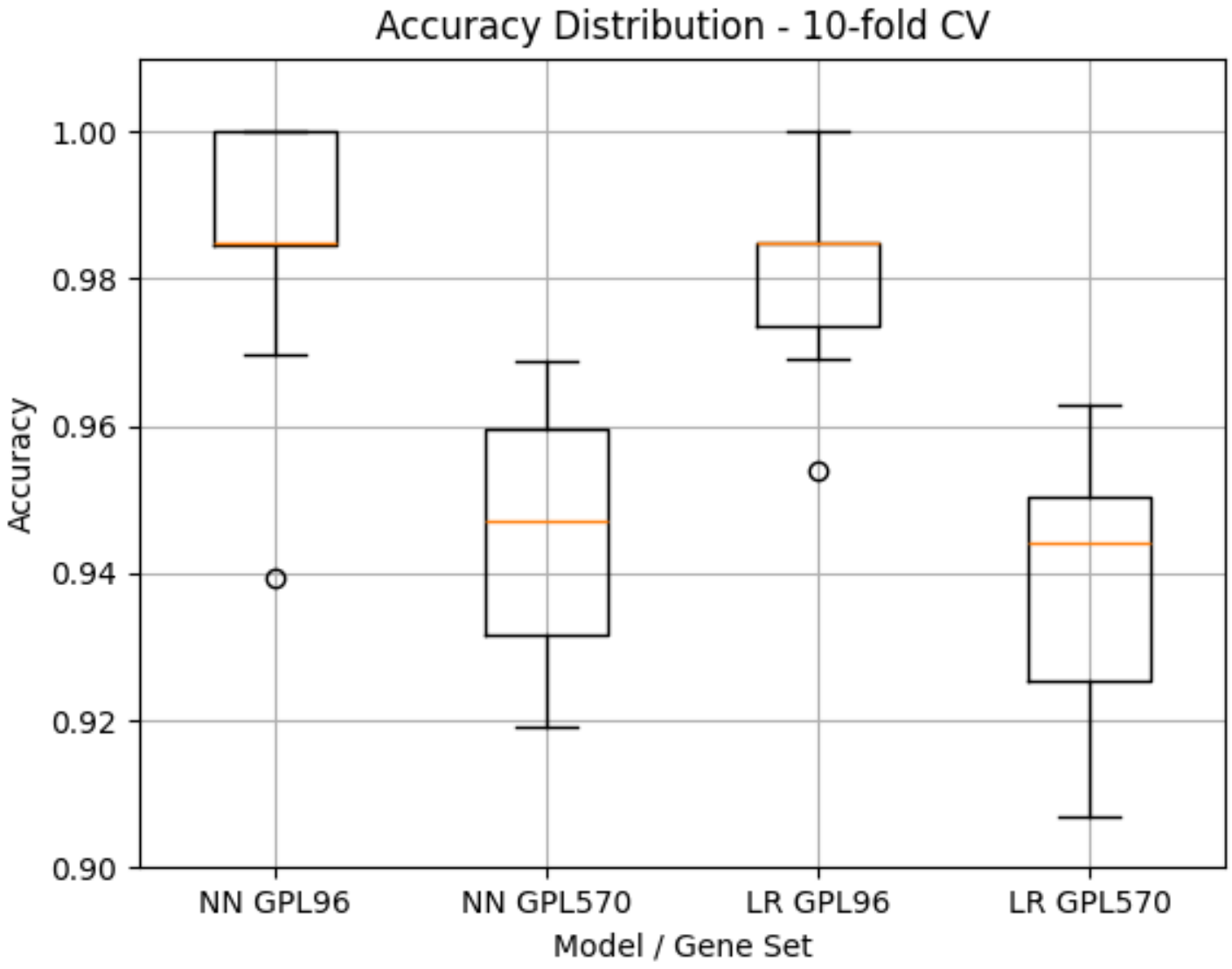
Background

- Microarrays measure the expression of thousands of genes simultaneously.
- Microarray data is being used for early diagnosis of cancer.
- Researchers are struggling to extract meaningful information from so much data.
- We propose an application of machine learning to make diagnoses from microarray data.

