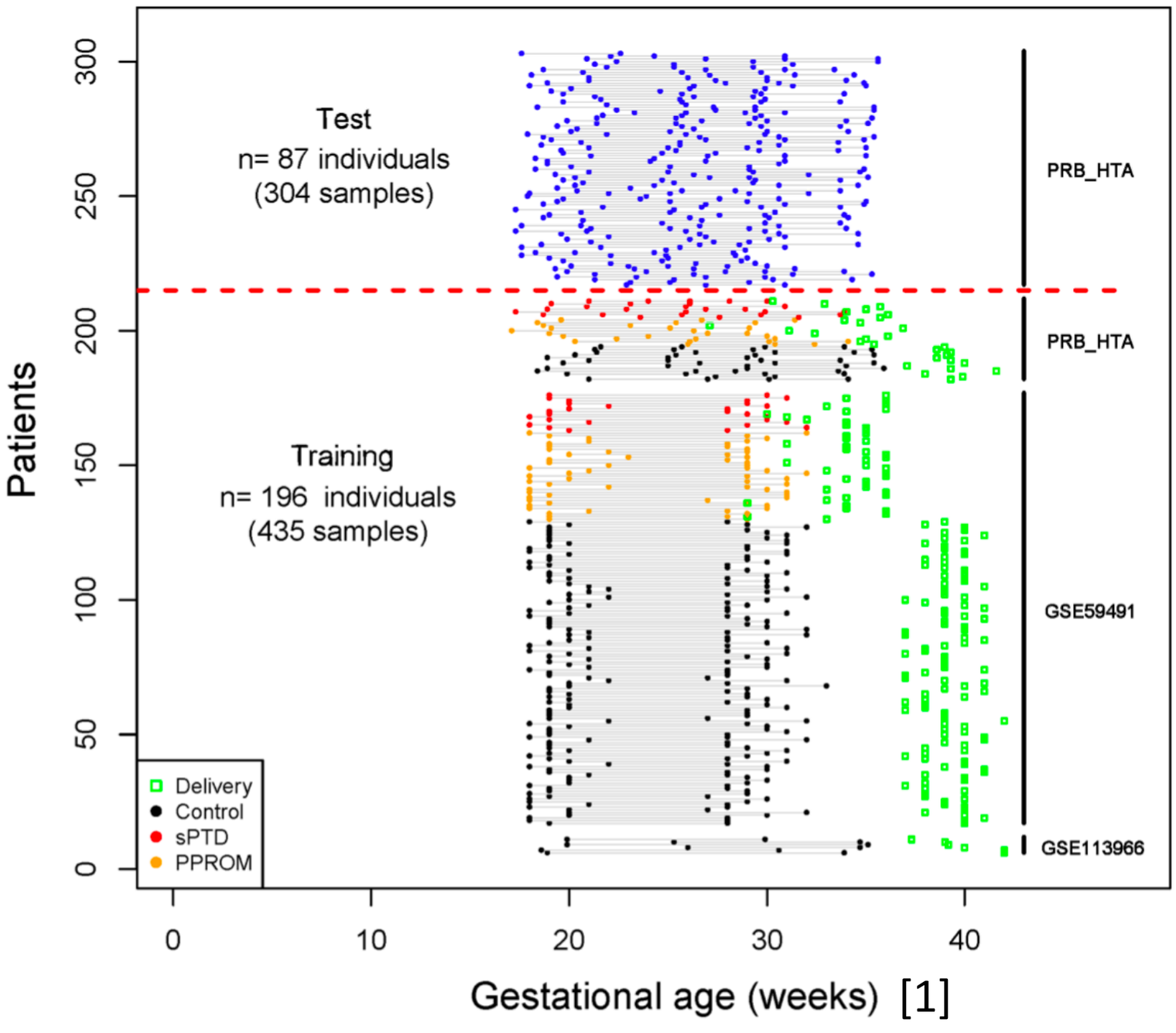
2nd Project Goal:

Build two classifiers that can predict preterm birth, **Control vs spontaneous preterm delivery (sPTD)** and **Control vs preterm premature rupture of membranes (PPROM)**, given the maternal whole blood transcriptome and metadata.

Subchallenge 2, Submission closed on December 5th.

