ECS 171 HW3

Wenhao Su 915297212

Prob1

I used Lasso. Lasso applies “absolute value of magnitude” of coefficient as penalty to the weights so that we could train our model with feature selections.

alpha 1e-06 has following mean MSE(5 Fold) of: 0.03690268648076217 and #features of 290

alpha 1e-05 has following mean MSE(5 Fold) of: 0.024802461496152246 and #features of 158

alpha 0.0001 has following mean MSE(5 Fold) of: 0.024895205907459535 and #features of 107

alpha 0.0005 has following mean MSE(5 Fold) of: 0.029237525453356544 and #features of 69

alpha 0.001 has following mean MSE(5 Fold) of: 0.043294803001845523 and #features of 50

alpha 0.01 has following mean MSE(5 Fold) of: 0.0605534235908767 and #features of 3

so alpha 1e-05 has the lowest MSE error and #number of features is 158

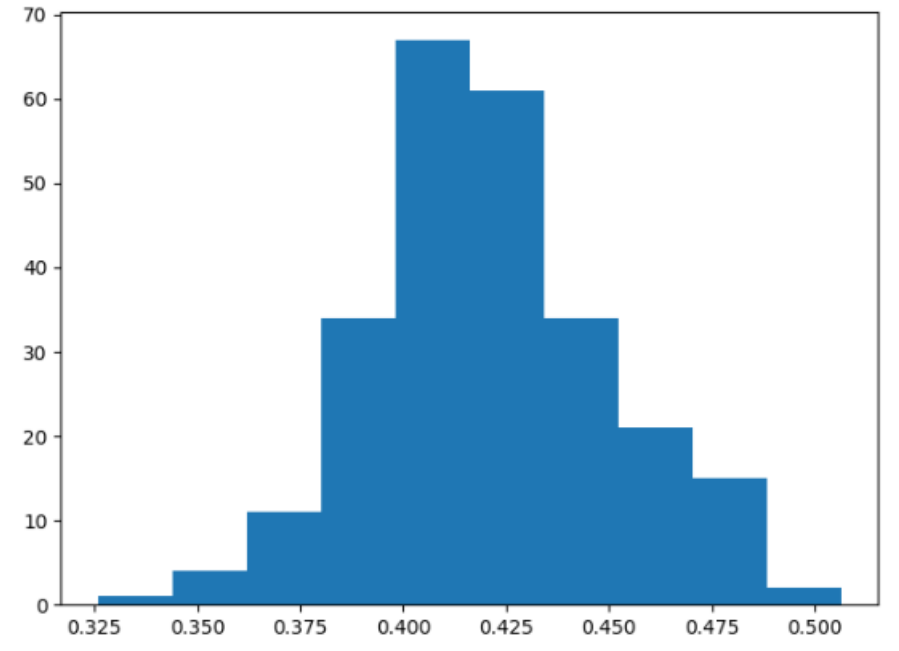
Prob2 Function implemented in prob2.py with extension of confidence interval

Prob3

methodology: Choose 30% percent of the samples as a training set and predict it with 1D vector based on the mean of all samples, get 250 predicted values and form a distribution.

