CBIO Hackathon: The Relation of Methylation to Breast Cancer

Adi Rabinovitz (314952540, adi_1209) Avi Kupinsky (318336070, avikupinsky)
Itay Ottenheimer (209493519, itayotten) Mia Segev Gal (208730960, miasegevgal)
Shay Cohen (207283094, shaycohen98)

1. Project Overview

Breast cancer is a type of cancer that forms in the cells of the breasts, potentially affecting both women and men, but is more common in women.
In our project, we aim to explore the relationship between methylation and breast cancer by analyzing a dataset that includes both healthy and unhealthy breast cell samples. Our objective is to determine if the methylation patterns of CpG sites differ significantly between these two groups. By identifying which CpG sites serve as critical features in this context, we intend to develop a classifier capable of predicting the health status of new cell samples based on their methylation profiles.

1.1. Hypothesis

Our hypothesis suggests a correlation between the methylation status of specific CpG sites and the development of breast cancer cells. We theorize that a different methylation rate at these critical sites may increase the likelihood of cell transformation from healthy to cancerous. Our hypothesis is based on how methylation affects gene activity. We think that unusual patterns of methylation might be significant for the development of breast cancer.

2. Data

2.1. Source: BRCA Dataset by TCGA [ref 12.8].

2.2. Properties: CpG sites name with the rate of methylation for each sample.

2.3. Scope

- The samples were taken from healthy breast cells and from a Tumor.
- The data is divided into two datasets: healthy and cancerous.
- In the healthy datasets there are 98 samples.
- In the cancerous datasets there are 722 samples.
- <u>Chromosome 19 dataset:</u>
    - There are 24671 rows representing the CpG sites across the genome in chromosome 19.
- <u>Whole genome dataset:</u>
    - There are 453526 rows representing the CpG sites across the whole genome.

3. WorkFlow High-Level Description

3.1. Pipeline

3.1.1. Train-test split (0.8-0.2)

3.1.2. Feature selection

3.1.2.1. <u>Method 1:</u> Select 100 features that have the highest Wasserstein distance.

3.1.2.2. <u>Method 2:</u> Select 100 random features.

3.1.2.3. <u>Method 3:</u> Select statistically significant features with $p - value < 0.05$ under the statistical model, later refined using the FDR method.

3.1.3. Classification Model Training

3.1.3.1. We have trained and tested the results of the following models:

3.1.3.1.1. XGboost

3.1.3.1.2. Logistic Regression

3.1.3.1.3. Random Forest

4. Model

4.1. Statistical model

4.1.1. Null Hypothesis

H_0 - CpG methylation site distribution is the same for healthy and cancerous cells.In our project, we focused on rejecting this hypothesis.

4.1.2. T-Test

4.1.2.1. T-test is a statistical method used to compare the means of two groups to see if they are significantly different from each other.
The formula for the t-value:

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- $\overline{x_1}$ and $\overline{x_2}$ are the average of each of the sample sets.
- s_p is the standard deviation of the differences of the paired data values.
- n_1 and n_2 are the sample sizes of the sample sets.

4.2. Feature selection

4.2.1. To find important CpG sites, we used 2 different methods. The first method we used is choosing the highest ranking features according to Wasserstein distance. We decided to use the first 100 sites. The second method was to find important features using t-test values,

that gave us a measure of how uncommon it is that both our distributions were the same for every feature.

4.2.2. Note: We tried to use M-values [ref 12.1] before performing the t-test but got ~300k features to use and couldn't run our model on all of them. After reducing the size of alpha multiple times we managed to get ~30k features, but again it was too heavy for our models.

5. Algorithms and Mathematical Methods

5.1. Wasserstein Distance

Wasserstein Distance is a measure of the distance between two probability distributions. It is also called Earth Mover's distance, because it is sometimes thought of as the minimum energy cost of moving and transforming a pile of dirt in the shape of one probability distribution to the shape of the other distribution.
Formal definition:

Definition 2.1. *Consider a metric space X endowed with distance function d_X. Then, for two multi-sets $A, B \subset X$, of size $s = |A| = |B|$, the* earthmover distance (EMD) *between A and B is defined as*

$$\mathrm{EMD}_X(A, B) = \frac{1}{s} \min_{\phi: A \to B} \sum_{x \in A} d_X(x, \phi(x))$$

where the minimum is taken over all bijections $\phi : A \to B$. The resulting metric is called EMD over X.