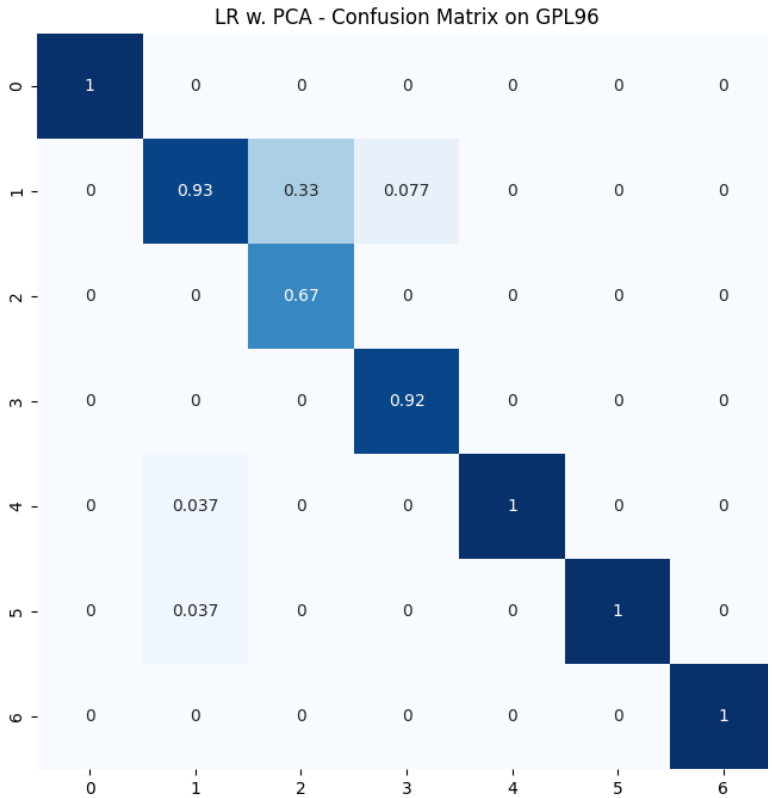
Procedure

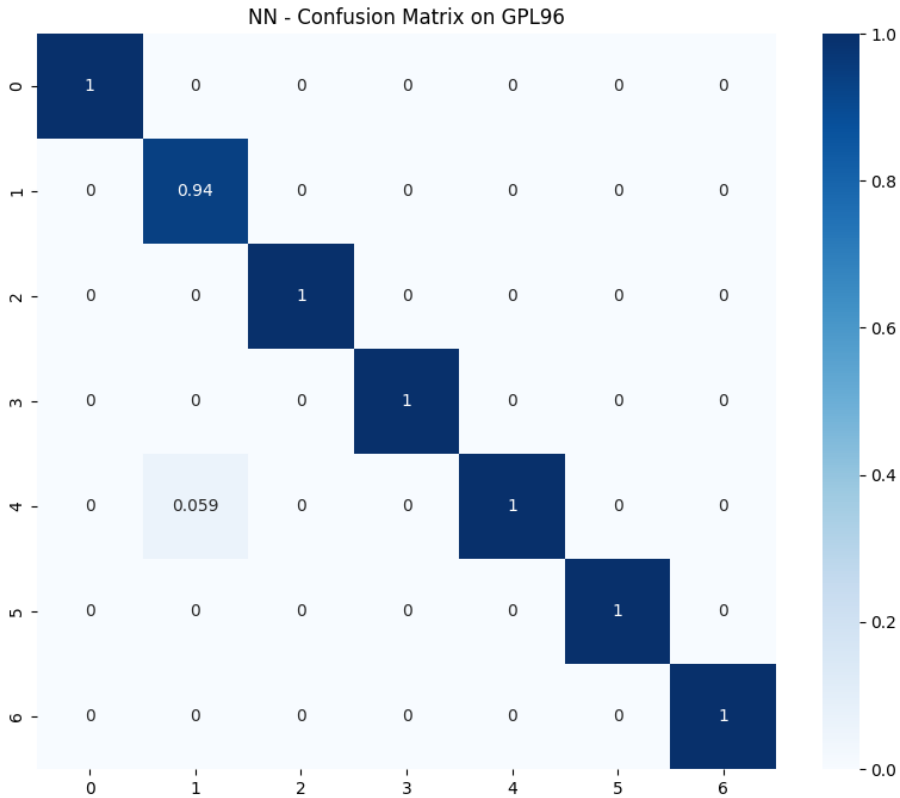
- We will perform multi-class classification to diagnose samples as healthy or cancerous and determine what type of cancer.
- We will compare traditional ML (logistic regression with PCA) and deep learning.
- We will compare different set of genes (as features) to determine which has more diagnostic power.

