Fisher Linear Discriminant and PCA analysis for Breast Cancer Detection.

Task 1: Fisher Linear Discriminant and PCA Analysis with 10 Components

- Apply Fisher linear discriminant to the first 2/3rd of the data (the training set) after reducing its dimensionality to 10 using PCA.
- Apply the same PCA transform to the remaining 1/3rd of the data (testing set) and then use the previously derived Fisher LDA to produce testing scores.
- Repeat this process without PCA shortening and produce all the training and validation ROC curves (4 altogether) and
- Also present the training and testing AUCs and d primes, and
- Plot the Scree graph.

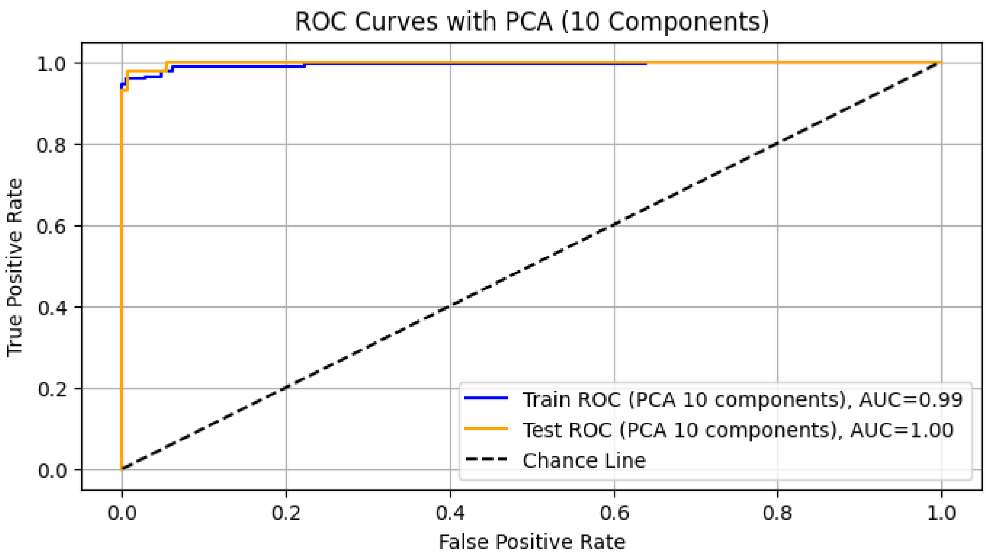
Results :

- **With PCA (10 Components) :**
  Explained variance by each principal component (10 components):
  [0.44629805 0.1926257 0.09980873 0.05816819 0.0512186
  0.0397819 0.0214634 0.01622204 0.0141189 0.01160413]

- **AUCs for PCA - reduced Data (10 components)**
  AUC (Train with PCA) = 0.993
  AUC (Test with PCA) = 0.998

- **d-prime for PCA-reduced data (10 components) :**
  d-prime (Train with PCA) = 3.126
  d-prime (Test with PCA) = 3.124

- **ROC Curves : PCA 10 components**



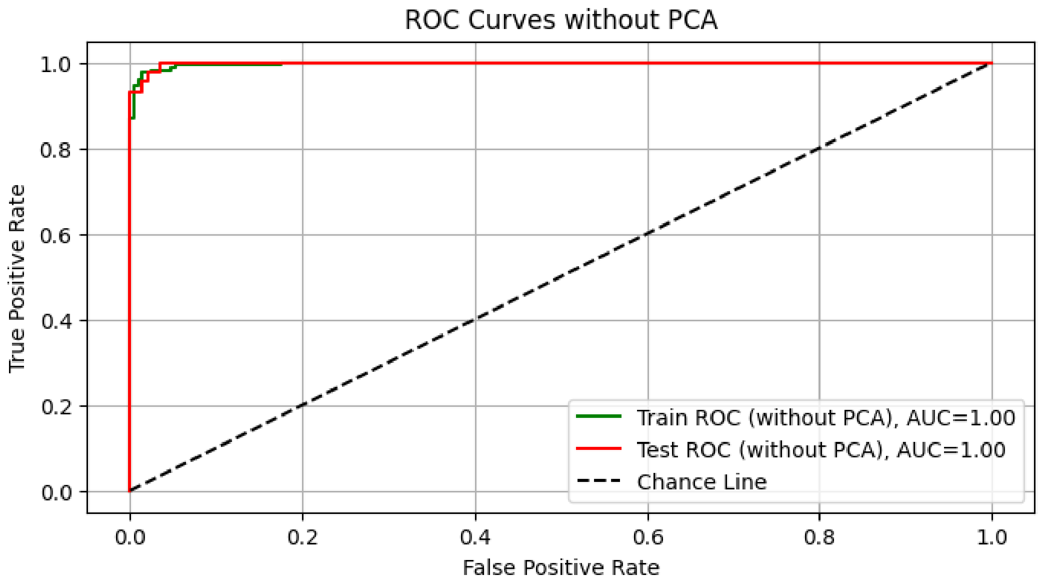
- The 10-component PCA reduction preserved a significant portion of variance, leading to strong classification performance. The ROC curves showed good separation, and AUC values for both training and testing were high, indicating reliable model accuracy.
- The d-prime values were also high, confirming excellent class separability.