5.2. Random Forest

Random Forest is an ensemble learning algorithm that builds multiple decision trees during training. Each tree in the forest is built independently, using a random subset of the features and a random subset of the training data. During prediction, each tree in the forest independently predicts the output, and the final prediction is determined by a majority vote (classification) or averaging (regression) of the predictions from all the trees. Random Forest is known for its robustness, scalability, and ability to handle high-dimensional data with minimal tuning.

5.3. XGBoost

XGBoost stands for eXtreme Gradient Boosting. It's a powerful machine learning library that enhances the performance of gradient-boosted decision trees. It's known for its efficiency, speed, and ability to handle large datasets. The "XG" emphasizes the extreme performance improvements and capabilities it brings to gradient boosting, including features like handling missing data, regularization to prevent overfitting, and options for parallel and distributed computing.

5.4. Logistic regression

The logistic function, also known as the sigmoid function, is used to ensure that the probabilities are bounded between 0 and 1. This method provides a framework for modeling the relationship between a binary dependent variable and continuous, categorical, or mixed independent variables, facilitating the understanding of how predictor variables influence the likelihood of various outcomes.

5.5. False Discovery Rate

FDR is a statistical method used to estimate the proportion of false positives among all positive findings or tests declared significant. This concept is especially important in fields where multiple hypothesis testing is common.

The FDR is the expected proportion of false discoveries (FP) among the total number of discoveries (both TP and FP):

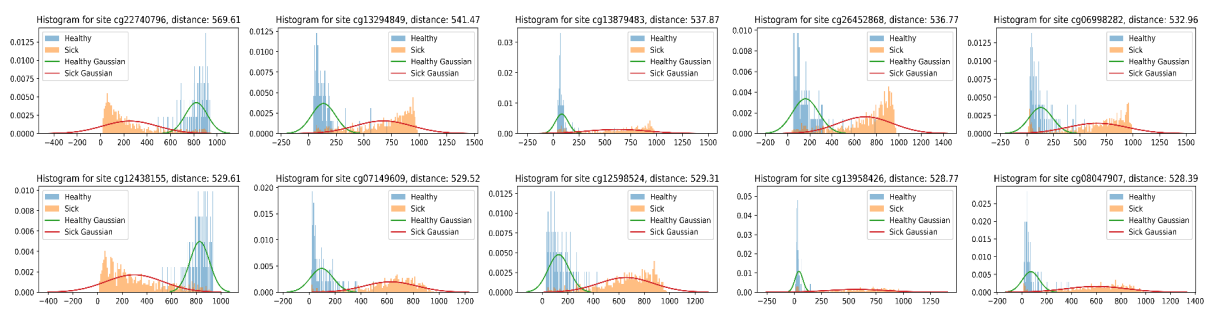
$$FDR = \frac{Expected\ number\ of\ false\ positives}{Total\ number\ of\ discoveries} = \frac{E[FP]}{TP+FP}$$

6. Feature Analysis

6.1. Features importance according to the Wasserstein distance

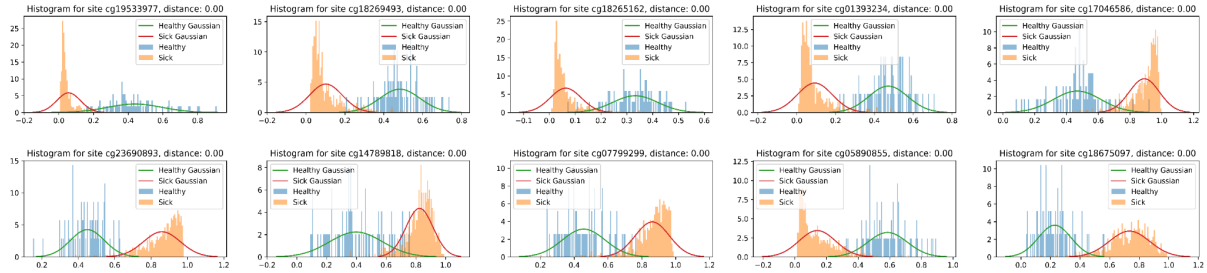
These are the distributions for the most/least indicative CpG sites, we can see that indeed they differ in their mean and the variance. This can explain why training on those features gives better results (as seen in section 7).

