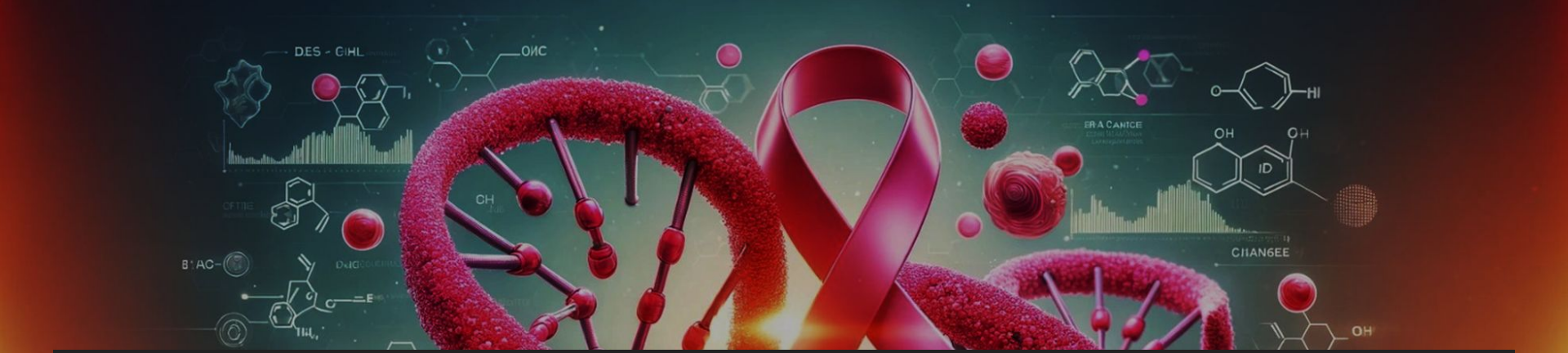
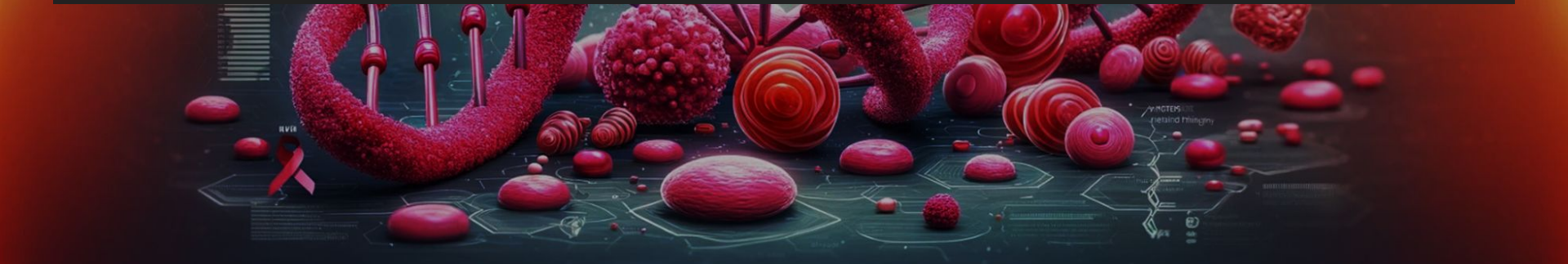


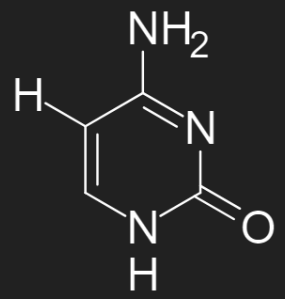
The Relation of Methylation to Breast Cancer

Adi Rabinovitz, Avi Kupinsky, Itay Ottenheimer,
Mia Segev Gal, Shay Cohen

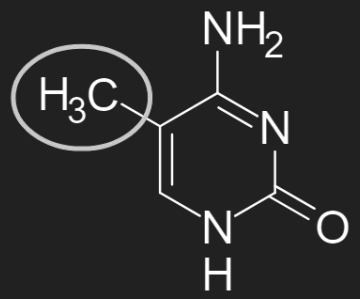


Methylation

- Methylation: TCGA
- Tags act like switches, turning genes on or off.
- Controls which proteins are made and when.
- Changes in methylation can affect health and diseases.
- Important for understanding how our bodies work.



cytosine



methylated cytosine

Project Overview

- Breast cancer affects both women and men, though it's more prevalent in women
- Our project aims to investigate the connection between methylation and breast cancer
- Our goal was to identify significant differences in methylation patterns of CpG sites between these groups.

