ECS171 Fall 2017
HW3 Report
Gabriel Yin
999885129

Problem 1

Optimal constrained parameter value: 0.00694
10-Fold cross-validation error: 0.03634
Number of non-zero features: 58

In Linear Regression, we aim to minimize the squared error. In Lasso method, we add a new constrain to the equation to prevent the value from going beyond a certain value. More specifically, we are trying to solve the following equation (Image taken from Wikipedia):

$$\min_{\beta_0,\beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t. \text{[1]}$$

In this case, y_i is the actual value of i-th sample, and $x_i^T B$ is the predicted value of the i-th input value. We add a new value of Beta as the penalty constrain and we further define that the sum of Beta cannot exceed a certain value t. In Ridge method, however, there is no constrain for the Beta value. I choose to use Lasso method to solve this problem. To get the optimal constrained parameter value, I used bic approach (please see code for details).

Problem 2

90^{th} Confident interval Lower: 0.8345 Upper: 0.9678
95^{th} Confident interval Lower: 0.8123 Upper: 0.9678

First of all, as we are dealing with a large dataset, we can assume that the data is distributed normally (i.e. normal distribution). To solve this problem, I used Ridge method instead of the first one. Then I use bootstrap method to calculate 90^{th} and 95^{th} confidence interval of the data. To do so, I first set the iteration number to be 1000. Then for each iteration, I randomly take a index (using np.random.choice) and use this to split the training dataset x and training dataset y. Similarly, based on the previously acquired training index, I get the test index and use it to split the test set value. In each iteration I calculate the score using prediction function and append it to a list.

To calculate confidence interval, I first calculate p, a percentile indicator.
P = 100 * ((1 – alpha_val)/2.0) and lower can be calculated as
low = max(0.0, np.percentile(data, int(p))). Similarly, to calculate the upper bound, p = 100 * (alpha_val + ((1.0 – alpha_val)/2.0)) and upper = min(1.0, np.percentile(data, int(p))).

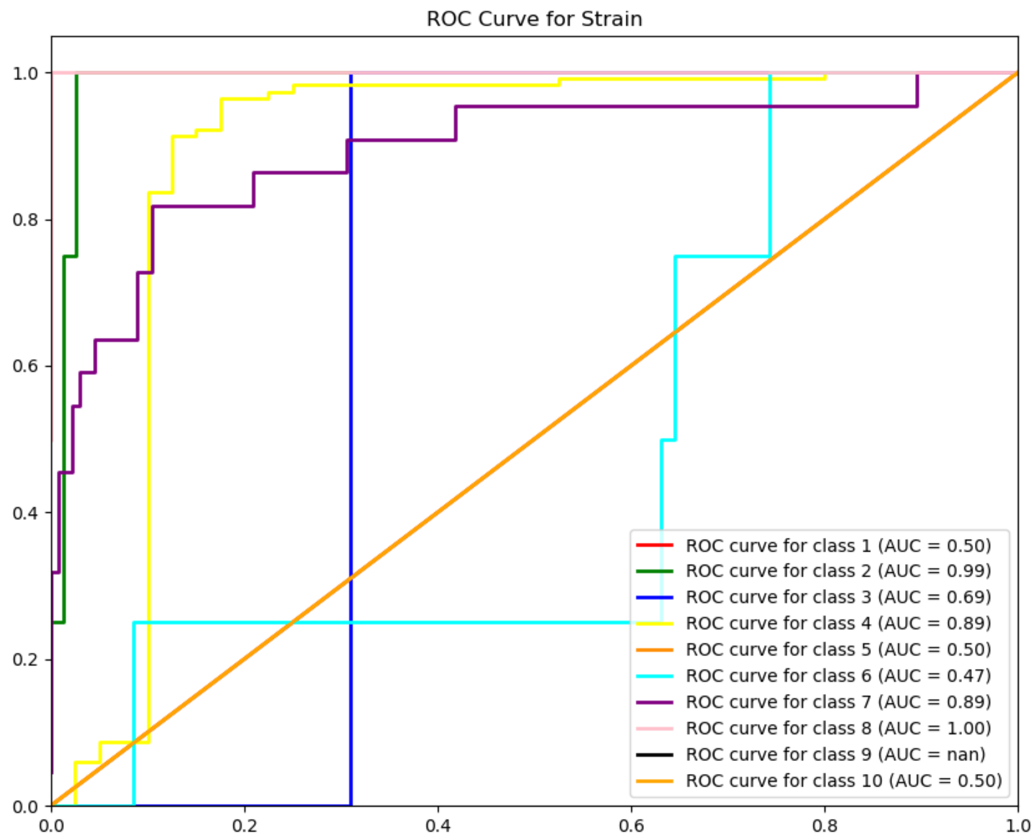
Problem 3

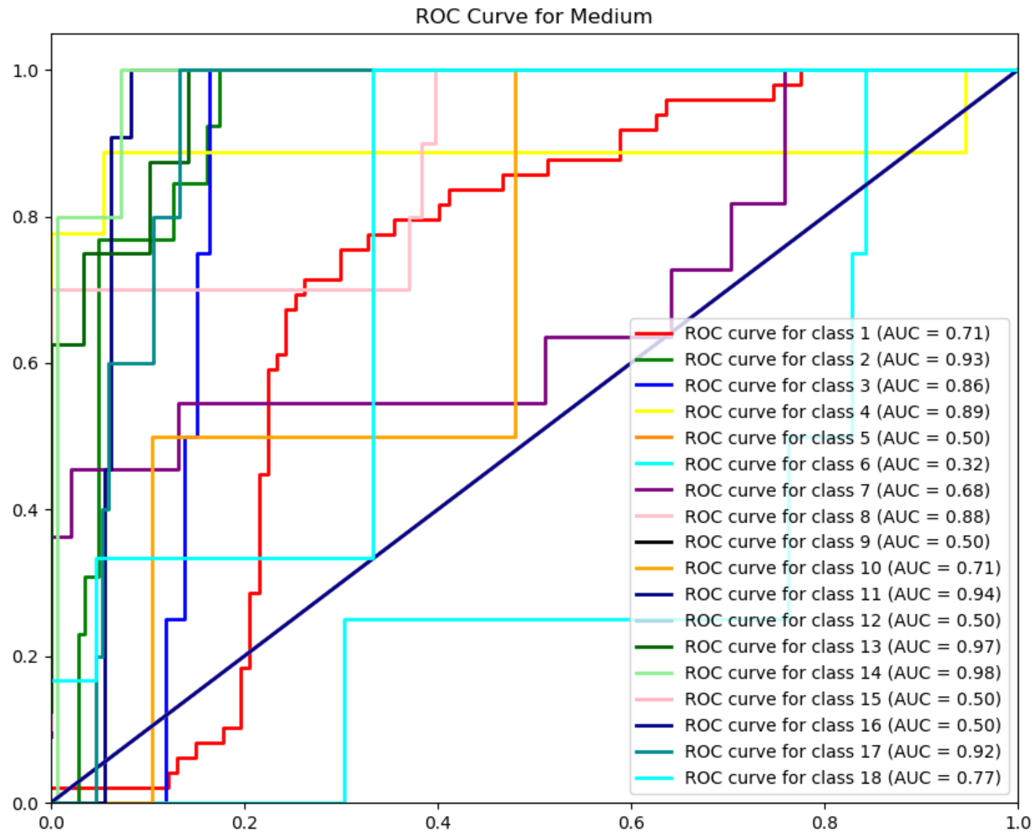
Based on my code, the result is 0.3936.

