**Assignment 3: Hierarchical Clustering**

**Analysis Report**

In this assignment, we initially trained an SVM classifier on the original dataset, achieving impressive results with cross-validation scores ranging from 0.875 to 0.9375, a mean cross-validation accuracy of 91.25%, and a validation score of 95%. These results indicate that the SVM performed exceptionally well on the original dataset.

Next, we explored dimensionality reduction using Agglomerative Hierarchical Clustering (AHC) with three different similarity measures: Euclidean, Minkowski, and Cosine. The optimal number of clusters determined by the silhouette score was 77 for both Euclidean and Minkowski measures, and 2 for the Cosine measure. However, the performance of the SVM classifier on the transformed datasets was significantly lower. The mean cross-validation scores were 77.91% for Euclidean, 67.08% for Minkowski, and 2.91% for Cosine. The validation scores were even more concerning, with 2.5% for Euclidean, 3.75% for Minkowski, and 2.5% for Cosine.

These results suggest that the dimensionality reduction using the optimal number of clusters did not capture the complexity of the data as effectively as the original features. The clusters formed by Agglomerative Clustering might not capture the nuanced variations in the data, resulting in clusters that are too broad and not representative of the underlying patterns needed for effective classification. Although we tried agglomerative clustering with three measures, only Euclidean and Minkowski showed better results compared to Cosine, which highlight how the choice of distance measure can significantly impact the clustering results, as each measure emphasizes different aspects of the data. Despite these results, we still did not reach the same level or higher than the mean cross-validation score of the original dataset. Therefore, it is important to recognize that we cannot solely relying on silhouette scores for determining the optimal number of clusters but also considering the specific requirements of the downstream task.

Techniques like Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), or Uniform Manifold Approximation and Projection (UMAP) are often more suitable for image data, as they are designed to preserve the structure of high-dimensional data in a lower-dimensional space more effectively.