Cross Validation Results

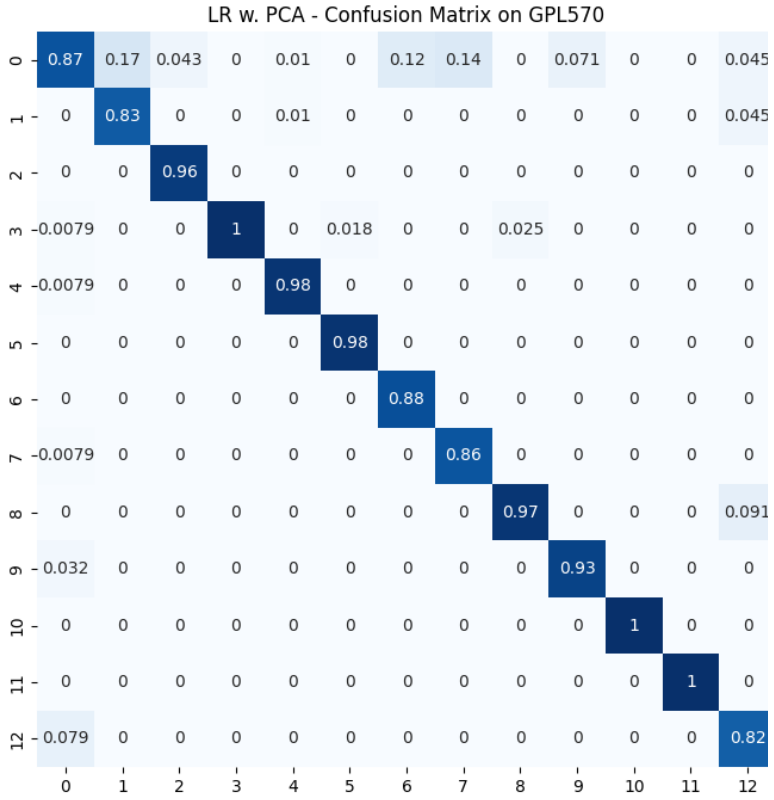


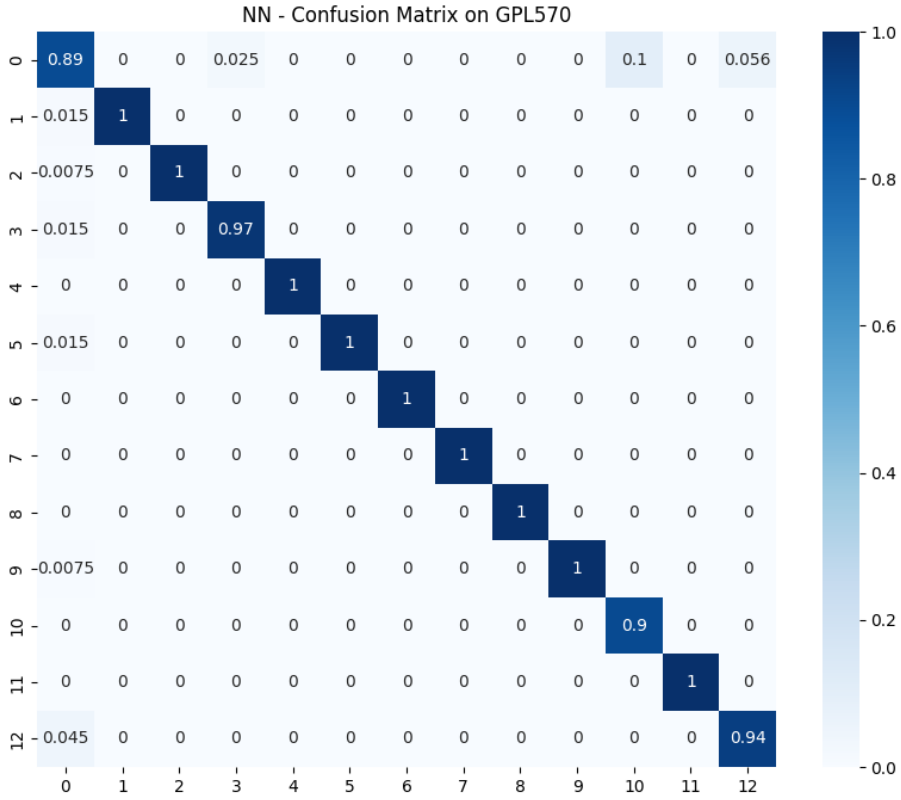
Confusion Matrix - GPL96



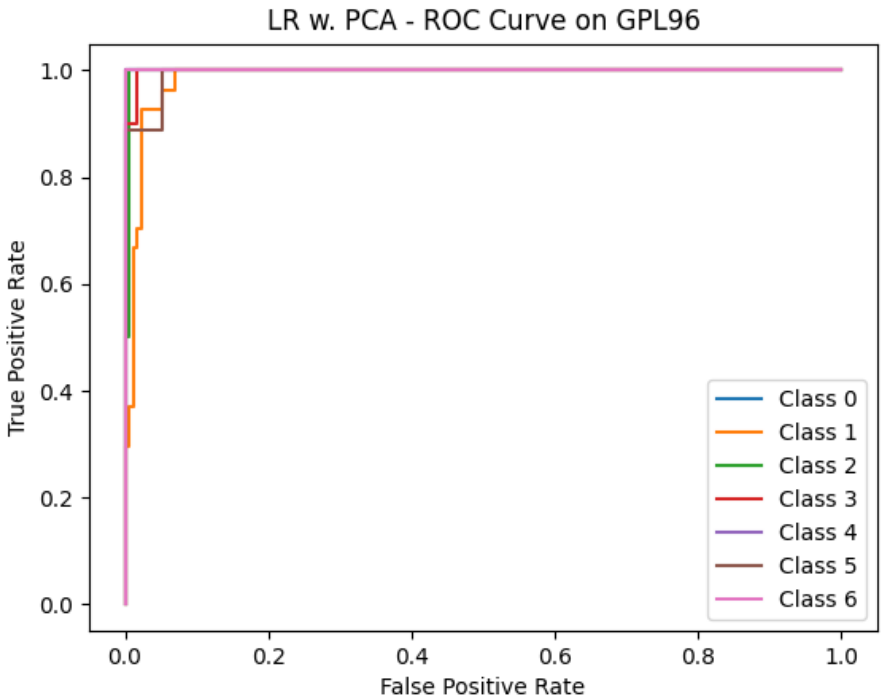


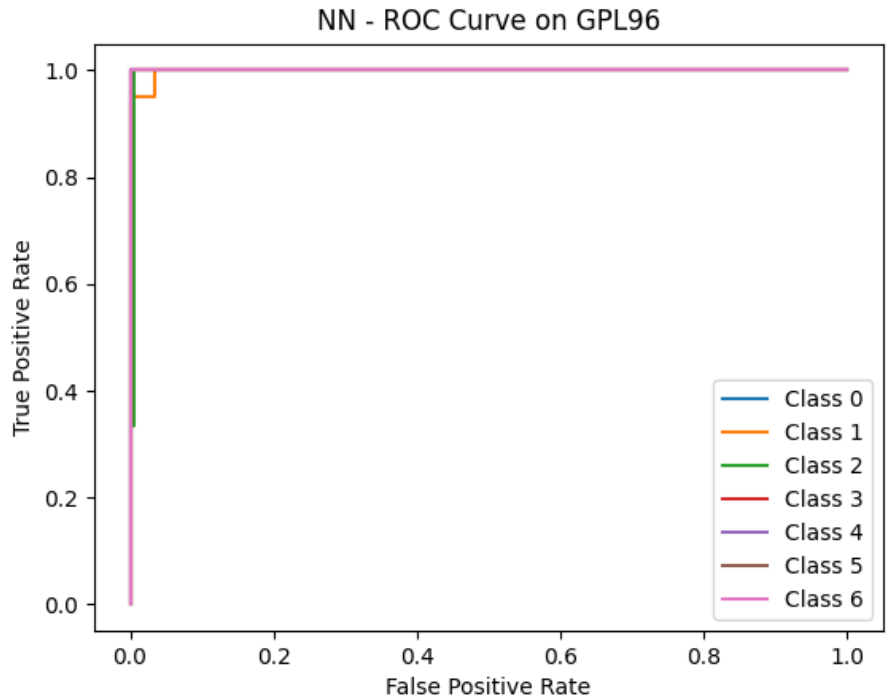
Confusion Matrix - GPL570