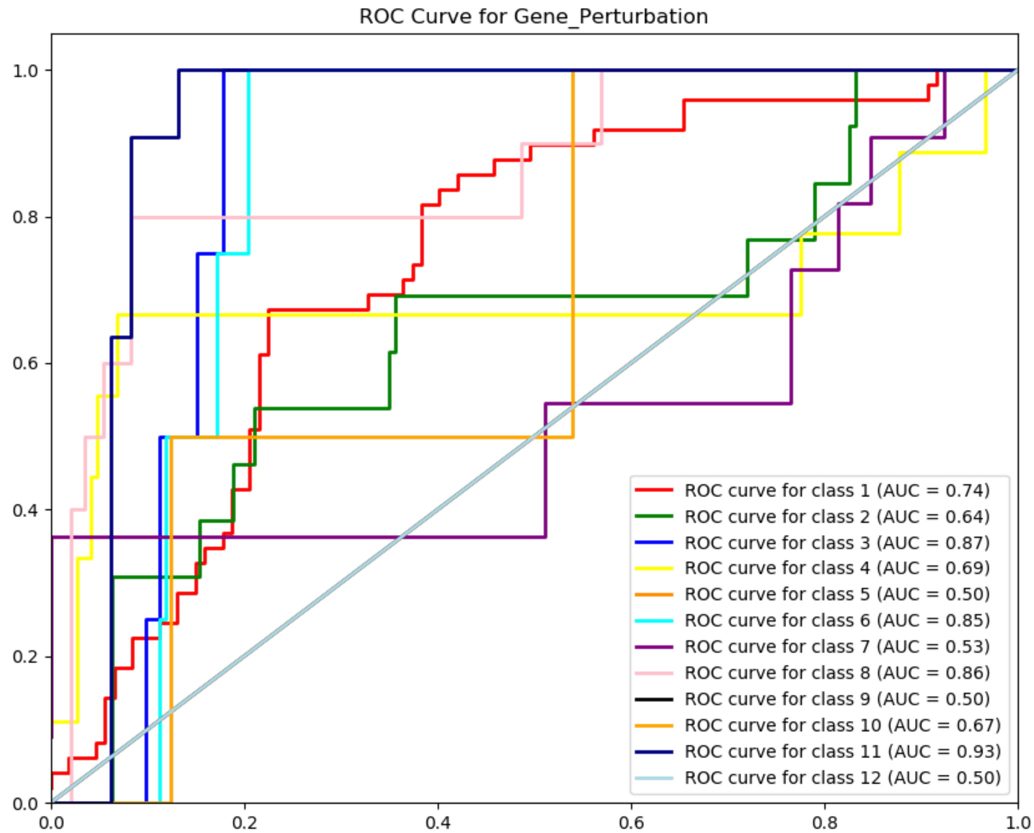
Problem 4

Strain: Best AUC: 1 AUPRC for the best AUC: 0.9342
Medium: Best AUC: 1 AUPRC for the best AUC: 0.8637
Environmental Perturbation: Best AUC: 1 AUPRC for the best AUC: 0.9134
Genetic Perturbation: Best AUC: 1 AUPRC for the best AUC: 0.9231

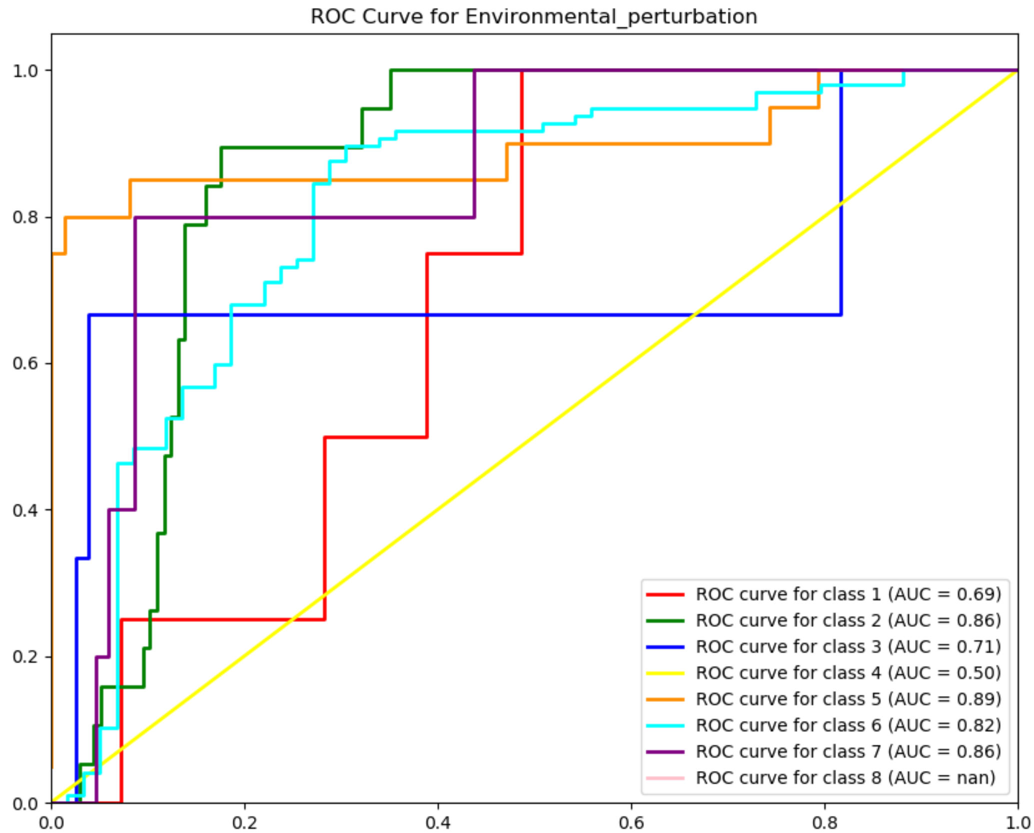
ROC Curve Graphs:



