ME 4990/6990 02 – Spring 2025
Homework #3 – Principal Component Analysis

- Please submit any Python code used as a separate file in addition to a text document (Word) which provides answers to the below problems. Writing should be in complete sentences with enough detail to demonstrate a thoughtful consideration of what is asked

The sklearn library includes a set of data containing image information on fine needle aspirates used to identify breast cancer. The dataset contains information gathered from the images as well as whether the mass was malignant (target=0) or benign (target=1). The included python template shows how to make arrays of the features and targets.

1. Perform a PCA transformation of the features and plot the individual and cumulative explained variance as a function of number of components.

2. Plot the first two components (component 1 along the x-axis and component 2 along the y-axis) for all of the data, differentiating between malignant and benign samples. Are the two cases well differentiated by just the first 2 components?

3. Given that a linear kernel was sufficient for differentiating the 2 cases when using an SVC, does the result for question 2 make sense? Explain.

4. Randomly select a subset containing 100 of the benign cases and perform PCA on this subset. Perform anomaly detection on the remaining benign and malignant cases uses this new PCA transform. You will want to consider the information loss from keeping only a finite number of components.

5. Determine a loss threshold and number of components to create a model for detecting potentially malignant cases using only PCA. Show the confusion matrix for this model.