Information about the Data

- Source: TCGA BRCA Dataset
- Properties: CpG sites named with methylation rates for each sample
- Scope:
  - Samples collected from healthy breast cells and tumors
  - Divided into two datasets: healthy (98 samples) and cancerous (722 samples)
- Chromosome 19 dataset:
  - 24,671 rows representing CpG sites across chromosome 19
- Whole genome dataset:
  - 453,526 rows representing CpG sites across the entire genome

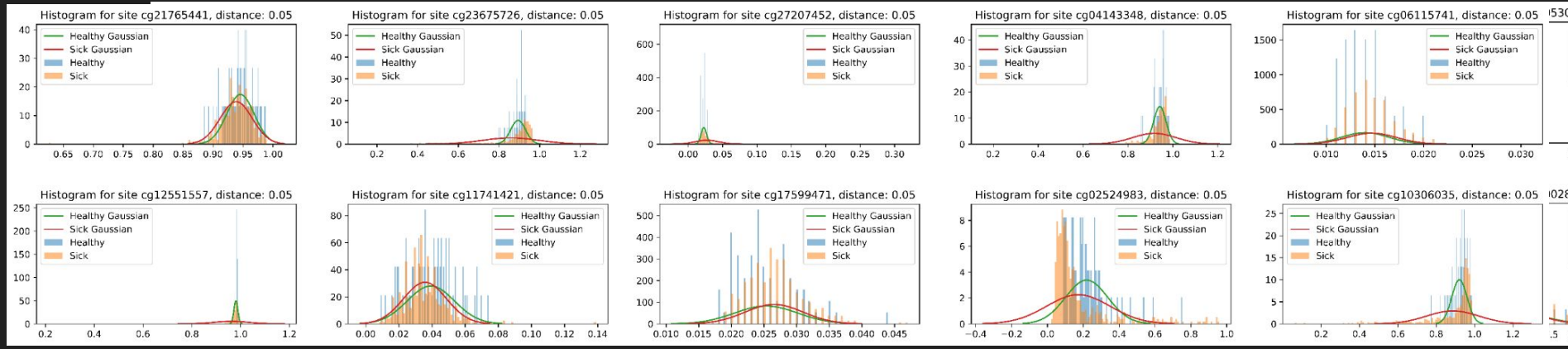
Feature selection

- Method 1: Select 100 features that have the highest Wasserstein distance.
- Method 2: Select 100 random features.
- Method 3: Select statistically significant features with p-value < 0.05 under the statistical model, later refined using the FDR method.