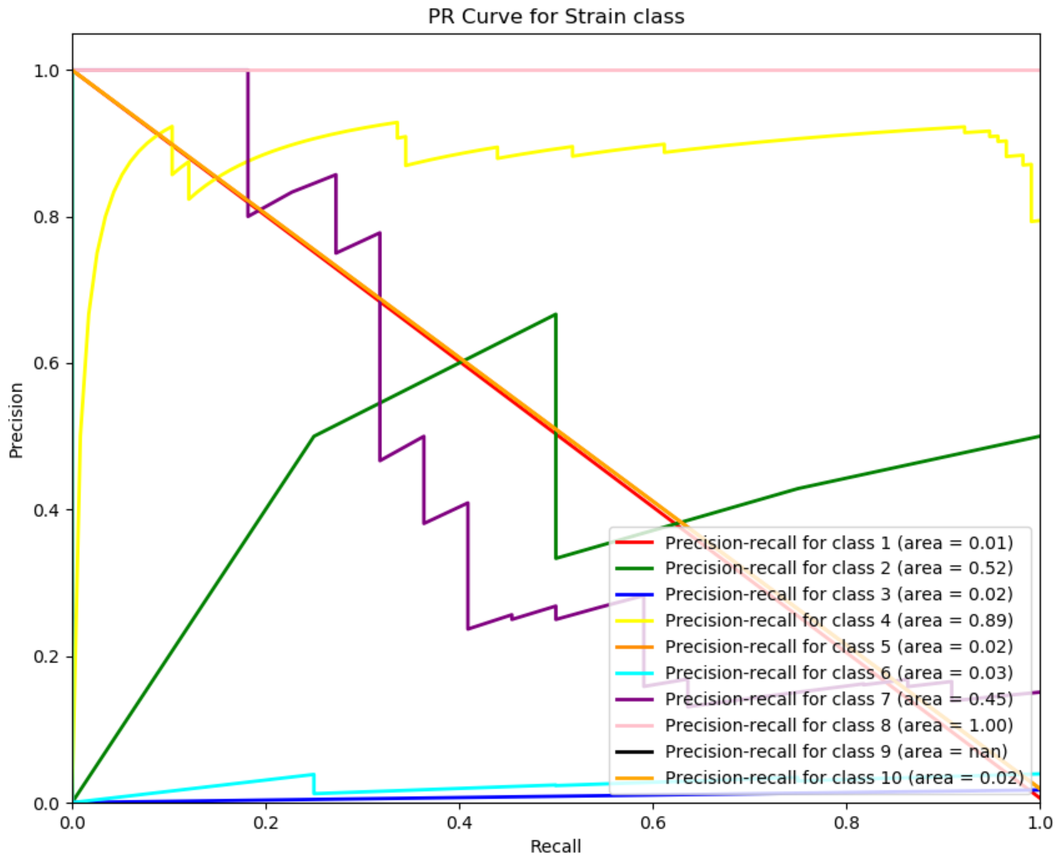

ROC Curve for Medium
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve for class 1 (AUC = 0.71)
ROC curve for class 2 (AUC = 0.93)
ROC curve for class 3 (AUC = 0.86)
ROC curve for class 4 (AUC = 0.89)
ROC curve for class 5 (AUC = 0.50)
ROC curve for class 6 (AUC = 0.32)
ROC curve for class 7 (AUC = 0.68)
ROC curve for class 8 (AUC = 0.88)
ROC curve for class 9 (AUC = 0.50)
ROC curve for class 10 (AUC = 0.71)
ROC curve for class 11 (AUC = 0.94)
ROC curve for class 12 (AUC = 0.50)
ROC curve for class 13 (AUC = 0.97)
ROC curve for class 14 (AUC = 0.98)
ROC curve for class 15 (AUC = 0.50)
ROC curve for class 16 (AUC = 0.50)
ROC curve for class 17 (AUC = 0.92)
ROC curve for class 18 (AUC = 0.77)

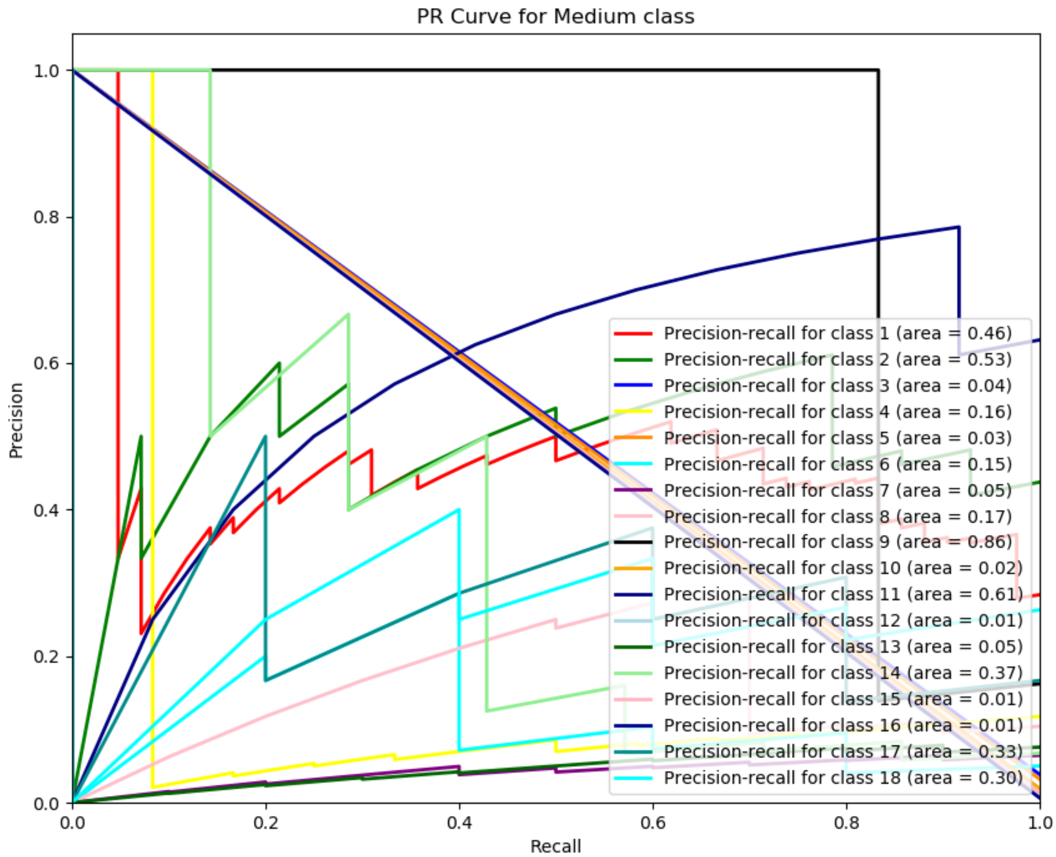


ROC Curve for Gene_Perturbation
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve for class 1 (AUC = 0.74)
ROC curve for class 2 (AUC = 0.64)
ROC curve for class 3 (AUC = 0.87)
ROC curve for class 4 (AUC = 0.69)
ROC curve for class 5 (AUC = 0.50)
ROC curve for class 6 (AUC = 0.85)
ROC curve for class 7 (AUC = 0.53)
ROC curve for class 8 (AUC = 0.86)
ROC curve for class 9 (AUC = 0.50)
ROC curve for class 10 (AUC = 0.67)
ROC curve for class 11 (AUC = 0.93)
ROC curve for class 12 (AUC = 0.50)