DREAM Challenge Data and Restrictions



<u>Microarray gene expression data:</u>

- **29459 genes** are in the matrix.
- The test/validation set contains 304 samples from 87 patients.
- The training set consists of 435 samples (**285 Control, 55 sPTD, and 95 PPROM**) from 196 patients.
- Notice that this is an imbalanced data.

<u>Metadata:</u>

- It contains **gestational age**, group label, microarray platform, and data sources.

<u>Restriction:</u>

- Develop parsimonious models using <100 unique genes.
- Allowed to make max 2 submissions.

Method

1. Select <100 genes using a volcano plot. We used fold change and p-value of t-test between control and experimental in the training data.

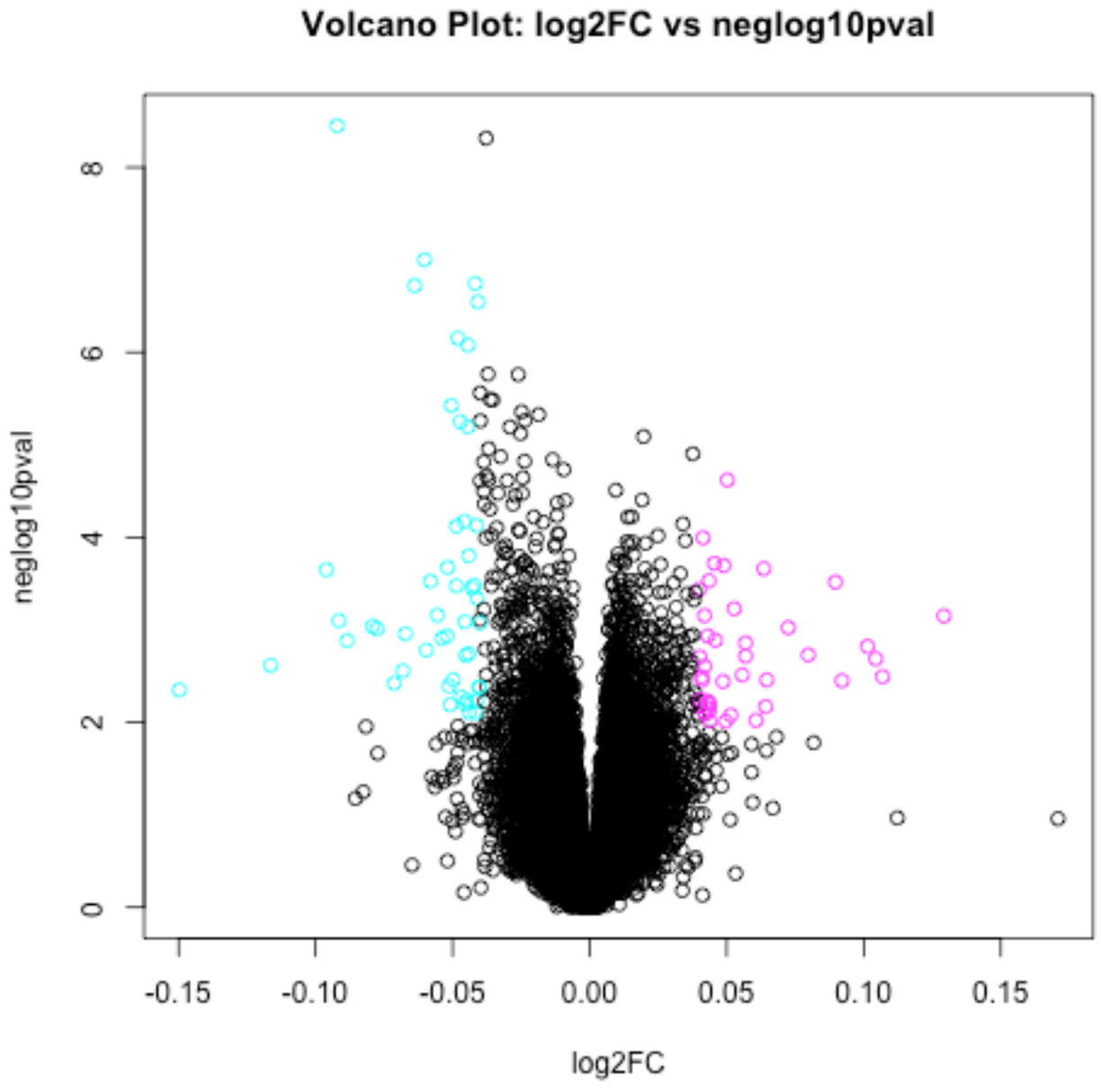
2. Train a denoising autoencoder (a neural network model) for dimensionality reduction using all samples with selected features.

