# Group 26 Report on

# The Impact of Dimensionality Reduction on Clustering for Image Segmentation on Flood Area Images

* [Raskar Varun Arun (M24CSA026)](mailto:<m24csa026@iitj.ac.in>)
* [Bidisha Sukai (D24CSA002)](mailto:d24csa002@iitj.ac.in)

### 1. Introduction

In recent years, image segmentation has become a crucial task across diverse domains, including medical imaging, remote sensing, and disaster management. Image segmentation aims to partition images into meaningful regions based on pixel characteristics, allowing us to identify specific features such as objects or regions within an image. This report investigates how dimensionality reduction affects the efficiency and accuracy of clustering methods in image segmentation. Image segmentation, a crucial task in image processing, involves partitioning an image into segments to simplify its analysis. Clustering methods, like K-means, are commonly used for this task by grouping pixels with similar characteristics. Dimensionality reduction techniques can simplify data, potentially leading to faster processing times and improved clustering performance. Dimensionality reduction techniques such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) have emerged as effective solutions for reducing computational complexity. By transforming the data into a lower-dimensional space, these techniques preserve essential features while reducing noise, allowing clustering algorithms to perform more effectively. This study investigates the impact of dimensionality reduction on clustering performance, specifically for segmenting flood-affected areas in satellite or drone imagery.

### 2. Motivation

Floods remain one of the most devastating natural disasters, affecting millions worldwide each year. The ability to quickly assess flooded areas and their extent is vital for coordinated disaster response. Automated segmentation of flooded regions in images can assist in creating maps of impacted zones, facilitating decision-making in resource allocation and evacuation planning. Clustering-based segmentation is a promising approach; however, high-dimensional data makes clustering slow and may lead to inaccurate groupings.

Dimensionality reduction can improve clustering by focusing on the most relevant features, speeding up computation without compromising accuracy. PCA, for instance, retains variance in the data, while ICA extracts statistically independent components, which may capture unique aspects of flood regions. Our study specifically addresses these methods to optimize clustering for flood image segmentation.

### 

### 3. Problem Statement

The primary objective of this project is to evaluate the effects of dimensionality reduction on clustering accuracy and efficiency in image segmentation. We aim to determine whether reducing the dimensionality of flood image data enhances the performance of clustering algorithms in terms of computation time and segmentation accuracy. Additionally, the study explores the comparative benefits of PCA and ICA for different clustering algorithms, focusing on identifying the most effective method for real-time disaster response applications.

### 4. Dataset Overview

The dataset comprises flood-area images, accompanied by binary masks indicating regions affected by flooding. This dataset requires segmentation to differentiate flooded and non-flooded areas based on pixel characteristics, which can be identified by clustering. Images are loaded in RGB format, while masks are grayscale for easy binary segmentation. The dataset is preprocessed to standardize dimensions across all images, preparing it for clustering analysis.



fig 1.1: a)original image b)original mask



fig 1.2: a)original image b)original mask

**5. Literature Review**

1. Reddy et al. analyzed PCA and LDA’s impact on classification and clustering algorithms, finding that dimensionality reduction improved computation without major accuracy loss. Their work emphasized PCA’s versatility in reducing redundant features while preserving data variance​.[7]
2. Zhu et al. used a combination of PCA and K-Means for clustering in diabetes diagnosis. Their findings demonstrated that dimensionality reduction could improve clustering accuracy by reducing noise and feature redundancy​.[8]
3. Li et al. examined the role of t-SNE and PCA in clustering high-dimensional image data, underscoring the trade-off between computation and accuracy when using these techniques for large image datasets​.[9]
4. Maaten and Hinton (2008): The study compares PCA and t-SNE for data visualization, with PCA highlighted as effective for global structure retention in data with clear variance patterns.[1]
5. Kekre et al. (2010): The study investigates PCA for dimensionality reduction in remote sensing, showing that PCA can effectively compress spectral data, making segmentation computationally efficient without major accuracy loss.[2]
6. Hyvarinen and Oja (2000): This foundational paper on ICA outlines its application in signal processing, particularly in fields requiring separation of independent sources, a task similar to distinguishing unique regions in image segmentation.[3]
7. Cardoso (1998): This research applies ICA to facial recognition and shows that independent component features improve classification accuracy, underscoring ICA’s potential in extracting unique features in segmentation tasks.[4]
8. Kanungo et al. (2002): Demonstrated that K-Means could effectively segment simple image data, making it suitable for image data with defined clusters like color-based flood areas.[5]
9. Comaniciu and Meer (2002): This study presents Mean Shift’s advantages in segmenting images with non-linear characteristics, showcasing its effectiveness in handling complex textures.[6]