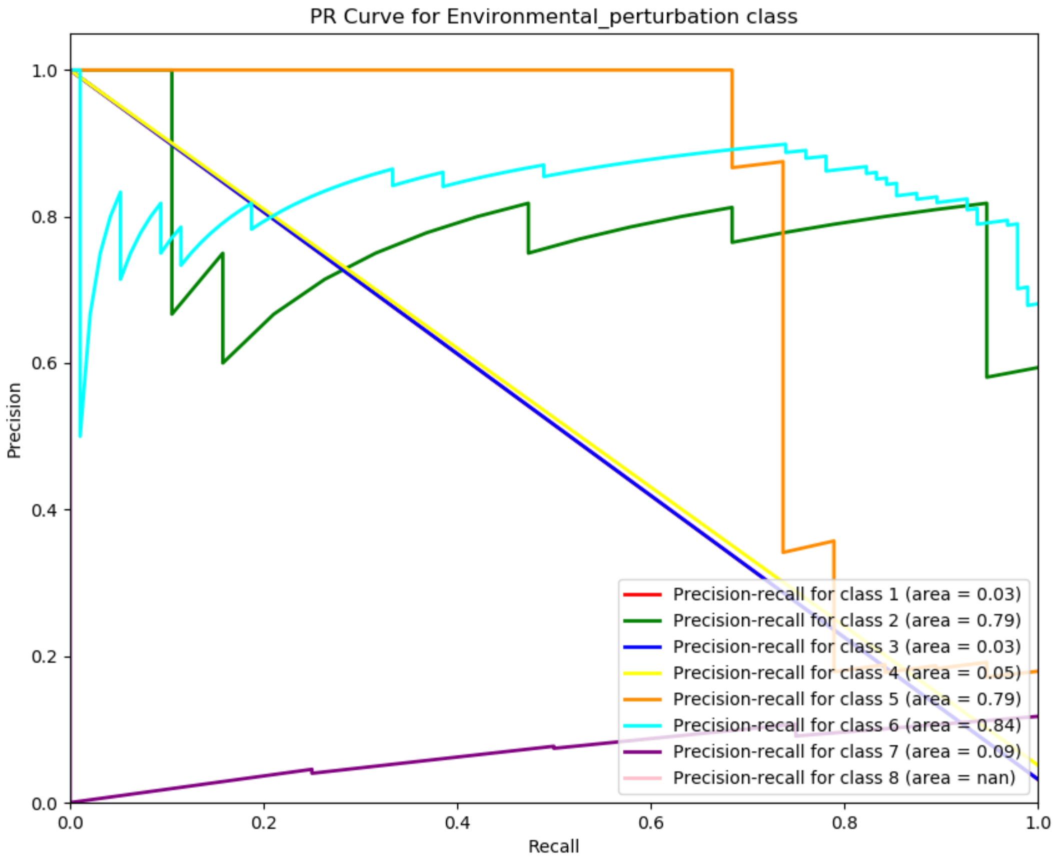


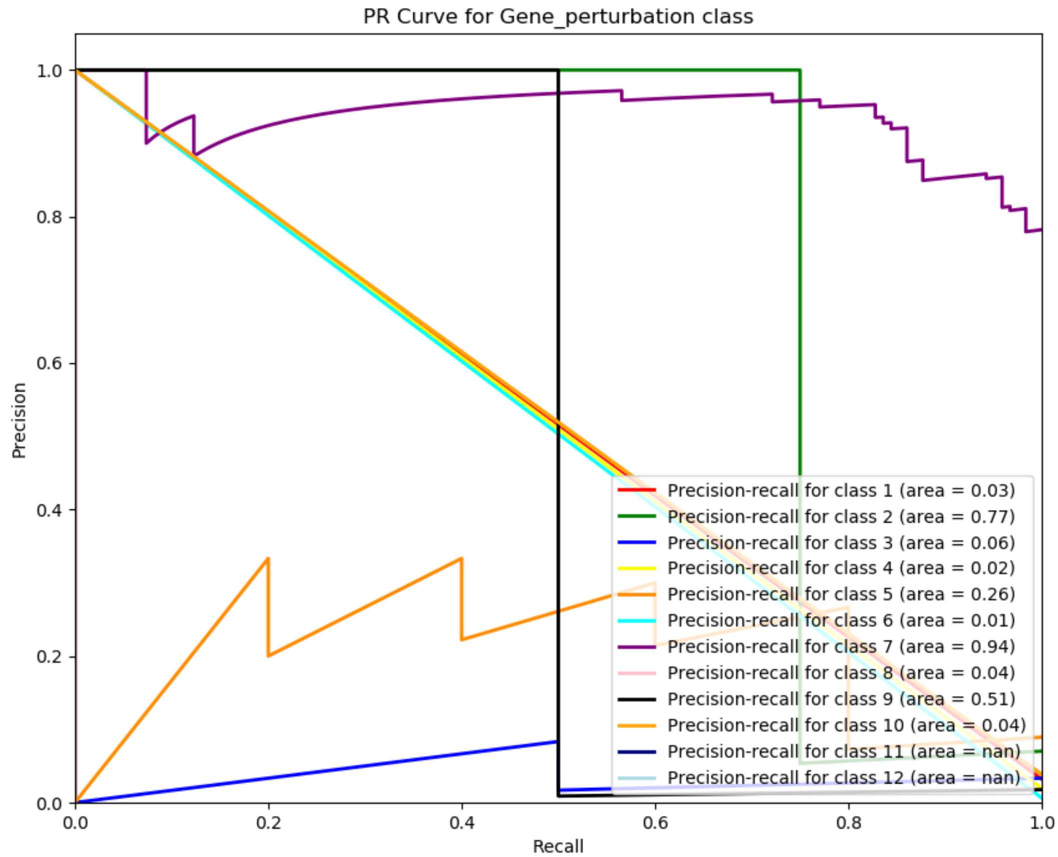
PR Curve Graphs:


PR Curve for Strain class
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Precision-recall for class 1 (area = 0.01)
Precision-recall for class 2 (area = 0.52)
Precision-recall for class 3 (area = 0.02)
Precision-recall for class 4 (area = 0.89)
Precision-recall for class 5 (area = 0.02)
Precision-recall for class 6 (area = 0.03)
Precision-recall for class 7 (area = 0.45)
Precision-recall for class 8 (area = 1.00)
Precision-recall for class 9 (area = nan)
Precision-recall for class 10 (area = 0.02)



PR Curve for Medium class
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Precision-recall for class 1 (area = 0.46)
Precision-recall for class 2 (area = 0.53)
Precision-recall for class 3 (area = 0.04)
Precision-recall for class 4 (area = 0.16)
Precision-recall for class 5 (area = 0.03)
Precision-recall for class 6 (area = 0.15)
Precision-recall for class 7 (area = 0.05)
Precision-recall for class 8 (area = 0.17)
Precision-recall for class 9 (area = 0.86)
Precision-recall for class 10 (area = 0.02)
Precision-recall for class 11 (area = 0.61)
Precision-recall for class 12 (area = 0.01)
Precision-recall for class 13 (area = 0.05)
Precision-recall for class 14 (area = 0.37)
Precision-recall for class 15 (area = 0.01)
Precision-recall for class 16 (area = 0.01)
Precision-recall for class 17 (area = 0.33)
Precision-recall for class 18 (area = 0.30)

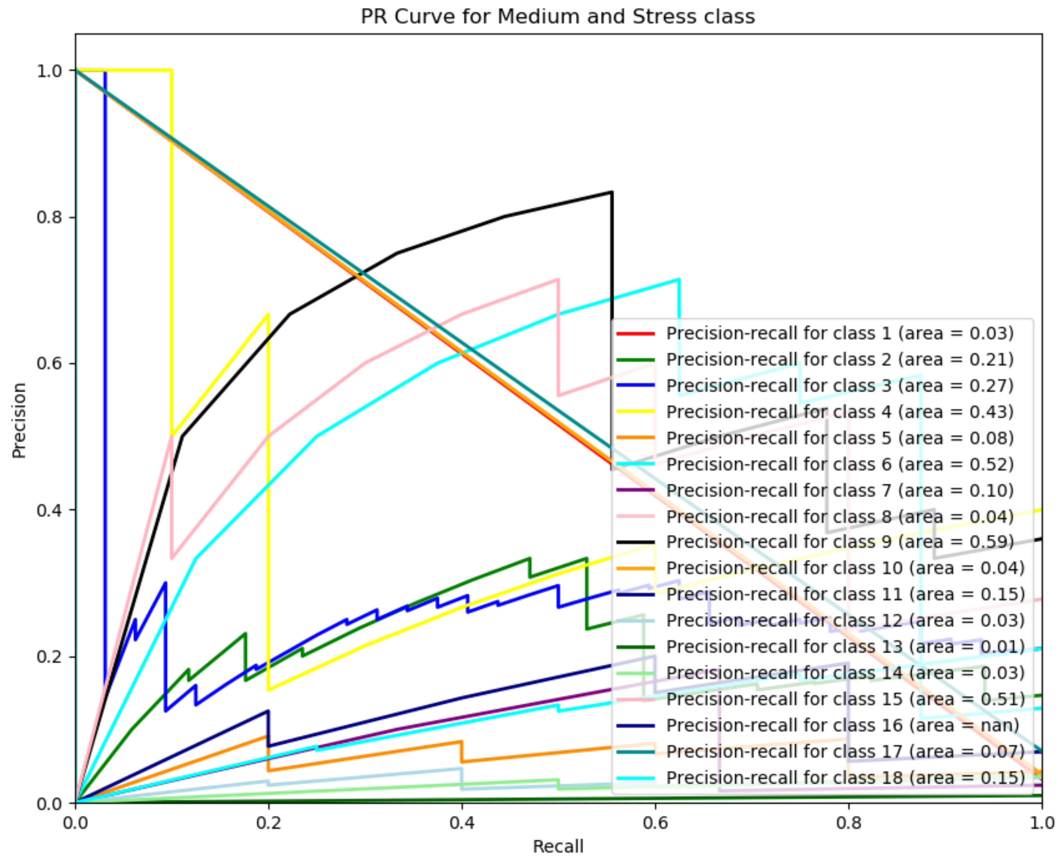

PR Curve for Environmental_perturbation class
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Precision-recall for class 1 (area = 0.03)
Precision-recall for class 2 (area = 0.79)
Precision-recall for class 3 (area = 0.03)
Precision-recall for class 4 (area = 0.05)
Precision-recall for class 5 (area = 0.79)
Precision-recall for class 6 (area = 0.84)
Precision-recall for class 7 (area = 0.09)
Precision-recall for class 8 (area = nan)