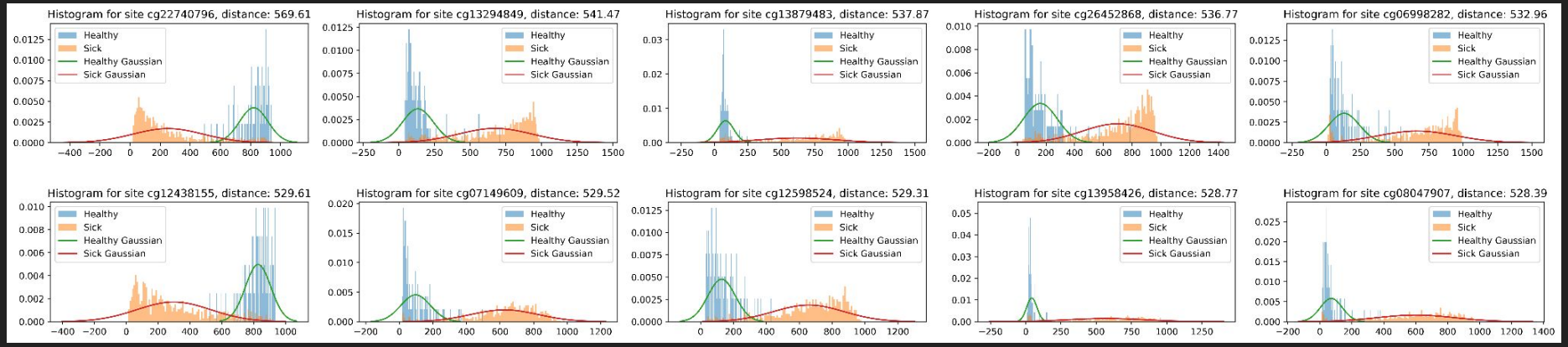
Failed attempt

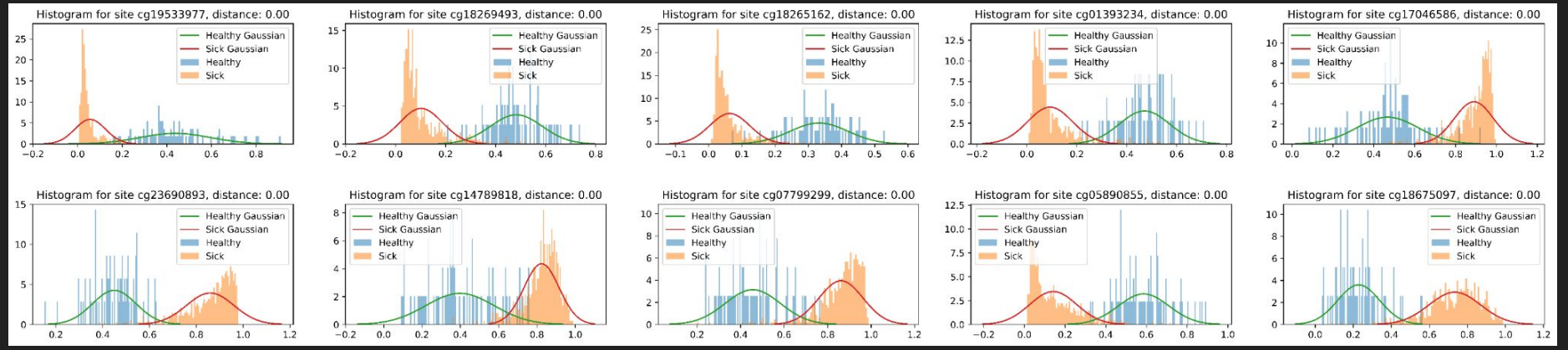
We tried to use M-values to calculate a more accurate t-test results, but after calculating them the FDR method returned ~300k features.