6.1.1. The 10 most important features according to Wasserstein distance:

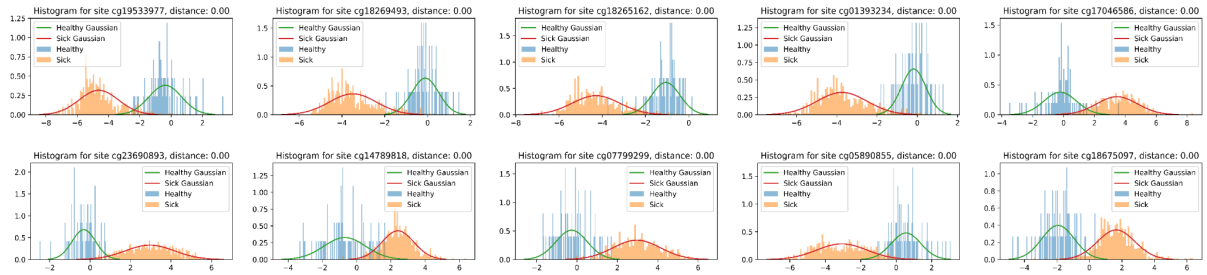


6.2. Features importance according to the p-value

6.2.1. The 10 most important features according to p-value:



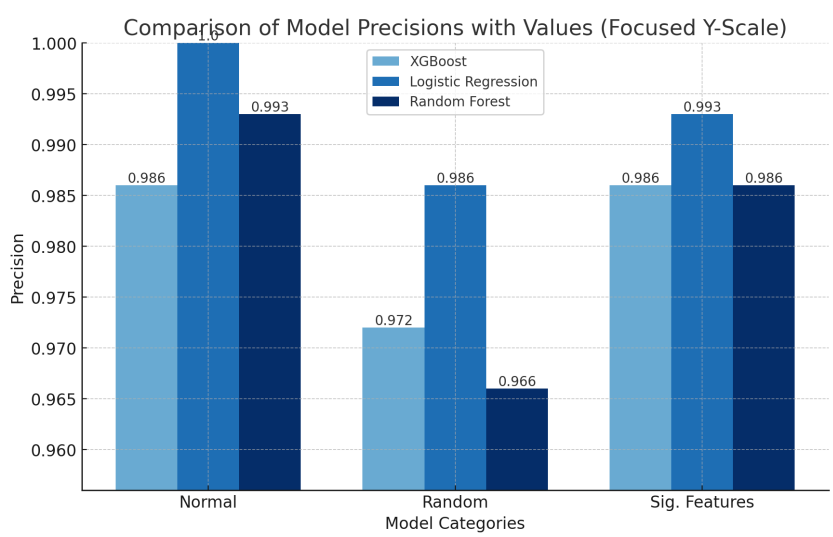
6.2.2. The 10 most important features according to p-value when using M-values transformation on the data:



7. Results

7.1. Precision

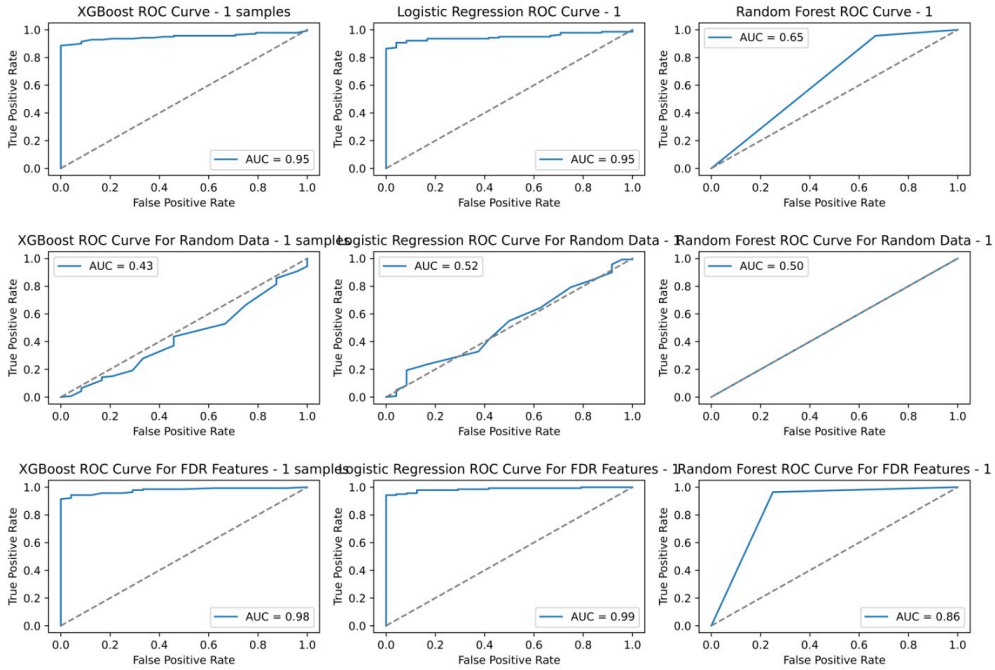
7.1.1. We present our results for each of our models, on the 3 datasets we created using our method for feature selection (as described in section 3.2).