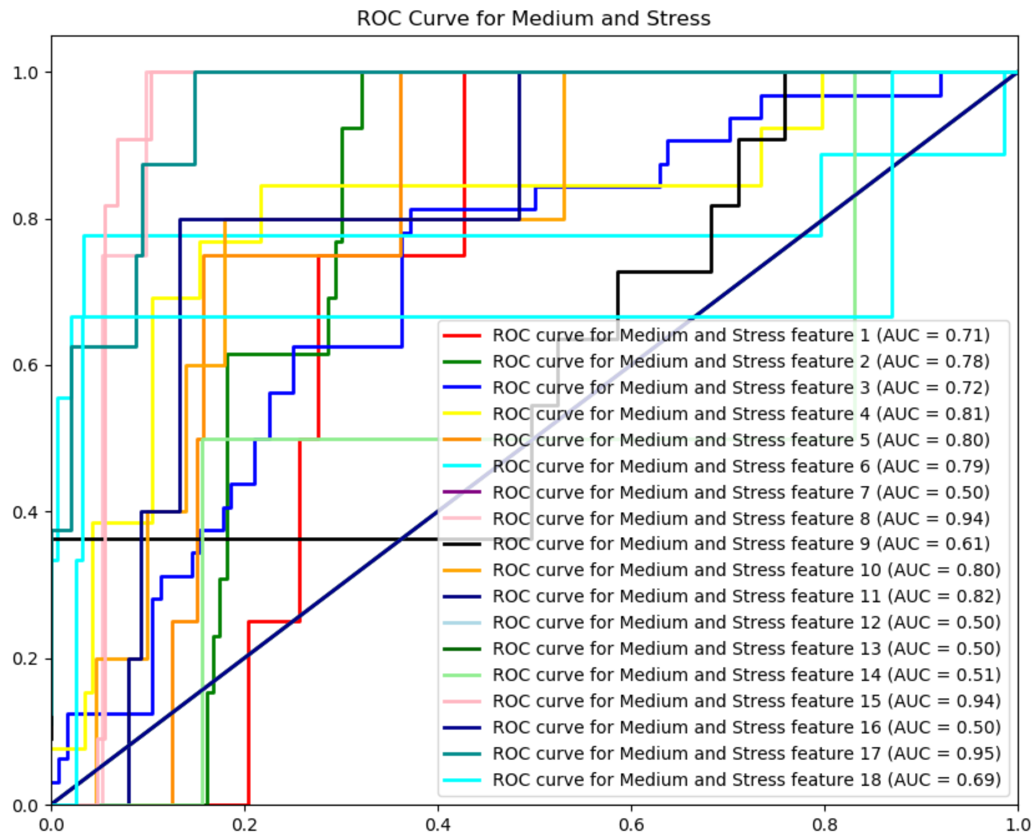
Problem 5

Before process the data, I merge those two classes together and obtain the data. Based on the results, it is more efficient and useful to use single classifier in my case since the combined one fail to recognizes some classes as their features are ignored in the process of feature selection.

Best Combined Classifier AUC: 0.67 AUPRC for best AUC: 0.0



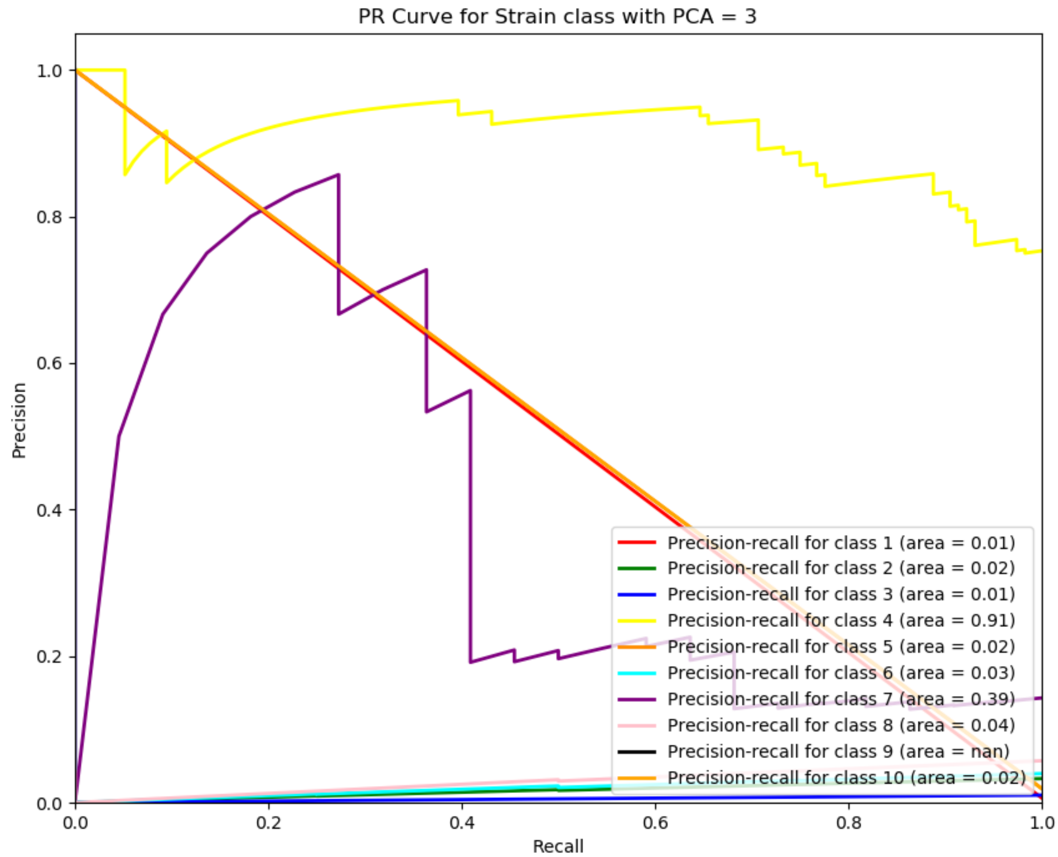
PR Curve for Medium and Stress class
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Precision-recall for class 1 (area = 0.03)
Precision-recall for class 2 (area = 0.21)
Precision-recall for class 3 (area = 0.27)
Precision-recall for class 4 (area = 0.43)
Precision-recall for class 5 (area = 0.08)
Precision-recall for class 6 (area = 0.52)
Precision-recall for class 7 (area = 0.10)
Precision-recall for class 8 (area = 0.04)
Precision-recall for class 9 (area = 0.59)
Precision-recall for class 10 (area = 0.04)
Precision-recall for class 11 (area = 0.15)
Precision-recall for class 12 (area = 0.03)
Precision-recall for class 13 (area = 0.01)
Precision-recall for class 14 (area = 0.03)
Precision-recall for class 15 (area = 0.51)
Precision-recall for class 16 (area = nan)
Precision-recall for class 17 (area = 0.07)
Precision-recall for class 18 (area = 0.15)