assumption: dataset mean equals to predictions mean

95.0% confidence interval 0.367 and 0.484

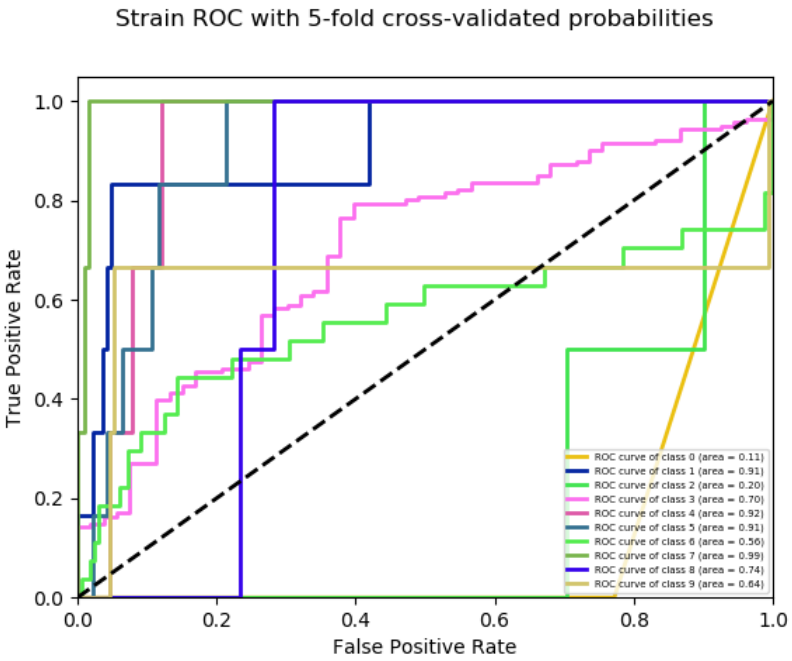


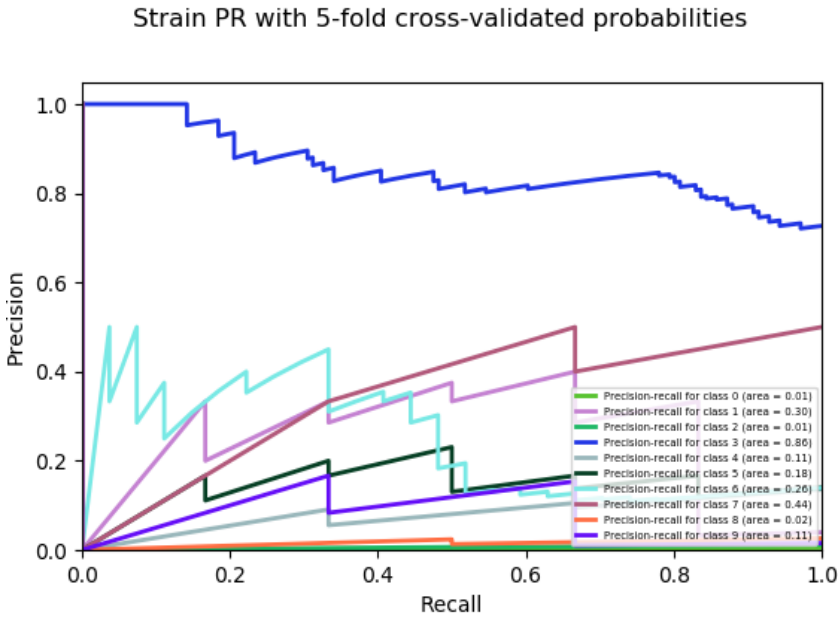
Prob4

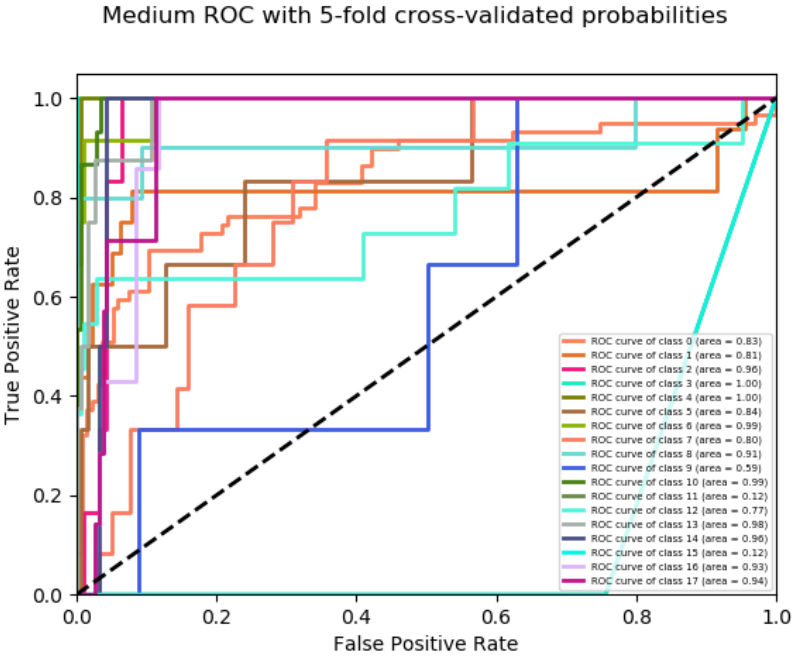
Mean AUC for two individual separate Medium and Perturb classifiers is 0.7764990965822531