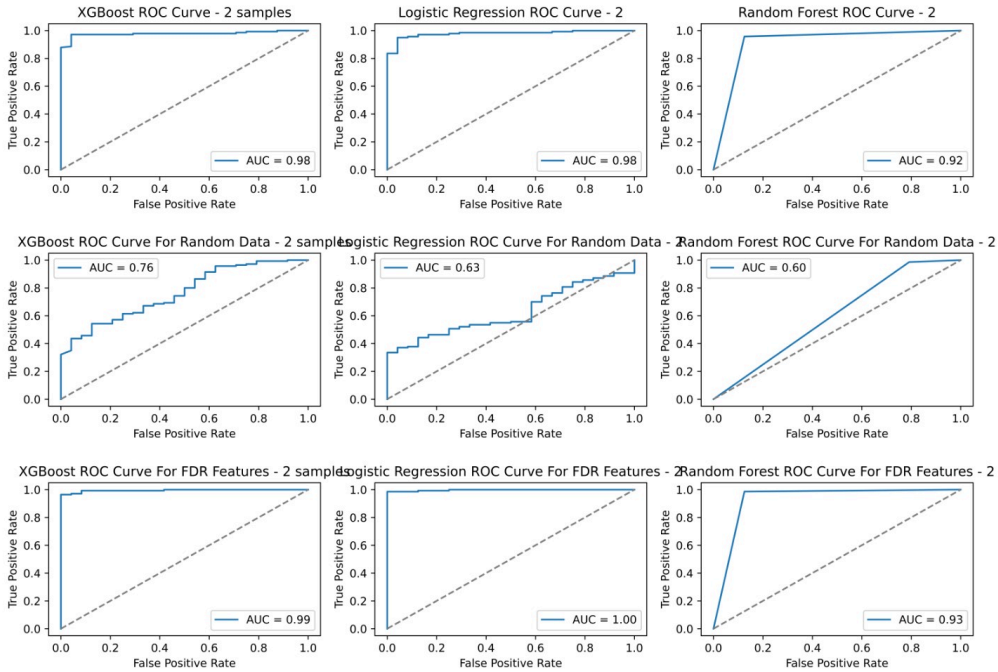
7.2. ROC

7.2.1. We compare ROC results for different amounts of features used to train our models. Every section describes taking the best X features according to each of the methods.