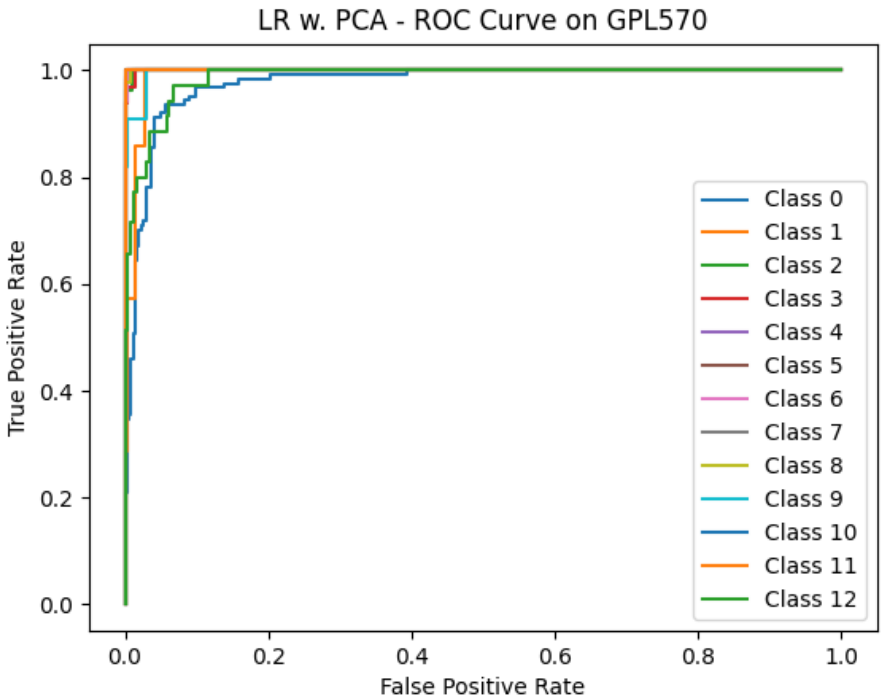


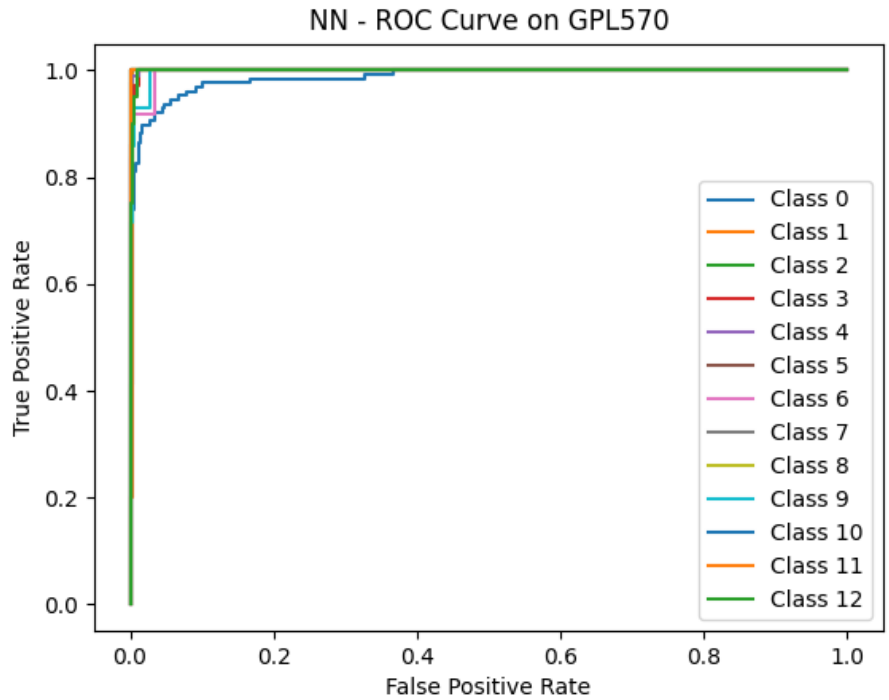
ROC Curve - GPL96





ROC Curve - GPL570





Open Questions and Next Steps

- If we restrict the GPL570 to the 7 classes present in GPL96, will performance improve? To address the information content of the gene sets, we need to perform this experiment.
- "Beautify" plots by adding labels to each class (healthy, lung cancer, etc.).
- Draw conclusions and prepare a presentation.