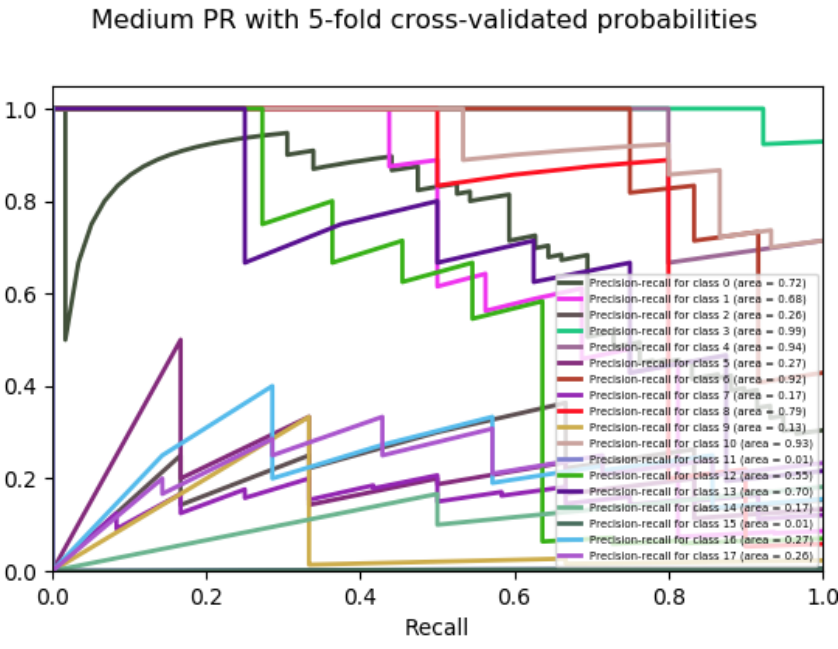
Mean AUPRC for two individual separate Medium and Perturb classifiers is 0.34215973748661166