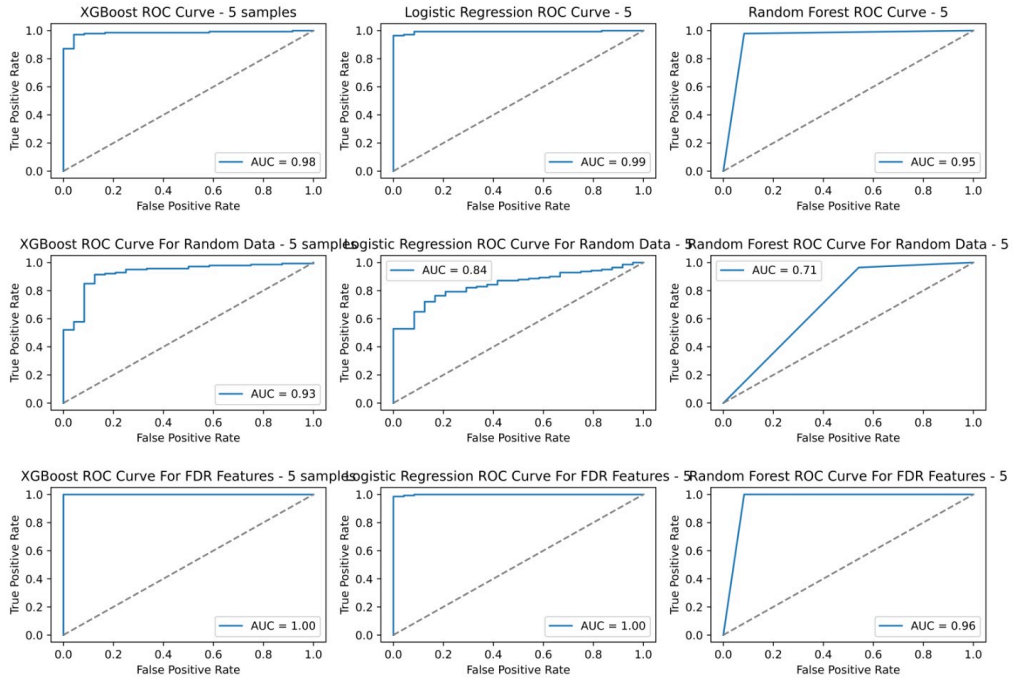
7.2.2. Using the best feature (X=1):



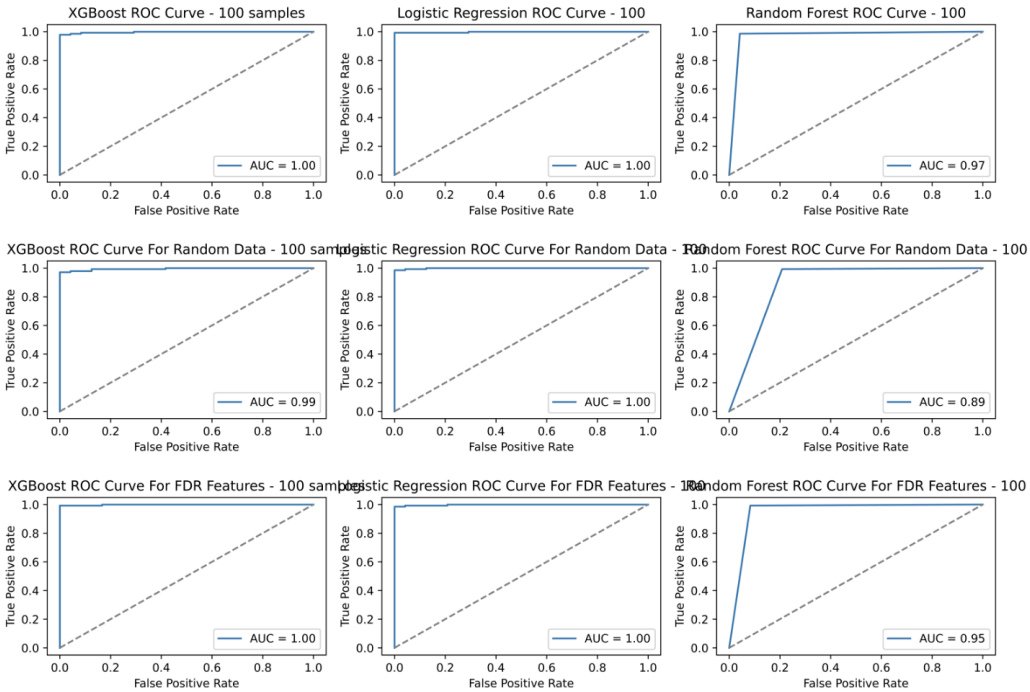
7.2.3. Using 2 best features (X=2):