Least Significant Features

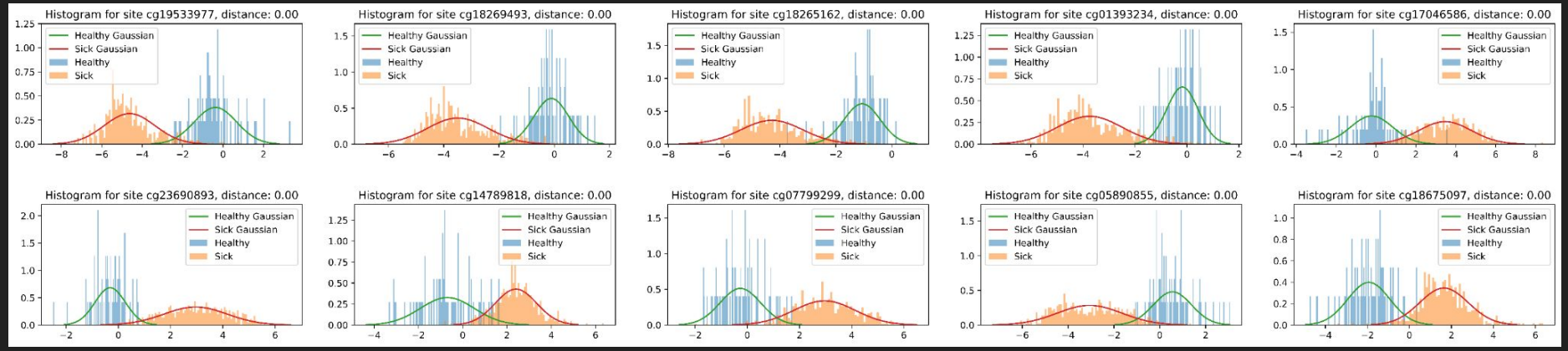


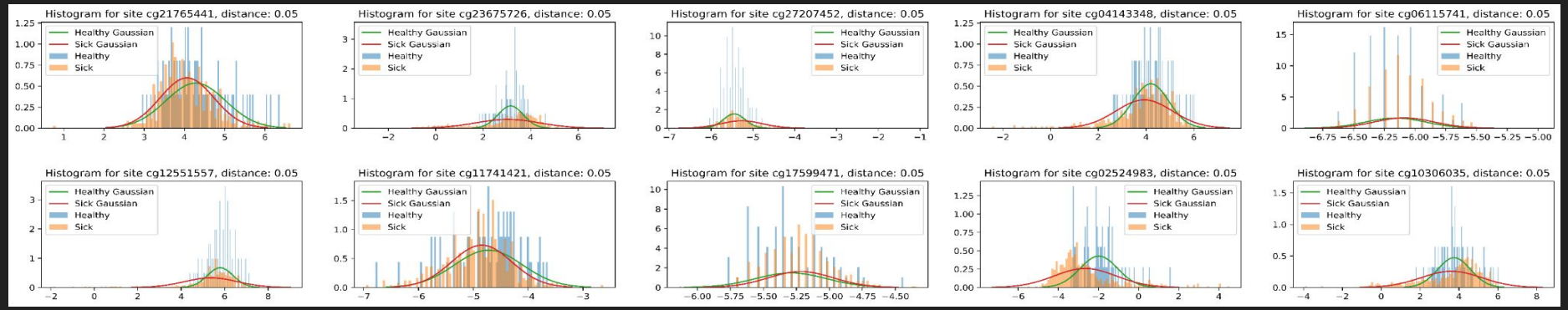
Most Significant features





Most/least Significant features - using M-values