- Overall, the 10-component PCA reduction was effective in maintaining model performance while simplifying the feature space.

- **Without PCA :**

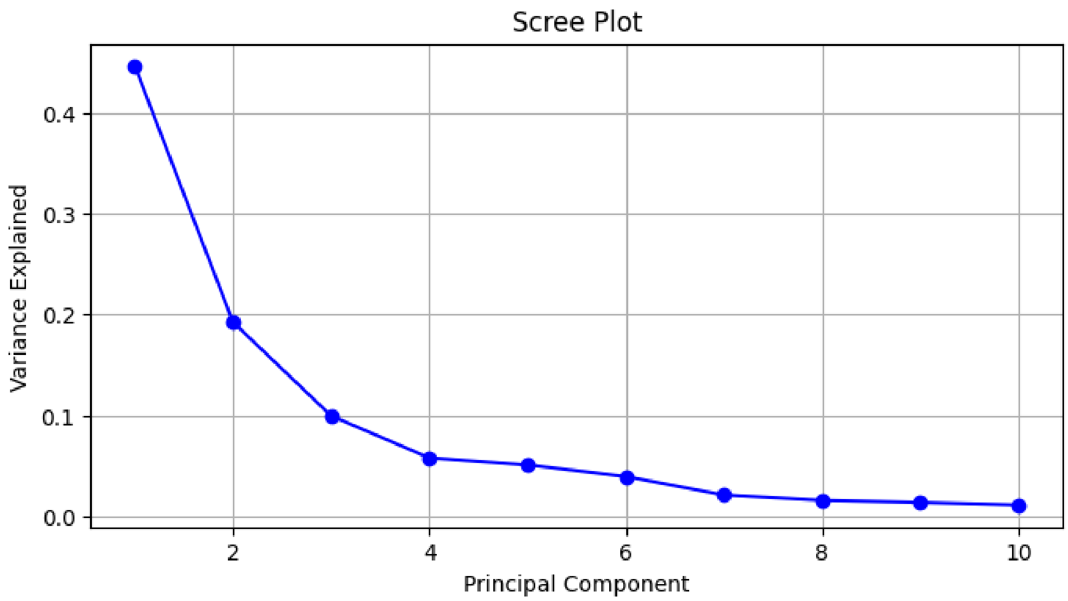
  - **AUCs for Original Data (Without PCA)**
    AUC (Train without PCA) = 0.997
    AUC (Test without PCA) = 0.998
  - **d-prime for full data without PCA :**
    d-prime (Train without PCA) = 3.699
    d-prime (Test without PCA) = 3.674

  - **ROC Curves : Without PCA**



  - Using LDA on the full dataset (without PCA) served as a baseline and showed comparable AUC and d-prime values to the 10-component PCA-reduced model.
  - This indicates that PCA did not significantly impact the classification power, while it did help reduce data complexity.

  - **Scree Plot :** A Scree plot is generated to show variance explained by each principal component, useful for selecting an optimal number of components.



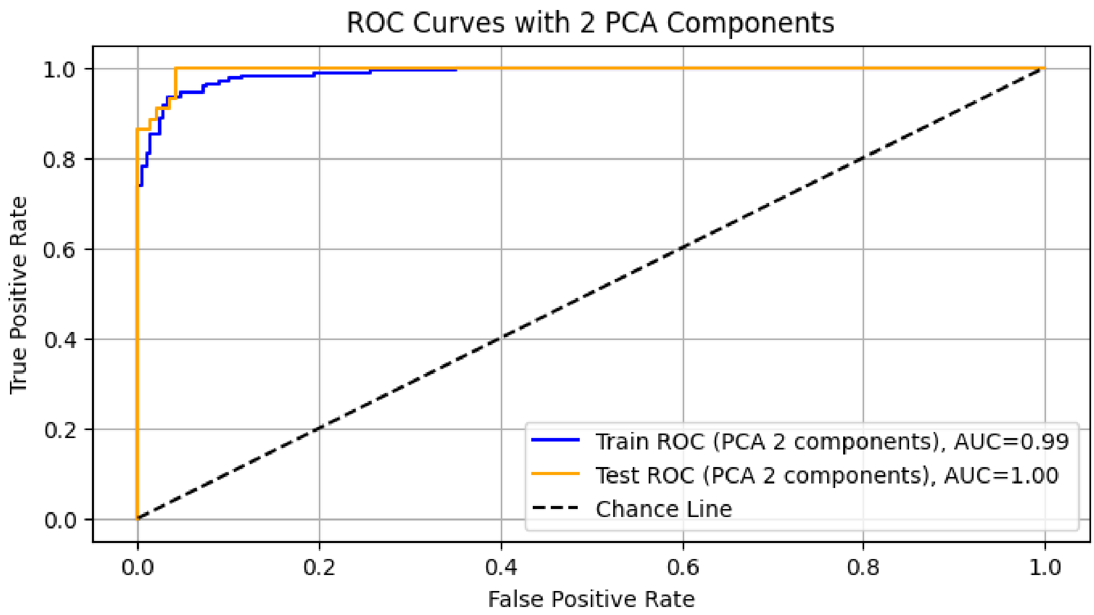
Task 2: Repeat with PCA to 2 and 5 components

