table2.txt | 10 | 62 | 0.0128 |
| 2 | Table3.txt | 29 | 62 | 0.0003 |

For the 3rd Dimensionality reduction method, I chose Singular Value Decomposition (SVD) because it breaks down the data into a series of orthogonal components and their corresponding singular values. SVD components were retained based on a cumulative energy threshold of 95%, much like in PCA. By discarding singular values below this cutoff, dimensionality was effectively decreased but the fundamental structure of the data was maintained. Since this method does not rely on the covariance matrix, it can handle zero entries more effectively, making it especially useful for sparse datasets.
The original and DCT-reconstructed images are shown below to demonstrate the successful dimensionality reduction. We are able to recreate the image while keeping most of the information with lesser dimensions.

## Table1:

Original Image - Table1

Reconstructed Image - Table1

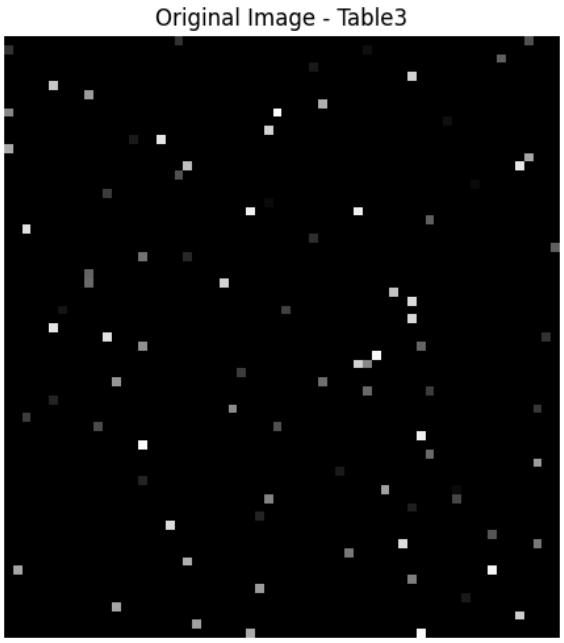


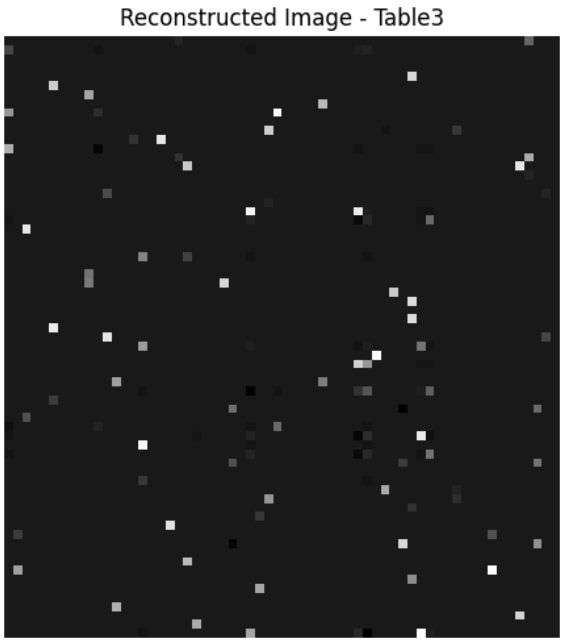
## Table2:

Original Image - Table2

Reconstructed Image - Table2

**Table3:**




3. Discuss the resulting dimensionalities (e.g., are they consistent for all the methods? If they are, why? If they are not, why?) (30 pts.)

Answer :

The resulting dimensionalities are not consistent across methods due to the distinct mechanisms each employs in capturing data variance. For dense datasets, PCA and DCT prioritize variance and energy compaction differently, which results in somewhat different reduced dimensions (Tables 1 and 2). In contrast, DCT and SVD demonstrated a larger dimensionality reduction, demonstrating its effectiveness in sparse data representation, while PCA kept more dimensions to capture variance dispersed across original attributes for the sparse dataset (Table 3). Among all datasets, SVD had the greatest variation in dimensional reduction, indicating that it is sensitive to the underlying structure of the data. This discrepancy emphasizes how crucial it is to select an approach that minimizes dimensionality while preserving reconstruction quality, one that is in line with the features of the data.

The reconstructed images from reduced dimensions using these methods demonstrate that dimensionality reduction can effectively capture the core information of datasets. Although the reconstructions are not perfect, the low reconstruction errors indicate a good balance between dimensionality reduction and information retention.