3. Combine the features from the mid layer of autoencoder with gestational age information from the metadata.

4. Train multiple models and assess their performance using training data using 5-fold cross-validation.

5. Make a prediction using test/validation data.

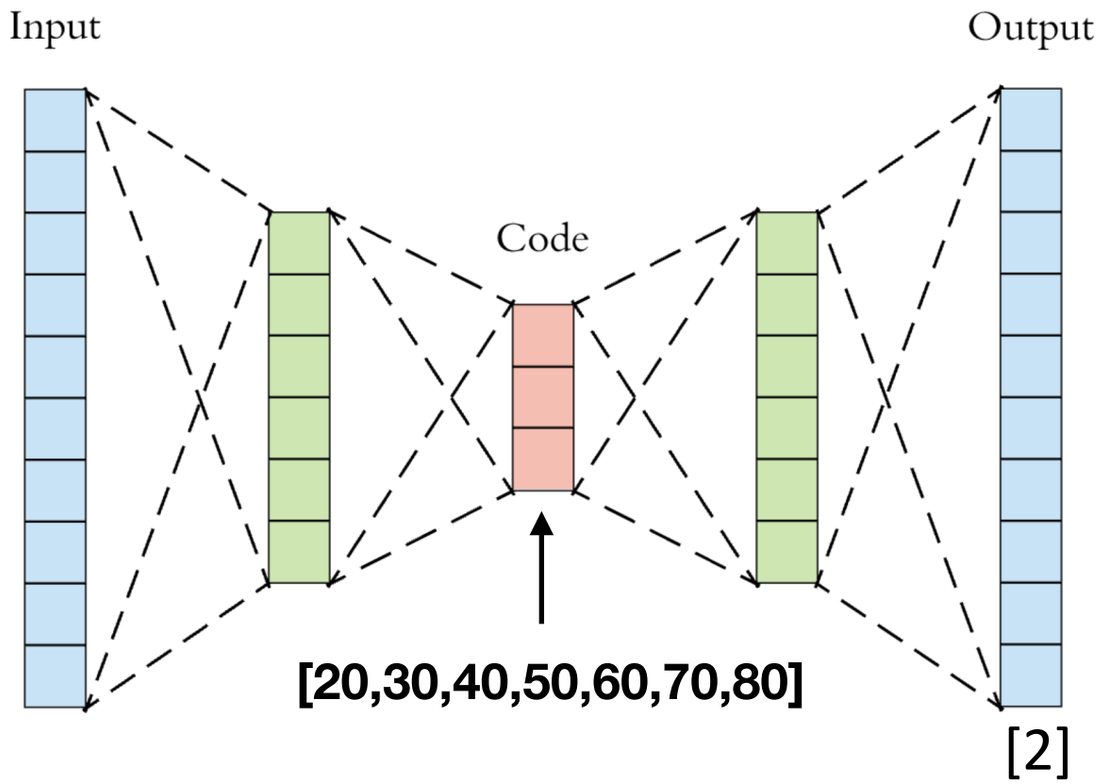
Selection of <100 genes



91 genes were selected by
fold change and p-value from Student's t-test
between Control and Experimental (sPTD + PPROM).