Use Parameter fold=5

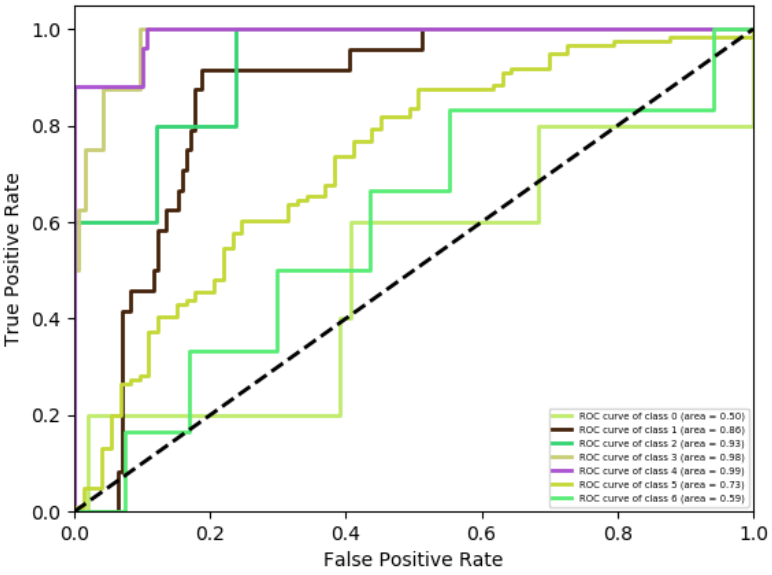




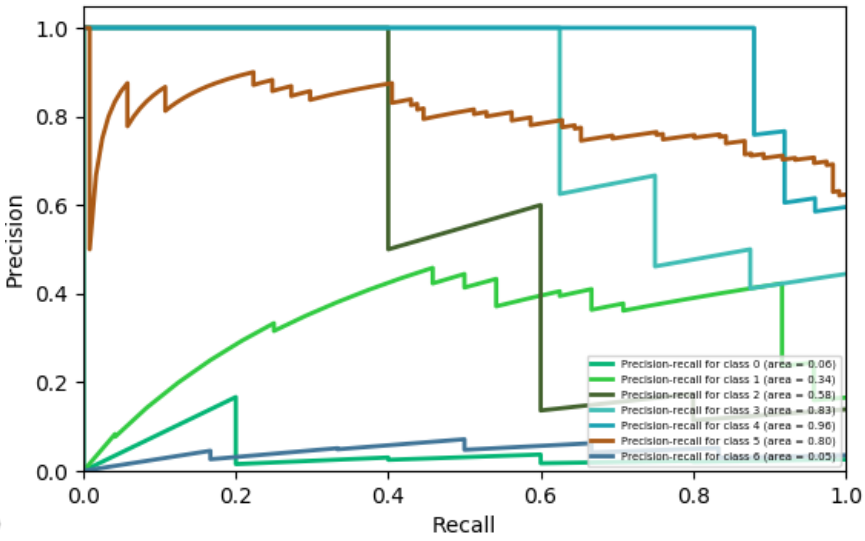




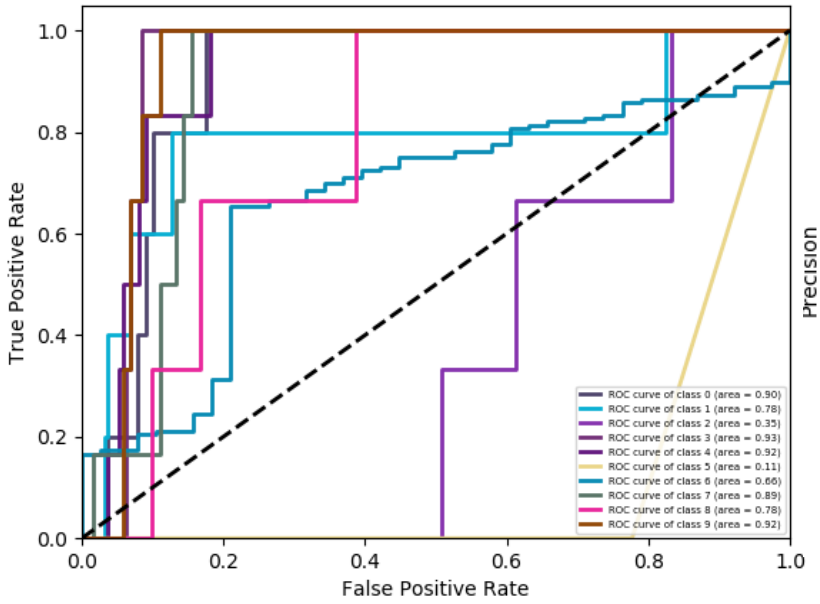
Stress ROC with 5-fold cross-validated probabilities