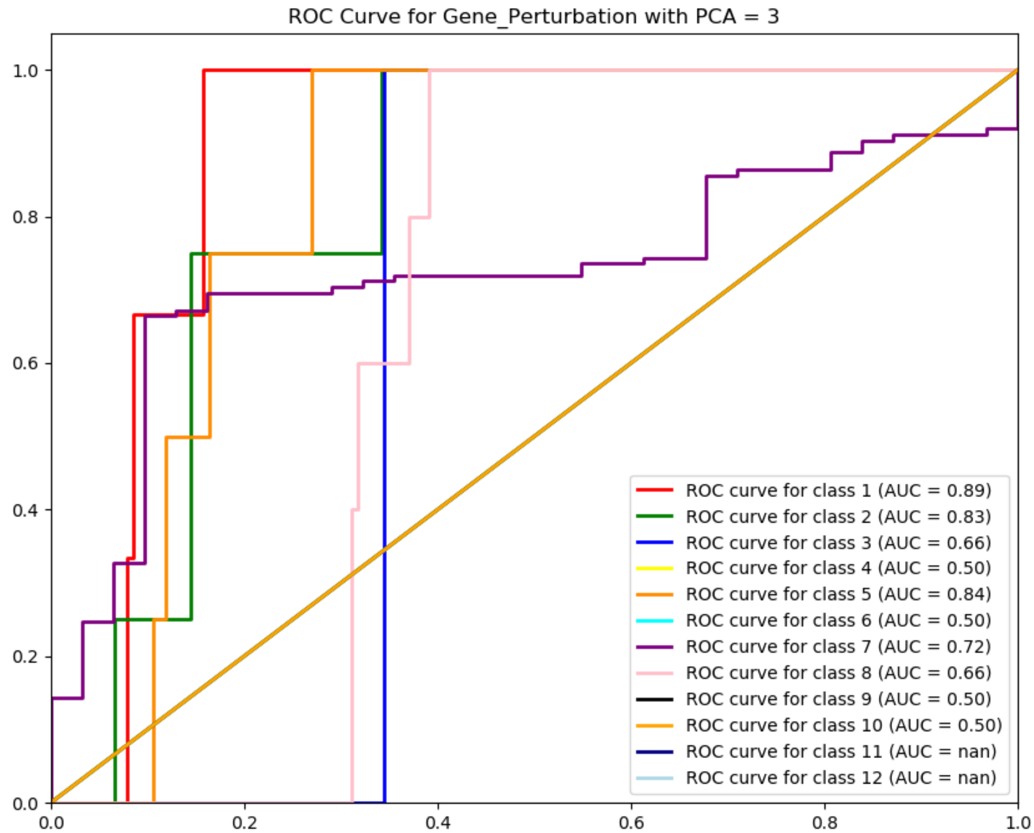
Problem 6

This problem can be done in a similar fashion as problem 4. The only difference is that we are using PCA with value of three instead of the LinearSVC method to do the dimensionality reduction. PCs retain most of the classification performance as before.

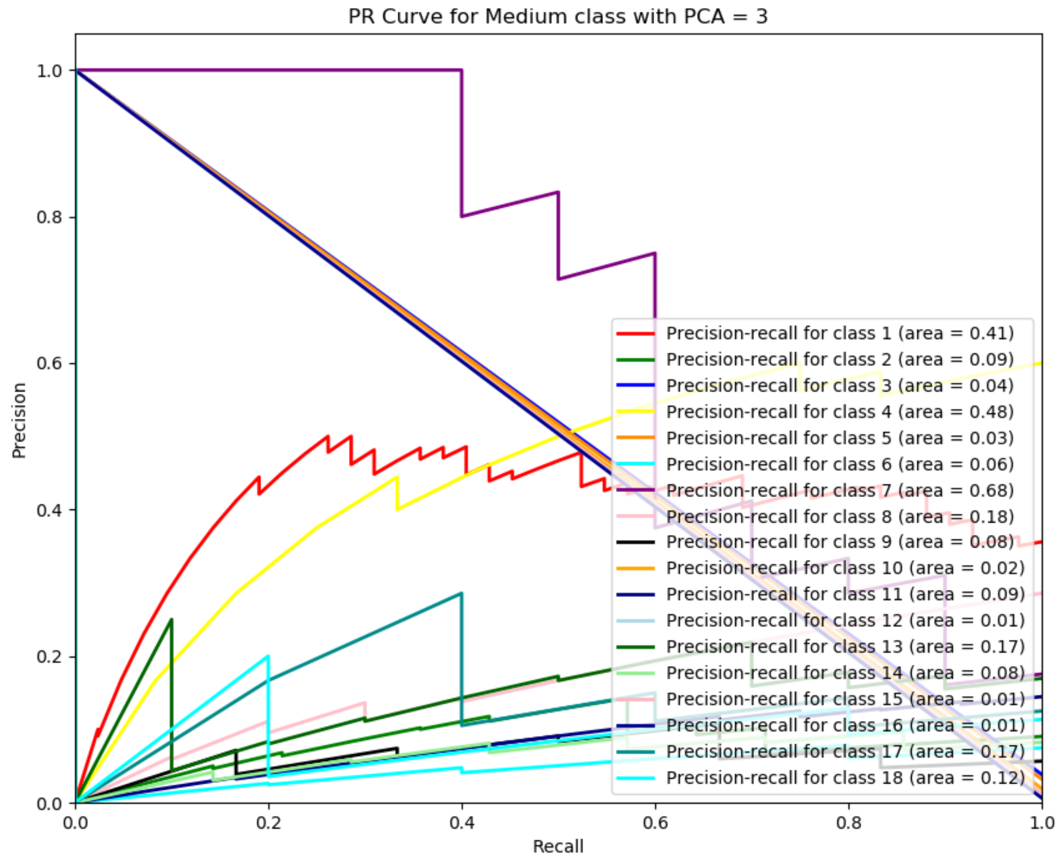
Best AUC value: 1 AUPRC for best AUC value: 0.9951