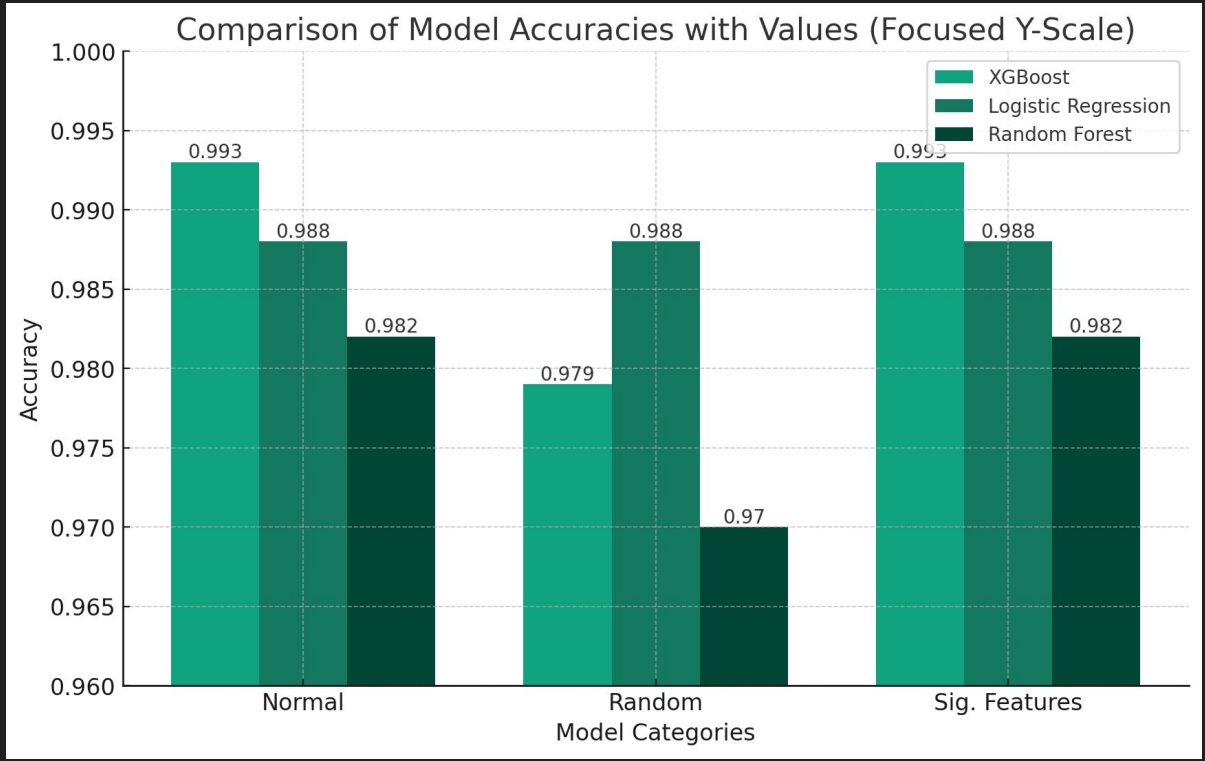


Classification Model Training

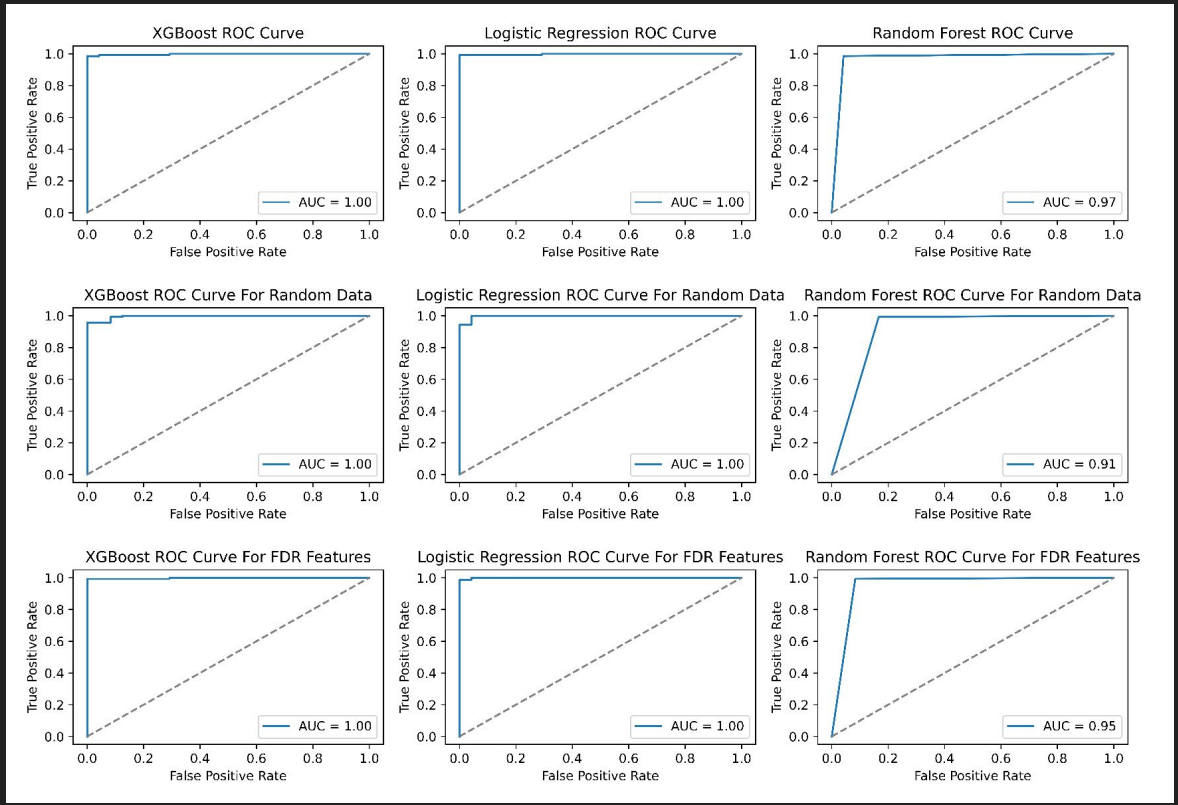
We have trained and tested the results of the following models