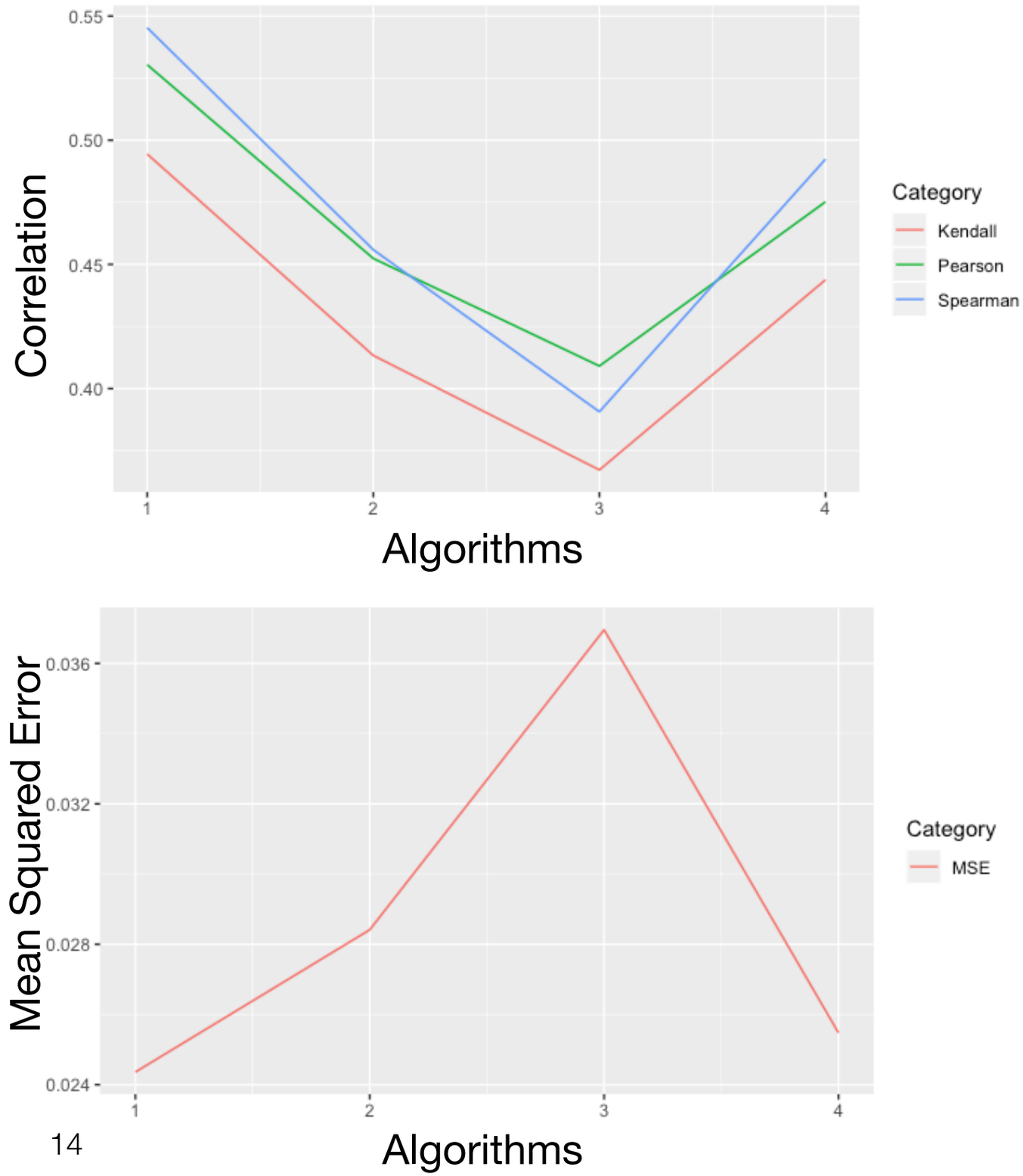
Training Denoising Autoencoder

Fixed:
Activation function = 'relu'
Optimization function = ‘adam’
hidden_size1 = 200
hidden_size2 = 80



Algorithms:

1. **7-layer autoencoder with Noise=0.4**
2. **5-layer autoencoder with Noise=0.4**
3. **3-layer autoencoder with Noise=0.4**
4. **7-layer autoencoder with Noise=0.2**