7.2.4. Using 5 best features (X=5):



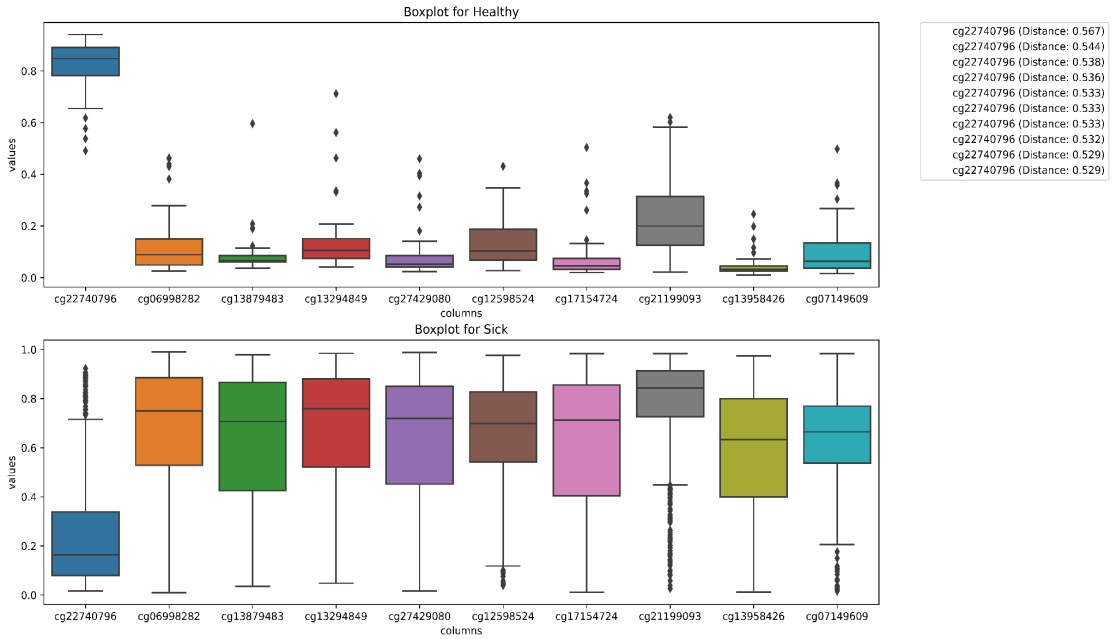
7.2.5. Using 100 best features (X=100):



7.3. Analysis of Methylation Rate in Relation to Distribution Distance

7.3.1. To identify the most informative CpG sites for our classification model, we used the Wasserstein Distance as a selection criterion.

Our objective was to explore the relationship between methylation rates and this distance metric, aiming to ascertain whether the selected CpG sites show patterns of hypo- or hypermethylation.



Upon analyzing the top 10 CpG sites with the most significant differences as indicated by the Wasserstein Distance we observed:

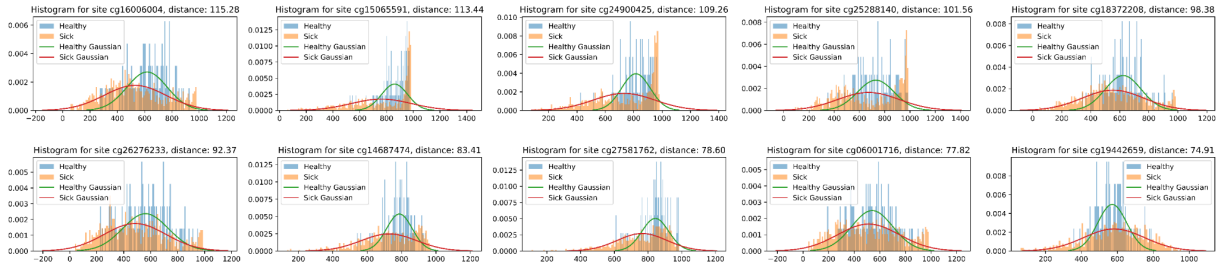
1. A greater variability in methylation rates among cancerous cells compared to healthy cells.
2. Healthy cells generally exhibited lower methylation rates, whereas cancerous cells demonstrated a tendency for higher methylation rates.