- **AUC values for both PCA - Reduced Data (2 and 5 Components)**
  AUC (Train with 2 PCA components) = 0.988
  AUC (Test with 2 PCA components) = 0.996
  AUC (Train with 5 PCA components) = 0.993
  AUC (Test with 5 PCA components) = 0.998

- **d-prime values for both PCA-Reduced Data (2 and 5 Components)**
  d-prime (Train with 2 PCA components) = 2.650
  d-prime (Test with 2 PCA components) = 2.583
  d-prime (Train with 5 PCA components) = 3.042
  d-prime (Test with 5 PCA components) = 2.972

- **PCA with 2 Components:**
  Train AUC = 0.99
  Test AUC = 1.00
  Train d-prime = 2.65
  Test d-prime = 2.58

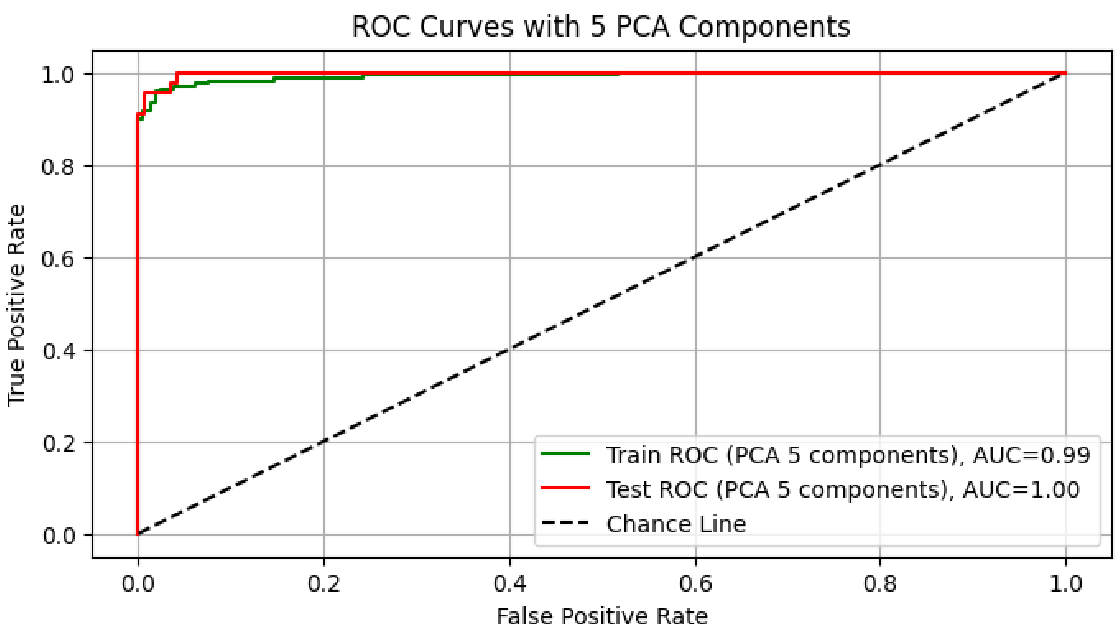
- **ROC Curves : for PCA with 2 Components**



- Reducing the dataset to only 2 components resulted in a noticeable loss of variance, which negatively impacted model accuracy. The AUC and d-prime values were lower, reflecting limited class separability and reduced performance.
- The 2-component PCA reduction was insufficient to retain critical information for effective classification.

- **PCA with 5 Components:**
  Train AUC = 0.99
  Test AUC = 1.00
  Train d-prime = 3.04
  Test d-prime = 2.97

- **ROC Curves : for PCA with 5 Components**



- Using 5 components provided a better balance, retaining more variance than the 2-component reduction while still simplifying the dataset.

- Although the performance improved over the 2-component reduction, the AUC and d-prime values were still somewhat lower than with 10 components, indicating that some discriminative information was lost.

Conclusion : The results suggest that reducing the dimensionality to 5 components offers a moderate improvement over 2 components but does not perform as well as the 10-component reduction.