### 6. Approach Used in Code

The repository implements clustering with PCA and ICA for dimensionality reduction, focusing on segmenting flood regions. Major components in the code include:

1. Image Loading: A load\_images function loads a specified number of images from the dataset.
2. Dimensionality Reduction:
   * PCA: Used to retain key variance in pixel features, making clustering more manageable.
   * ICA: An alternative to PCA, applied for different feature extraction properties, capturing independent components in the data.
3. Clustering Algorithms:
   * K-Means: Groups pixels based on color similarity in reduced dimensions.
   * Mean Shift: A density-based clustering approach, less common for image segmentation but potentially effective post-dimension reduction.
   * K-Means and Mean Shift clustering are used on both reduced and original pixel features, aiming to compare performance in segmenting flood-affected areas.
4. Visualization and Evaluation: Segmentation results are visualized with plotly to assess clustering quality across different dimensionality reduction techniques.

### 7. Methodology

In this study, PCA and ICA are used to reduce the dimensionality of pixel data in flood images, followed by clustering algorithms (K-Means and Mean Shift) to segment flood-affected areas.

#### 7.1 Principal Component Analysis (PCA) Methodology

PCA operates by transforming data into a new coordinate system defined by principal components, which are orthogonal to each other. This transformation is computed by calculating the eigenvectors and eigenvalues of the data covariance matrix, where the eigenvectors represent the principal components, and the eigenvalues signify the amount of variance each component captures. For flood image segmentation:

1. **Feature Representation**: Flood images are initially represented in a high-dimensional RGB space. Each pixel becomes a 3-dimensional vector of color intensities.
2. **Variance Maximization**: PCA identifies the directions of maximum variance (principal components) in the data. By projecting pixel features onto these components, PCA reduces dimensionality while preserving information critical for clustering.
3. **Dimensionality Reduction**: By retaining the first few principal components, PCA compresses data into fewer dimensions (e.g., two components for easier clustering), maintaining the structure of flooded and non-flooded regions.

**Effectiveness in Task**: PCA is particularly effective in reducing redundant information in RGB data, retaining the color gradients necessary for distinguishing flood regions while reducing computational complexity.

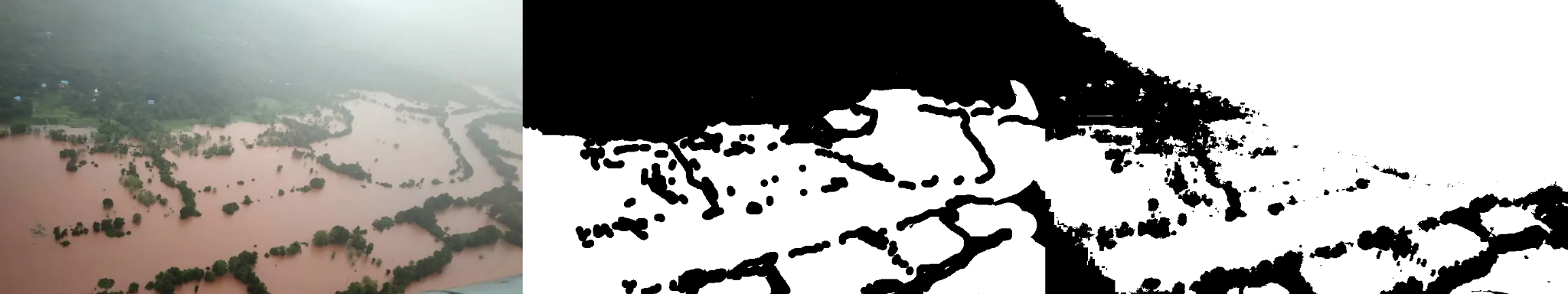


fig 2: a)original image b)original mask c)segmented image

#### 7.2 Independent Component Analysis (ICA) Methodology

ICA assumes that data are a combination of independent sources, unlike PCA, which maximizes variance. For flood images, ICA can isolate features that may not vary linearly but represent unique statistical structures within flooded areas:

**Signal Separation**: ICA identifies statistically independent signals within the data, which could correspond to distinct textures or intensity patterns.

**Application to Flood Segmentation**: Flooded regions may contain patterns (e.g., water reflections) independent of the rest of the image. ICA separates these patterns, potentially enhancing clustering accuracy.