8. Relation to Known BRCA Genes

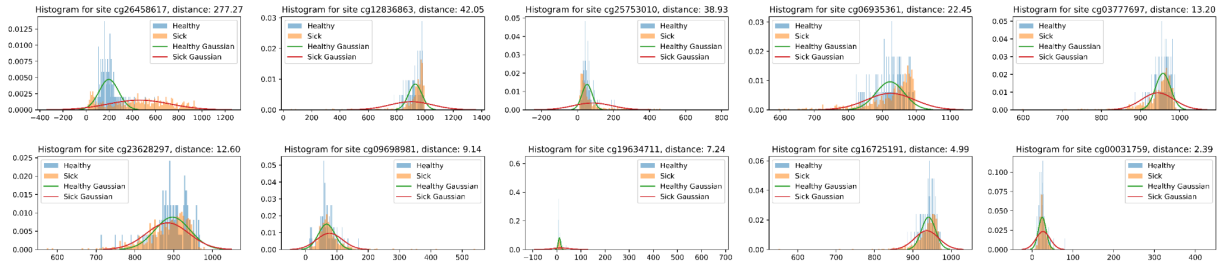
8.1. BRCA1 and BRCA2

8.1.1. BRCA1 and BRCA2 are unrelated proteins, but both are normally 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. We wanted to check the methylation levels of various CpG sites across the BRCA1 and BRCA2 checking if methylation levels could be a sufficient indicator for BRCA mutation.

8.2. The 10 most important features according to Wasserstein distance in BRCA1:



8.3. The 10 most important features according to Wasserstein distance in BRCA2:



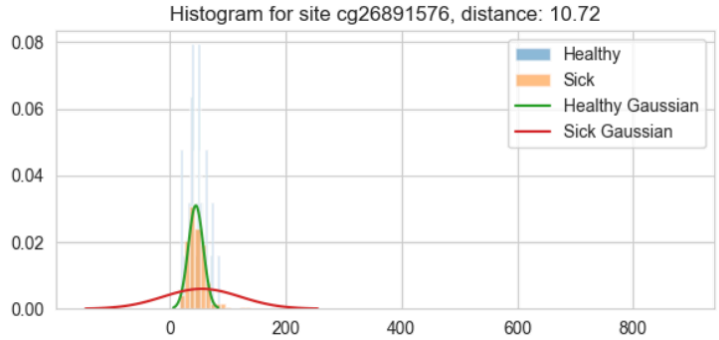
8.4. Findings

8.4.1. When analyzing distributions calculated over Wasserstein Distance, the data suggests that the sites exhibit a lower distance score compared to other genomic sites. When investigating BRCA1/2 sites in relation to cancer, there are two potential primary causes: either the proteins are mutations contributing to the cancer or the site experiences hypermethylation, leading to its muting. Consequently, some patients may develop cancer due to a mutation in BRCA1/2, resulting in their methylation at the CpG site associated with BRCA1/2 falling within the healthy range.

8.4.2. Exploring TSS200 BRCA1

8.4.2.1. TSS200 indicates the start of the BRCA1 promoter, we know that a high level of methylations could cause the promoter to be less active and thus cause cancer.

8.4.2.2. We attempted to analyze the CpG site: cg26891576, this site is in the TSS200 range.



8.4.2.3. Analyzing the result indicates the same conclusion. The cancer could be caused by a mutation in BRCA1/2 and not hypermethylation at the promoter.

9. Conclusions

9.1. All models achieved high accuracy Using 100 features, with XGBoost and Logistic Regression showing slightly better performance, suggesting they might be a better option to use for clinical applications and diagnosis of breast cancer.

9.2. 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 our hypothesis.

9.3. As seen in section 7.2, some CpG sites are highly informative for classification and by using even 2 "strong" features we can achieve very good results (almost identical to using 100 features).

9.4. Moving forward, these conclusions underscore the potential of DNA methylation as a non-invasive biomarker for breast cancer screening and diagnosis.

9.5. 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.

10. Previous Papers

10.1. DNA methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation [ref 11.6]

10.1.1. The paper describes a complex network of DNA methylation modifications within breast cancer tissues. Their work of profiling 1,538 breast tumors, uncovered the replication-linked clock and epigenomic instability, highlighting its correlation with tumor

aggressiveness, TP53 mutations, and prognostic outcomes. Our study was a more targeted analysis of methylation's diagnostic potential. By integrating machine learning algorithms–namely, Random Forest, Logistic Regression, and XGBoost–we not only achieved a predictive accuracy nearing 0.98 but also underscored the potential of these selected CpG sites as robust biomarkers for breast cancer diagnosis. Our results align and might also extend the insights provided by Batra et al. by offering a practical application of computational models to enhance the specificity and reliability of disease diagnostics through epigenetic markers.