- XGboost
- Logistic Regression
- Random Forest

Results

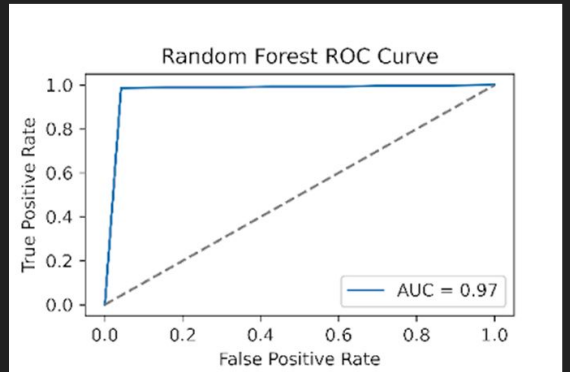

Comparison of Model Accuracies with Values (Focused Y-Scale)
XGBoost
Logistic Regression
Random Forest
Accuracy
1.000
0.995
0.990
0.985
0.980
0.975
0.970
0.965
0.960
0.993
0.988
0.982
0.979
0.988
0.97
0.993
0.988
0.982
Normal
Random
Sig. Features
Model Categories

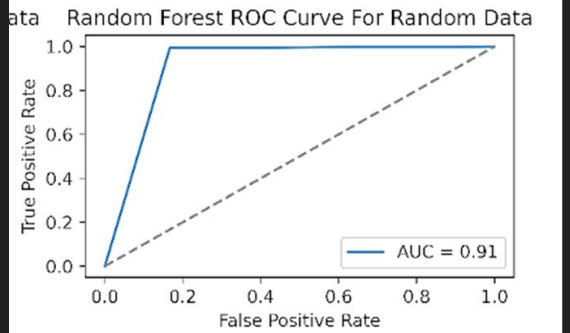
Results


XGBoost ROC Curve
AUC = 1.00
Logistic Regression ROC Curve
AUC = 1.00
Random Forest ROC Curve
AUC = 0.97
XGBoost ROC Curve For Random Data
AUC = 1.00
Logistic Regression ROC Curve For Random Data
AUC = 1.00
Random Forest ROC Curve For Random Data
AUC = 0.91
XGBoost ROC Curve For FDR Features
AUC = 1.00
Logistic Regression ROC Curve For FDR Features
AUC = 1.00
Random Forest ROC Curve For FDR Features
AUC = 0.95
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