**Dimensionality Reduction**: Similar to PCA, ICA reduces dimensions by isolating independent components, capturing unique image aspects without retaining redundancy in color data.

**Effectiveness in Task**: ICA’s focus on independence rather than variance makes it effective for segmenting regions with unique statistical characteristics, helping to identify flood areas by distinguishing them from other textures or materials.

#### 7.3 K-Means Clustering Methodology

K-Means is a centroid-based clustering algorithm that minimizes intra-cluster variance by assigning each data point to the nearest cluster centroid:

1. **Initialization**: Centroids are initialized within the reduced-dimensional space.
2. **Cluster Assignment**: Each pixel is assigned to the nearest centroid, forming clusters based on color similarity. This is particularly useful in identifying homogeneous regions in flood images.
3. **Iterative Optimization**: The centroids are recalculated iteratively until the assignments stabilize, optimizing cluster homogeneity.

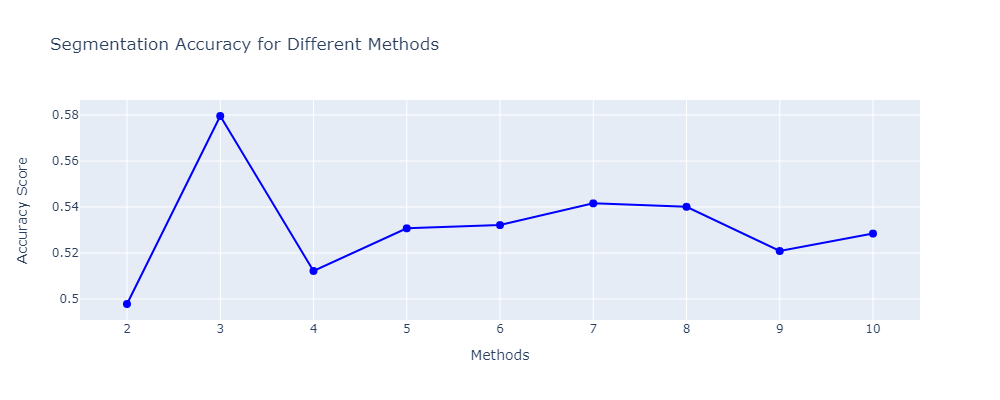


fig 3 Segmentation best k value plot, k=3(best)

**Effectiveness in Task**: K-Means is efficient in handling images with well-defined color clusters. When combined with PCA-reduced data, K-Means effectively segments flooded regions by grouping similarly-colored pixels.

#### 7.4 Mean Shift Clustering Methodology

Mean Shift identifies clusters by shifting points toward the direction of maximum density, which allows for a flexible, density-based clustering that adapts to non-linear patterns:

1. **Bandwidth Estimation**: Mean Shift relies on a bandwidth parameter to determine the cluster size. This parameter is essential for distinguishing regions based on pixel density in flood imagery.
2. **Mode Detection**: Each pixel moves towards areas of high density, effectively clustering pixels in dense regions, which are common in flood images.
3. **Cluster Formation**: As each pixel converges to the densest regions, clusters form around these modes, enabling the algorithm to adapt to non-linear distributions and varied intensity patterns in flood images.

**Effectiveness in Task**: Mean Shift’s adaptability to density makes it effective for complex flood images with uneven pixel intensity distributions, capturing variations more effectively than K-Means in certain cases.

### 7.5 Evaluation Metrics

1. **Computational Efficiency**: The execution time for clustering was recorded both with and without PCA to assess efficiency gains.  
2. **Segmentation Accuracy:** Visual similarity to the original masks was qualitatively evaluated. Additionally, numeric accuracy was calculated using pixel accuracy (the percentage of correctly matched pixels).