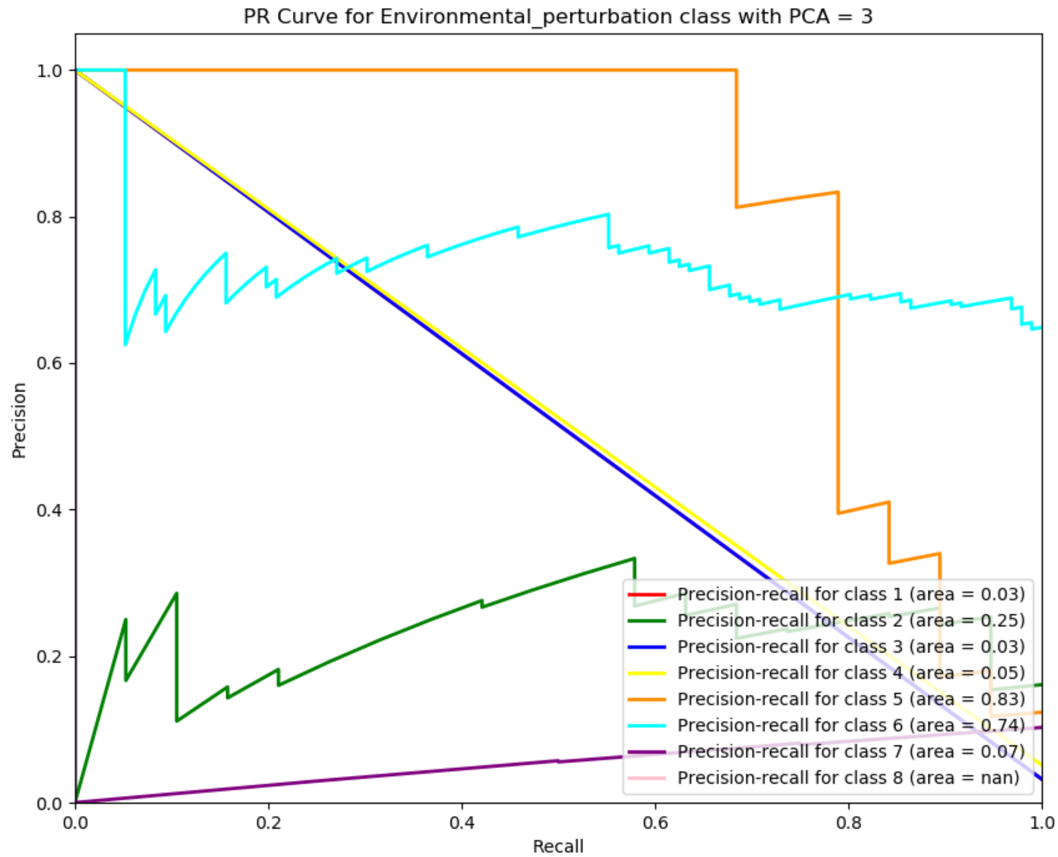


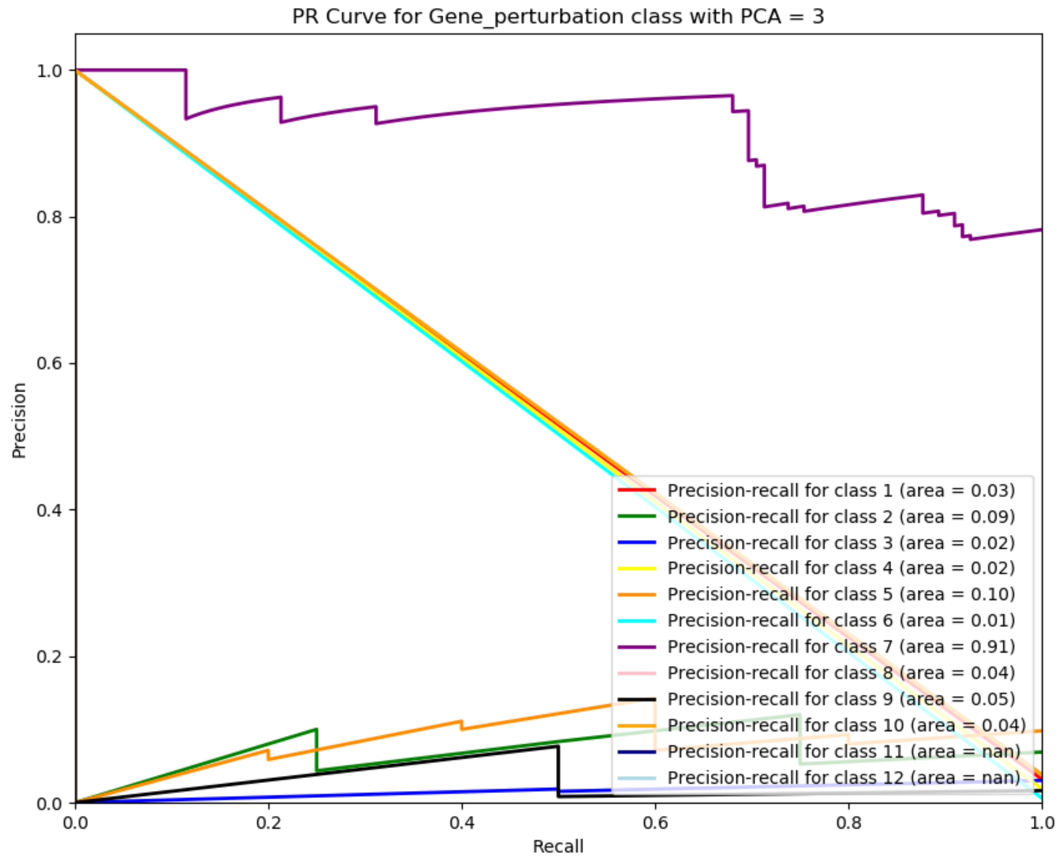
PR Curve for Strain class with PCA = 3
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Precision-recall for class 1 (area = 0.01)
Precision-recall for class 2 (area = 0.02)
Precision-recall for class 3 (area = 0.01)
Precision-recall for class 4 (area = 0.91)
Precision-recall for class 5 (area = 0.02)
Precision-recall for class 6 (area = 0.03)
Precision-recall for class 7 (area = 0.39)
Precision-recall for class 8 (area = 0.04)
Precision-recall for class 9 (area = nan)
Precision-recall for class 10 (area = 0.02)

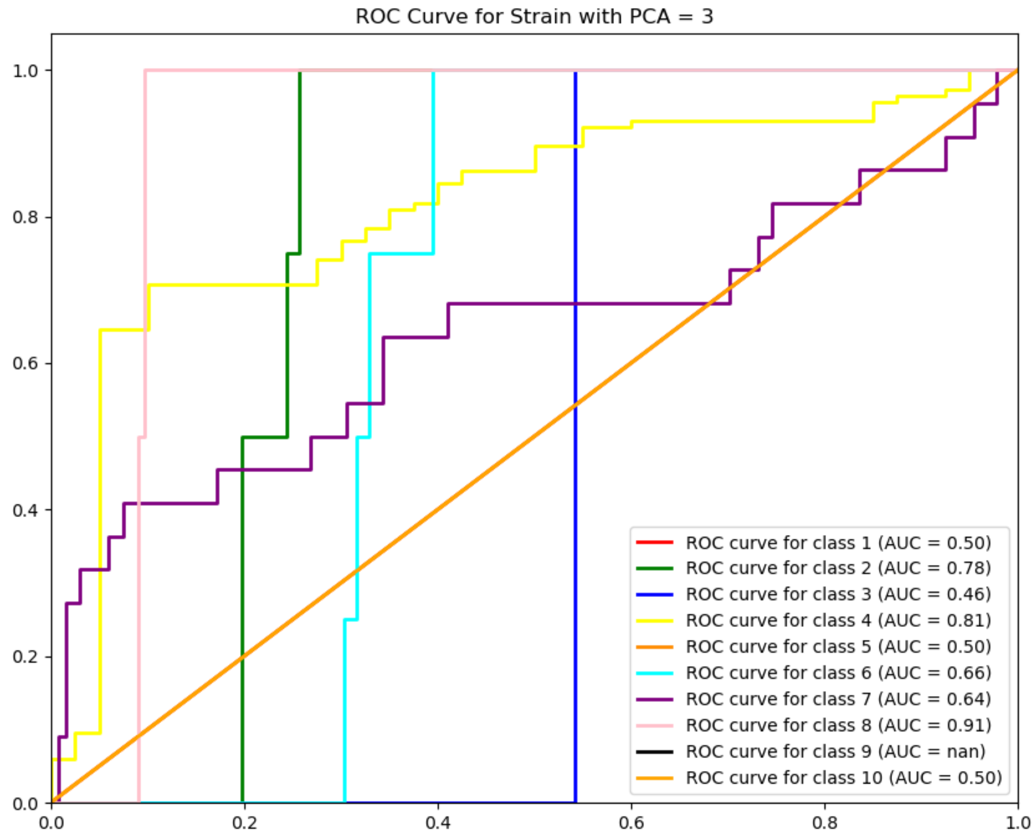


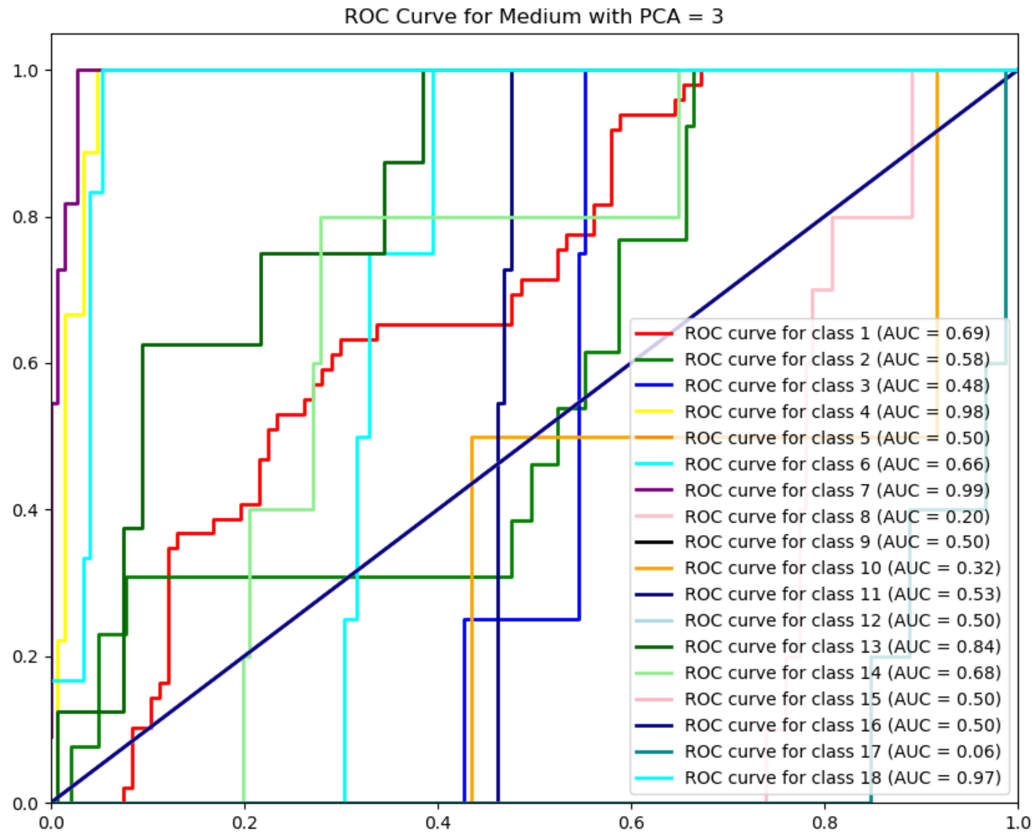
ROC Curve for Gene_Perturbation with PCA = 3
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve for class 1 (AUC = 0.89)
ROC curve for class 2 (AUC = 0.83)
ROC curve for class 3 (AUC = 0.66)
ROC curve for class 4 (AUC = 0.50)
ROC curve for class 5 (AUC = 0.84)
ROC curve for class 6 (AUC = 0.50)
ROC curve for class 7 (AUC = 0.72)
ROC curve for class 8 (AUC = 0.66)
ROC curve for class 9 (AUC = 0.50)
ROC curve for class 10 (AUC = 0.50)
ROC curve for class 11 (AUC = nan)
ROC curve for class 12 (AUC = nan)