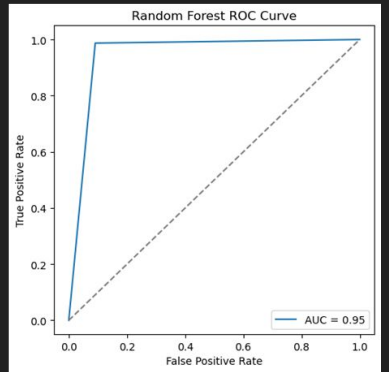
BRCA Compared to COAD

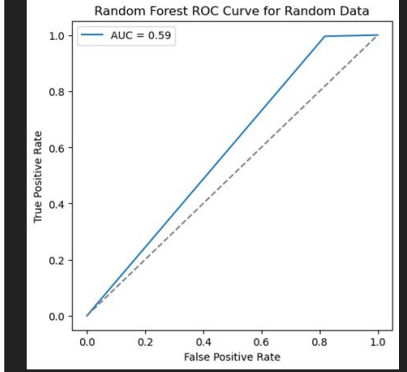
BRCA





COAD



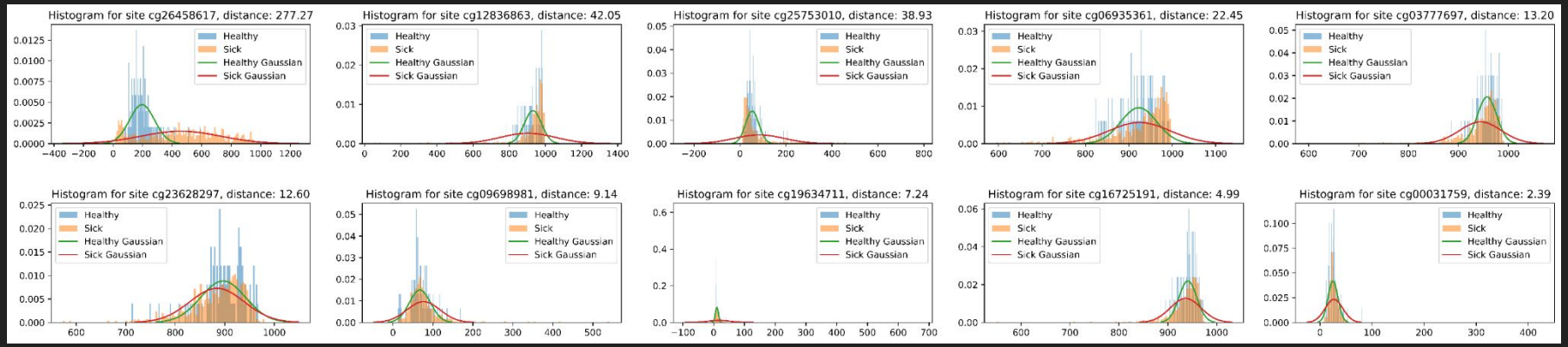


BRCA1 and BRCA2

- BRCA1 and BRCA2 are proteins, both are expressed in the breast and other tissue cells, where they help repair damaged DNA, or destroy cells if DNA cannot be repaired.
- If BRCA1 or BRCA2 itself is damaged by a BRCA mutation, damaged DNA is not repaired properly, and this increases the risk of breast cancer.

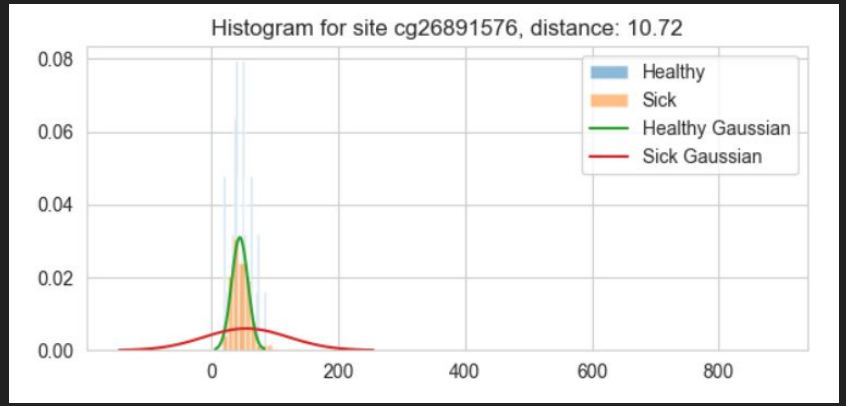
BRCA1 and BRCA2

Results for BRCA2:



BRCA1 and BRCA2

analyze the CpG site：cg26891576, this site is in the TSS200 range



Conclusion

- All models **achieved high accuracy**, with XGBoost and Logistic Regression showing slightly better performance.
- The consistency in model accuracy across normal and significant feature datasets indicates that methylation patterns at these CpG sites are highly indicative of breast cancer status, supporting the hypothesis that altered methylation is a hallmark of cancerous transformation in breast cells.

Conclusion

- By employing robust machine learning methods with data from both healthy individuals and those diagnosed with breast cancer, we can develop a model surpassing pathologists in detecting breast cancer.

Future steps

1. Weighting the samples in the models could help address the imbalance between sick and healthy samples.
2. Exploring other cancer types could provide valuable insights into DNA methylation patterns in breast cancer patients.