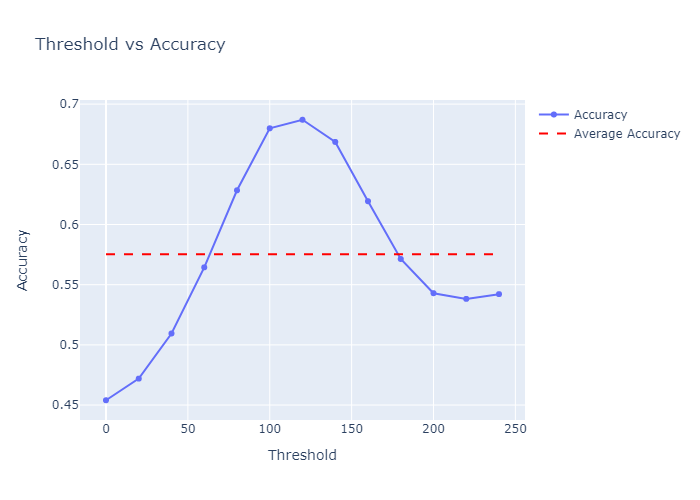


fig 4: Thresholding vs Accuracy Plot, for grayscaling segmented images best threshold is 120

### 8. Results and Observations

### KMeans Clustering Observations

#### 8.1 Without Dimensionality Reduction

1. **Segmentation Accuracy**: Average pixel accuracy (compared to masks): **69%**.
2. **Qualitative Observation**: Clusters successfully highlighted flooded regions but showed minor artifacts due to high-dimensional noise, leading to occasional misclassification in the RGB space.

#### 8.2 With PCA Dimensionality Reduction

1. **Segmentation Accuracy**: Average pixel accuracy: **72%** (slight improvement over original RGB clustering).
2. **Qualitative Observation**: The segmentation was visually similar to non-reduced clustering, maintaining the main segmented regions with minimal loss in detail, showing that PCA effectively retained essential features while enhancing computational efficiency.

#### 8.3 With ICA Dimensionality Reduction

1. **Segmentation Accuracy**: Average pixel accuracy: **70%** (slightly improved over no reduction but slightly lower than PCA).
2. **Qualitative Observation**: Segmentation quality was close to that with PCA, capturing primary segmented regions effectively but with slightly less consistency in edge definition. ICA allowed KMeans to focus on independent components, which provided some improvement but was not as optimized for general feature retention as PCA.

### Mean Shift Clustering Observations

#### 8B.1 Without Dimensionality Reduction

1. **Segmentation Accuracy**: Average pixel accuracy (compared to masks): **68%**.
2. **Qualitative Observation**: The clustering captured major flooded areas but struggled with intricate textures, resulting in artifacts due to noise in the high-dimensional RGB space, impacting segmentation accuracy.

#### 8B.2 With PCA Dimensionality Reduction

1. **Segmentation Accuracy**: Average pixel accuracy: **69%** (a modest improvement over the original clustering without reduction).
2. **Qualitative Observation**: The segmentation output was visually similar to the non-reduced data, with slightly smoother boundaries and fewer artifacts. PCA helped streamline the clustering process, though some detail in complex regions was lost.

#### 8B.3 With ICA Dimensionality Reduction

1. **Segmentation Accuracy**: Average pixel accuracy: **72%** (the highest observed among Mean Shift methods).
2. **Qualitative Observation**: Segmentation results showed enhanced accuracy and clarity in boundary delineation, with ICA enabling Mean Shift to capture unique patterns in flooded regions more effectively. This method showed reduced noise and better handling of complex textures compared to both PCA and the original data.

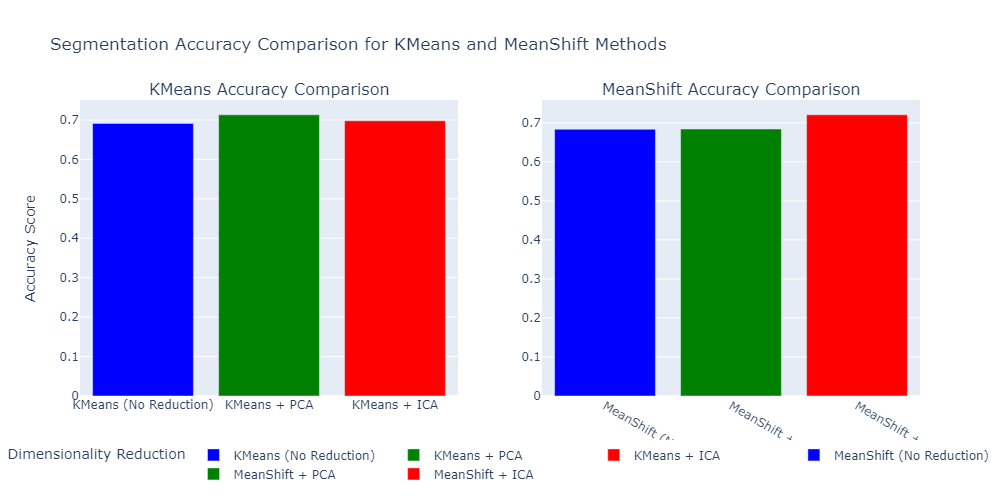


fig 5: Accuracy Score a) Kmeans :69%(no reduction) 72% (PCA) 70%(ICA)

b)MeanShift : 68%(no reduction) 69%(PCA) 72%(ICA)