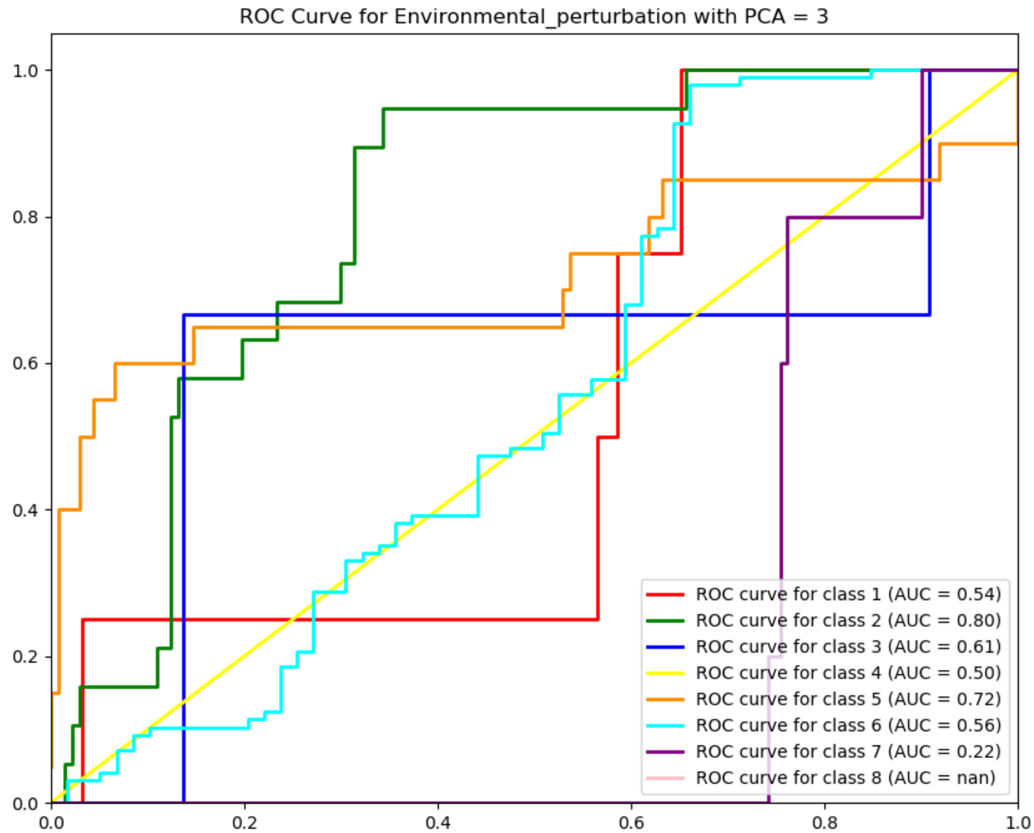

PR Curve for Medium class with PCA = 3
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Precision-recall for class 1 (area = 0.41)
Precision-recall for class 2 (area = 0.09)
Precision-recall for class 3 (area = 0.04)
Precision-recall for class 4 (area = 0.48)
Precision-recall for class 5 (area = 0.03)
Precision-recall for class 6 (area = 0.06)
Precision-recall for class 7 (area = 0.68)
Precision-recall for class 8 (area = 0.18)
Precision-recall for class 9 (area = 0.08)
Precision-recall for class 10 (area = 0.02)
Precision-recall for class 11 (area = 0.09)
Precision-recall for class 12 (area = 0.01)
Precision-recall for class 13 (area = 0.17)
Precision-recall for class 14 (area = 0.08)
Precision-recall for class 15 (area = 0.01)
Precision-recall for class 16 (area = 0.01)
Precision-recall for class 17 (area = 0.17)
Precision-recall for class 18 (area = 0.12)


PR Curve for Environmental_perturbation class with PCA = 3
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Precision-recall for class 1 (area = 0.03)
Precision-recall for class 2 (area = 0.25)
Precision-recall for class 3 (area = 0.03)
Precision-recall for class 4 (area = 0.05)
Precision-recall for class 5 (area = 0.83)
Precision-recall for class 6 (area = 0.74)
Precision-recall for class 7 (area = 0.07)
Precision-recall for class 8 (area = nan)


PR Curve for Gene_perturbation class with PCA = 3
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Precision-recall for class 1 (area = 0.03)
Precision-recall for class 2 (area = 0.09)
Precision-recall for class 3 (area = 0.02)
Precision-recall for class 4 (area = 0.02)
Precision-recall for class 5 (area = 0.10)
Precision-recall for class 6 (area = 0.01)
Precision-recall for class 7 (area = 0.91)
Precision-recall for class 8 (area = 0.04)
Precision-recall for class 9 (area = 0.05)
Precision-recall for class 10 (area = 0.04)
Precision-recall for class 11 (area = nan)
Precision-recall for class 12 (area = nan)


ROC Curve for Strain with PCA = 3
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve for class 1 (AUC = 0.50)
ROC curve for class 2 (AUC = 0.78)
ROC curve for class 3 (AUC = 0.46)
ROC curve for class 4 (AUC = 0.81)
ROC curve for class 5 (AUC = 0.50)
ROC curve for class 6 (AUC = 0.66)
ROC curve for class 7 (AUC = 0.64)
ROC curve for class 8 (AUC = 0.91)
ROC curve for class 9 (AUC = nan)
ROC curve for class 10 (AUC = 0.50)


ROC Curve for Medium with PCA = 3
ROC curve for class 1 (AUC = 0.69)
ROC curve for class 2 (AUC = 0.58)
ROC curve for class 3 (AUC = 0.48)
ROC curve for class 4 (AUC = 0.98)
ROC curve for class 5 (AUC = 0.50)
ROC curve for class 6 (AUC = 0.66)
ROC curve for class 7 (AUC = 0.99)
ROC curve for class 8 (AUC = 0.20)
ROC curve for class 9 (AUC = 0.50)
ROC curve for class 10 (AUC = 0.32)
ROC curve for class 11 (AUC = 0.53)
ROC curve for class 12 (AUC = 0.50)
ROC curve for class 13 (AUC = 0.84)
ROC curve for class 14 (AUC = 0.68)
ROC curve for class 15 (AUC = 0.50)
ROC curve for class 16 (AUC = 0.50)
ROC curve for class 17 (AUC = 0.06)
ROC curve for class 18 (AUC = 0.97)
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve for Environmental_perturbation with PCA = 3
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve for class 1 (AUC = 0.54)
ROC curve for class 2 (AUC = 0.80)
ROC curve for class 3 (AUC = 0.61)
ROC curve for class 4 (AUC = 0.50)
ROC curve for class 5 (AUC = 0.72)
ROC curve for class 6 (AUC = 0.56)
ROC curve for class 7 (AUC = 0.22)
ROC curve for class 8 (AUC = nan)