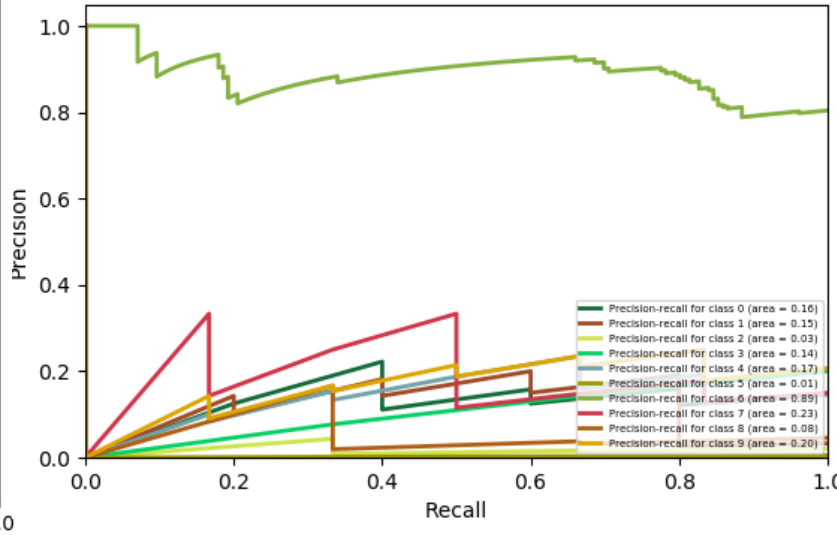
Stress PR with 5-fold cross-validated probabilities