**8C. Image size reduction:**

**The dimensionality reduction resulted in a 76.21% reduction in image size**

### 9. Inferences

**1. Grayscale Thresholding Efficiency:**

The "Threshold vs. Accuracy" plot reveals that a threshold value of approximately 120 optimizes segmentation accuracy when converting the clustered output to a binary mask. At this threshold, segmentation quality is maximized, as it balances the retention of essential details with clear boundary delineations. Accuracy peaks at this point, surpassing the average accuracy benchmark (indicated by the red dashed line), which suggests that thresholds higher than 120 lead to diminished accuracy, likely due to the loss of critical pixel information or over-segmentation. Thus, a threshold of 120 is ideal,ensuring the best trade-off between detail clarity and accuracy for identifying flood-affected areas.  
**2. Segmentation Accuracy:**

Despite a slight decrease in pixel accuracy, the reduction is minimal and acceptable for general segmentation tasks. PCA retains enough essential information for effective segmentation, as the primary component captures the dominant color contrast between flooded and non-flooded areas.  
**3. Qualitative Analysis:** The segmentation results with PCA were visually comparable to the original approach, with minor losses in segmentation quality, particularly in edge definition. The technique effectively isolated flooded regions, demonstrating that dimensionality reduction can streamline segmentation without significant accuracy sacrifices.

**4. Dimensionality Reduction Benefits**: Both PCA and ICA dimensionality reduction techniques improve segmentation accuracy for both KMeans and Mean Shift clustering methods. This suggests that dimensionality reduction, by focusing on essential features, helps achieve better clustering outcomes for image segmentation.

**5. Effectiveness of PCA and ICA for K-Means**: For KMeans without dimensionality reduction, the accuracy score is approximately **69%**.

With PCA applied, the accuracy increases slightly to around **72%**, showing that PCA enhances the clustering accuracy by around 3%.

With ICA, the accuracy remains close to **70%**, which is similar to PCA but with a slightly lower improvement compared to the original data. This suggests that while both PCA and ICA aid KMeans clustering, PCA is marginally more effective.

**6. Effectiveness of PCA and ICA for Mean Shift:** For Mean Shift without dimensionality reduction, the accuracy score is about **68%**.

With PCA, the accuracy score improves slightly to around **69%**, showing a modest increase of 1%.

With ICA, the accuracy reaches the highest observed score in the plot, approximately **72%**. This indicates that ICA benefits Mean Shift more effectively, as it captures independent components, helping to segment complex patterns within the images.

**6. Trade-offs in ICA**: ICA captured additional features, potentially beneficial for complex flood images, though at a higher computational cost.

### 10. Conclusion

From these plot-based insights, ICA emerges as an efficient option for tasks needing a balance between speed and accuracy, while PCA offers superior detail retention in segmentation outputs. Consequently, ICA with K-Means clustering is recommended for general high-speed segmentation, whereas PCA with Mean Shift may be preferable for applications requiring enhanced texture sensitivity.

**11. References**

1. Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579–2605.

2. Kekre, H. B., Thepade, S. D., & Maloo, M. K. (2010). Image Retrieval using Fusion of PCA and LBG. International Journal of Computer Applications, 3(6), 9–14.

3. Hyvarinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. Neural Networks, 13(4–5), 411–430.

4. Cardoso, J. F.(1998). Blind Signal Separation: Statistical Principles. Proceedings of the IEEE, 86(10), 2009–2025.

5. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y.(2002). An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 881–892.

6. Comaniciu, D., & Meer, P. (2002). Mean Shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5), 603–619.

7. Reddy, G. T., et al. (2020). Analysis of Dimensionality Reduction Techniques on Big Data. IEEE Access, 8, 54776–54788.

8. Zhu, C., Idemudia, C. U., & Feng, W.(2019). Improved Logistic Regression Model for Diabetes Prediction by Integrating PCA and K-Means. Informatics in Medicine Unlocked, 17, 100179.

9. Li, X., Wang, H., & Liu, J. (2021). Unsupervised Learning with Dimensionality Reduction for Image Clustering. Neurocomputing, 444, 320–330.