Prediction Result

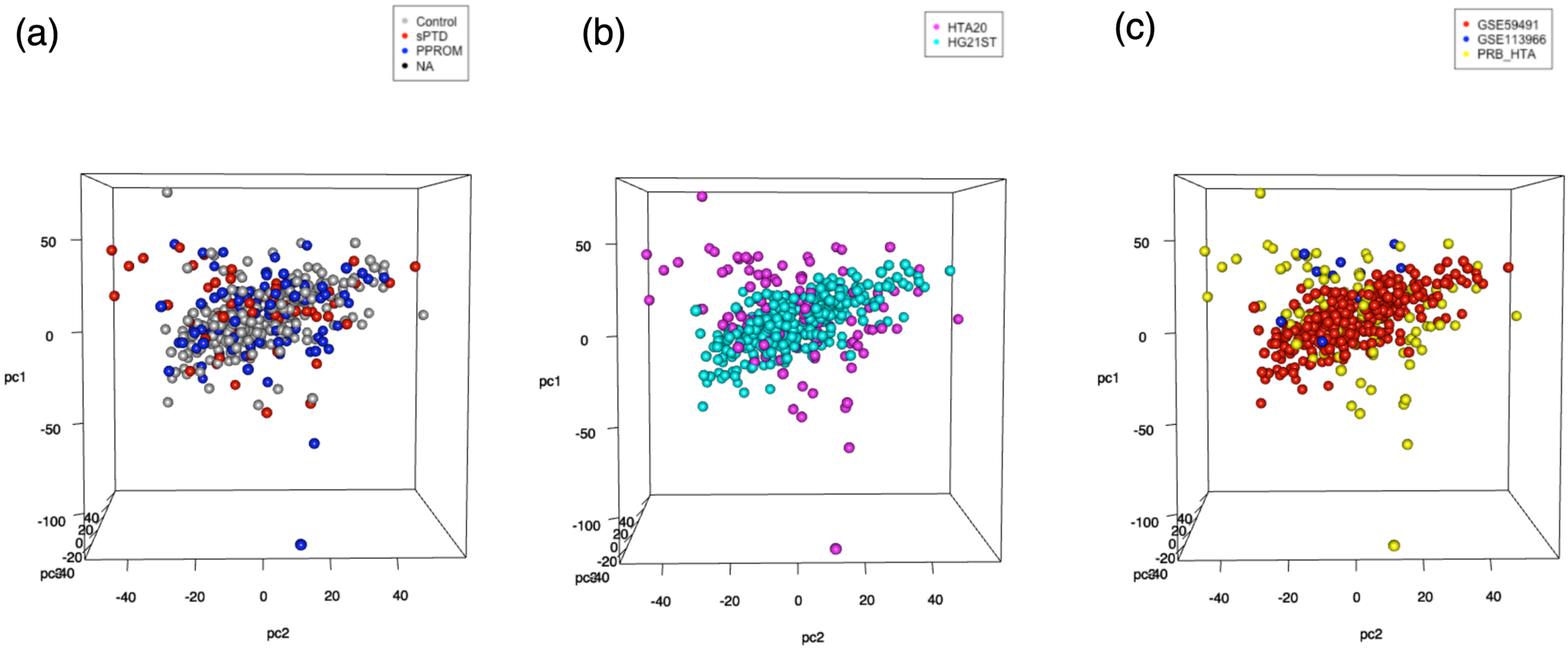
- Since we had an imbalanced data, we tried to add weights for the groups. This improved the F1 score, but not the accuracy.

- Our final model is the Logistic regression.

Submission Id	Submitter	Status	sPTD_AUC	sPTD_AUPR	PPROM_AUC	PPROM_AUPR	mean
9696251	Team ZO	SCORED	0.4433	0.3704	0.7438	0.6967	0.5636

- Our result ranks the 22th out of 122 submissions made. This can be verified on the Challenge website leaderboard.

Preliminary Visualization using Top 5000 genes with high variance and PCA



- No batch effects given three different sources and two different microarray platforms.
- No visible clusters by groups even after using PCA as a dimensionality reduction method.

Training Two Classifiers for Prediction

- 5-Fold Cross Validation
- Weights for mildly skewed data
  - 285 Control and 55 sPTD (1:3)
  - 285 Control and 95 PPROM (1:5)
- Regularization Parameter: C = [1, 3, 7, 10, 30, 100]
- **Logistic Regression:** Penalty = “L2”, Solver = “newton-cg”
- **Support Vector Machine:** Kernels = [“linear", “poly", rbf”]
- **Random Forest Classifier:** max_depth=5, criterion="entropy"

Future Directions

- The challenge organizers noticed that some teams cleverly used only a subset of provided data using the gestational age distribution. These teams got the top scores.

- While the organizers acknowledge that the challenge rule had no restriction on using a subset or whole data, they decided not to award teams based on the current leaderboard scores because this approach was not going to help solving their research question.

- They gave a new deadline of 01/05/2020 for a new submission of codes.

- “Briefly, we propose to use the analysis scripts that you are expected to provide (per challenge rules) and we will train and test the resulting models under several scenarios in which training and test sets do not feature differences in the GA sampling distributions.”

- This will give Team ZO an opportunity to improve the results.