10.2. A Nobel Promoter CpG-Based Signature for Long-Term Survival Prediction of Breast Cancer Patients [ref 11.7]:

10.2.1. This study developed a novel promoter CpG-based signature to predict long-term survival in breast cancer patients, using data from TCGA and validated in GEO datasets. In our algorithm we took the top 100 CpG sites and we checked what RefGene repeated and found that TSS and UTR are repeated a few times. TSS (Transcription Start Site): The point in the DNA where the transcription of a gene into RNA begins. It's crucial for controlling when and how much a gene is expressed. UTR (Untranslated Region): The part of mRNA located just after the coding sequence and before the tail end. It doesn't code for protein but plays a key role in regulating gene expression, including mRNA stability and translation efficiency. The concurrence between our data analysis and the existing research underscores a potentially vital correlation: alterations in the methylation patterns within TSS and UTR regions may significantly impact gene expression pathways involved in breast cancer development. This alignment highlights the importance of these regions in both the onset and progression of the disease, as well as their potential as targets for diagnostic and therapeutic strategies

11. Forward Perspectives in Breast Cancer Research

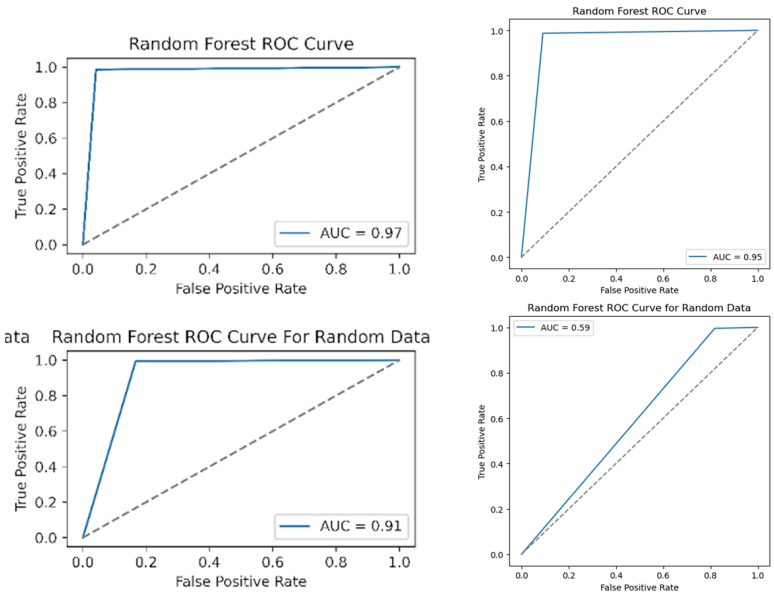
11.1. Despite breast cancer's prominence, the number of healthy individuals significantly surpasses those diagnosed with the disease. In the dataset, there's a notable imbalance between the number of sick and healthy samples, which raises concerns regarding representativeness.

11.2. Consideration of weighting the samples in the models could help address the imbalance between sick and healthy samples, ensuring more accurate and representative analyses.

11.3. After utilizing a subset of CpG sites with strong features, our results persist in their reliability, hinting at potential overfitting in the model. Further validation and exploration are crucial to ensure the broad applicability of our findings and to mitigate the risk of overfitting.

11.4. Exploring other cancer types could provide valuable insights into DNA methylation patterns in breast cancer patients.
The distinction between breast cancer patients and those with colon adenocarcinoma is emphasized here cancer concerning chromosome 19. The findings reveal that the random forest model trained on colon adenocarcinoma data with random feature selection performs worse compared to using the best feature selection method. This underscores the significance of feature selection in research, as it enhances the model's ability to analyze data and achieve higher precision.

BRCA COAD



12. References

12.1. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis | BMC Bioinformatics | Full Text

12.2. Wasserstein Distance, Contraction Mapping, and Modern RL Theory | by Kowshik chilamkurthy

12.3. XGBoost – What Is It and Why Does It Matter?

12.4. Earth Mover Distance over High-Dimensional Spaces

12.5. The Clinical and Pathological Profile of BRCA1 Gene Methylated Breast Cancer Women: A Meta-Analysis - PMC

12.6. DNA methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation | Nature Communications
12.7. A Nobel Promoter CpG-Based Signature for Long-Term Survival Prediction of Breast Cancer Patients
12.8. The site with the datasets we used