ME 4990/6990 02 – Spring 2025
Homework #2 – Support Vector Machines

- 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=1) or benign (target=0). The included python template shows how to make arrays of the features and targets.

1. What is the split of malignant vs benign for the dataset? What issues might you foresee using a dataset with this split?

2. Generate a support vector classifier (SVC) to determine if the input data indicates a malignant or benign cancer. Use the linear kernel and varying values of C. Plot the accuracy score as a function of C. Use cross validation. What value of C gives the best results?

3. Using the radial basis function kernel (rbf), determine the hyperparameters (C and gamma) which result in the best model.

4. Generate the most accurate SVC model you can (feel free to use other kernels and varying hyperparameters). What is the precision and recall of this model? Use a train/test split and give the precision and recall for the test data. Show the confusion matrix for both the training and test data.

5. If your trained model from question 4 is given a novel set of inputs and determines that the cancer is malignant, what is the chance that it is wrong? How does this relate to your answer for question 1?