Perturb ROC with 5-fold cross-validated probabilities



Perturb PR with 5-fold cross-validated probabilities



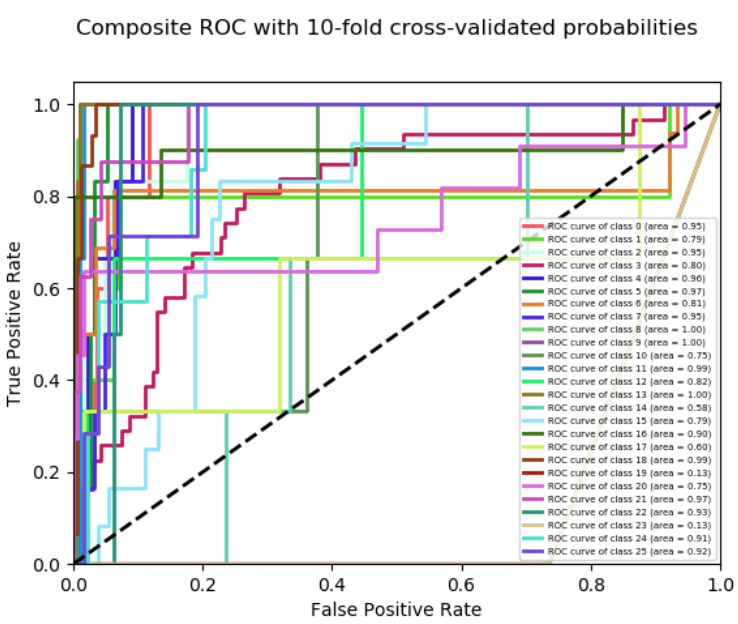
Prob5

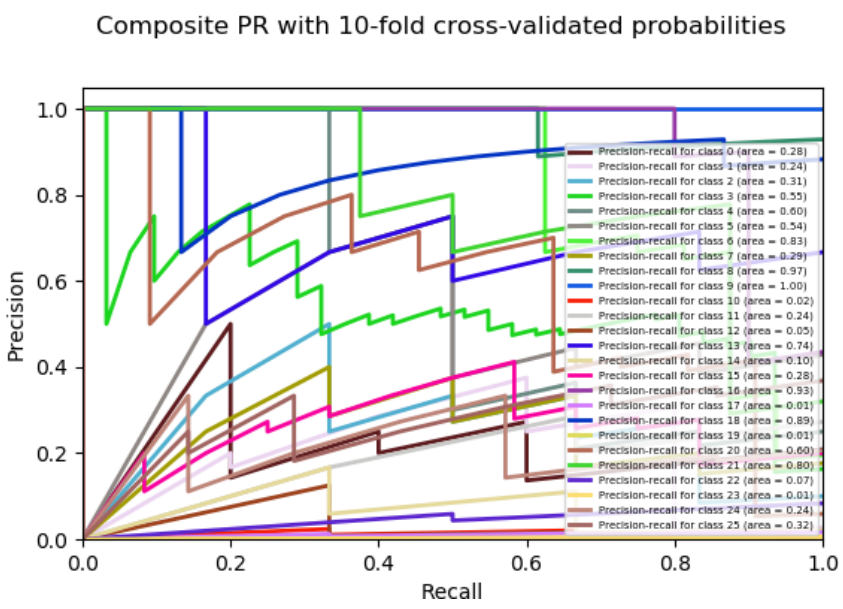
Mean AUC for two individual separate Medium and Perturb classifiers is (Use parameter fold = 10)

0.7764990965822531

Mean AUPRC for two individual separate Medium and Perturb classifiers is (Use parameter fold = 10)

0.34215973748661166





Mean AUC for all classes for simultaneous Medium and Perturb predictor is 0.8293204406469722

Mean AUPRC is 0.42048175776730895

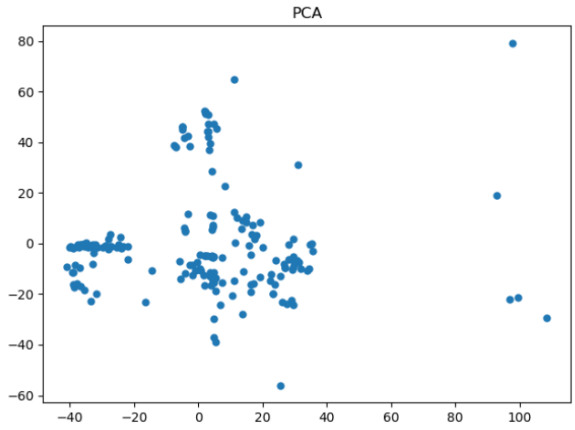
Therefore, it performs better than individual separate predictors, the composite also performs better than **Baseline performance** for the simple reason that baseline ROC curve is always below the majority of ROC of SVM. The precision is 0.16 which is always much lower for SVM Mean AUPRC.

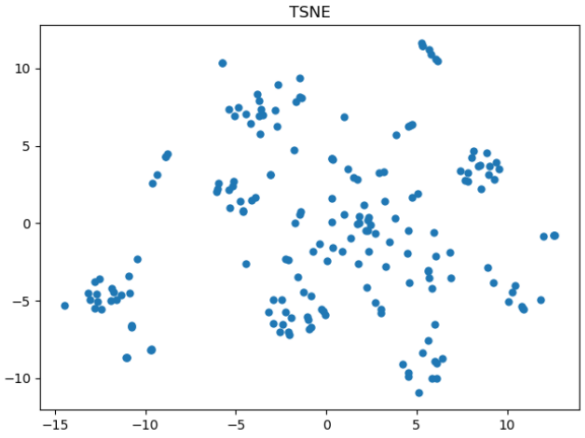
Baseline Most Frequent Label is 4 (MD001, na_WT)

Total #number of True Positive is 31

Precision for baseline is 0.15979381443298968

Prob6





Prob7

The measurement is using macro mean of each four classes.

Feature Selection performs better.

AUC	Strain	Medium	Stress	Perturb
Selection	0.666987	0.826806	0.785025	0.726192
PCA	0.413299	0.568566	0.601098	0.435965
TSNE	0.525602	0.594797	0.65248	0.459534

AUPRC	Strain	Medium	Stress	Perturb
Selection	0.227039	0.484483	0.522093	0.199837
PCA	0.123439	0.149326	0.203377	0.100456
TSNE	0.100456	0.186318	0.231389	0.110626