Epi-C

Target:

DNA methylation, is an important epigenetic modification, is associated with gene silencing, and the primary methylated sequence in vertebrates is CpG.

We predicted the prognostic properties (Clinical Stage) of cancer using CpG methylation data. We analyze the evaluation metrics of the stage prediction pipeline made using Machine Learning. In this study, we tried to identify DNA methylation markers to differentiate early-stage samples from the late-stage samples in cancers. We further probe the correlation of these biomarkers with overall survival (defined as the time from the first day of cancer diagnosis until the day of death by any cause) with DNA methylation levels.

Created Machine Learning pipeline to predict

Ovarian Cancer:

Ovarian cancer (OC) causes significant morbidity and mortality as neither detection nor screening of OC is currently feasible at an early stage. Difficulty to promptly diagnose OC in its early stage remains challenging due to non-specific symptoms in the early stage of the disease, their presentation at an advanced stage and poor survival. Therefore, improved detection methods are urgently needed. In this study, we summarize the potential clinical utility of epigenetic signatures like DNA methylation which play an important role in ovarian carcinogenesis and discuss its application in the development of diagnostic, prognostic, and predictive biomarkers.

Data

The Cancer Genome Atlas methylation, CPG Makers: Illumina HumanMethylation27

Ovarian: Samples: 616

Kidney Samples: 414

For binary Classification between early-stage and late-stage cancer, the data was given Labels:

0: early-stage (Stage 1,2, 3a)

1: late-stage (Stages 3b, 3c, 4)

Handling Missing Values:

Samples with missing labels were removed

Missing values in the methylation data were imputed with mean

Training and Testing:

4 Fold validation was used to split the data every time. SMOTE was used with training data to upsample the minority class. The data was trained and tested on 6 different models: SVM, Random Forest, Logistic Regression, K nearest, Balanced Bagging, RUS

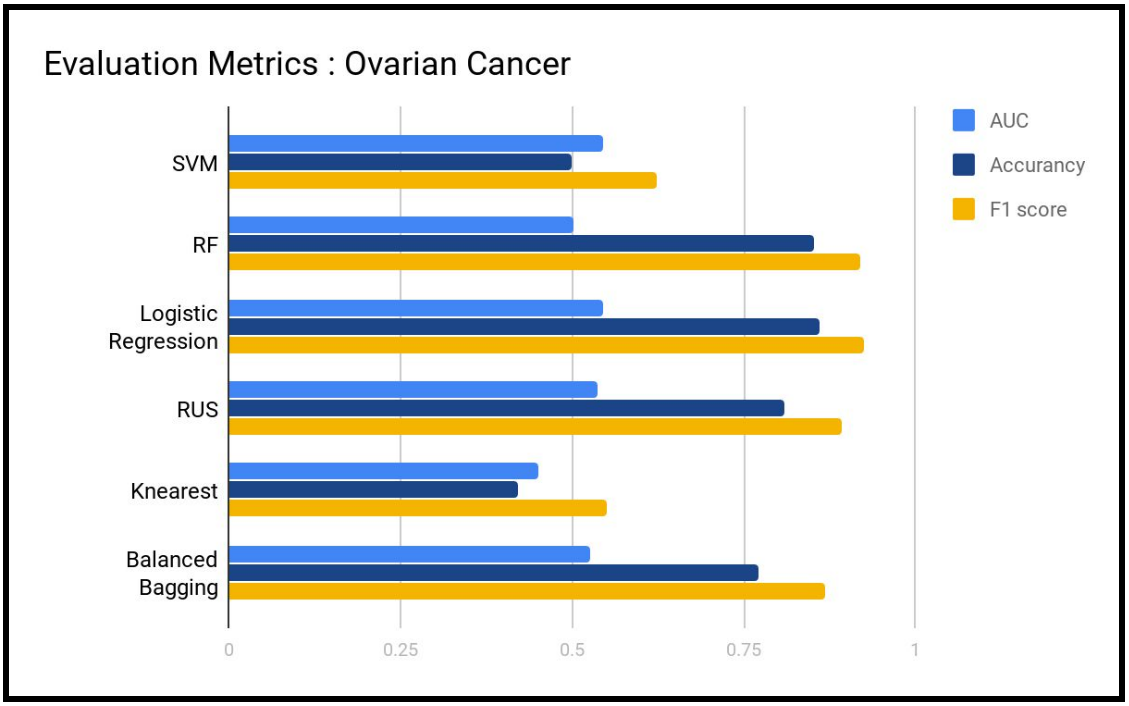
Results:

Ovarian Cancer:

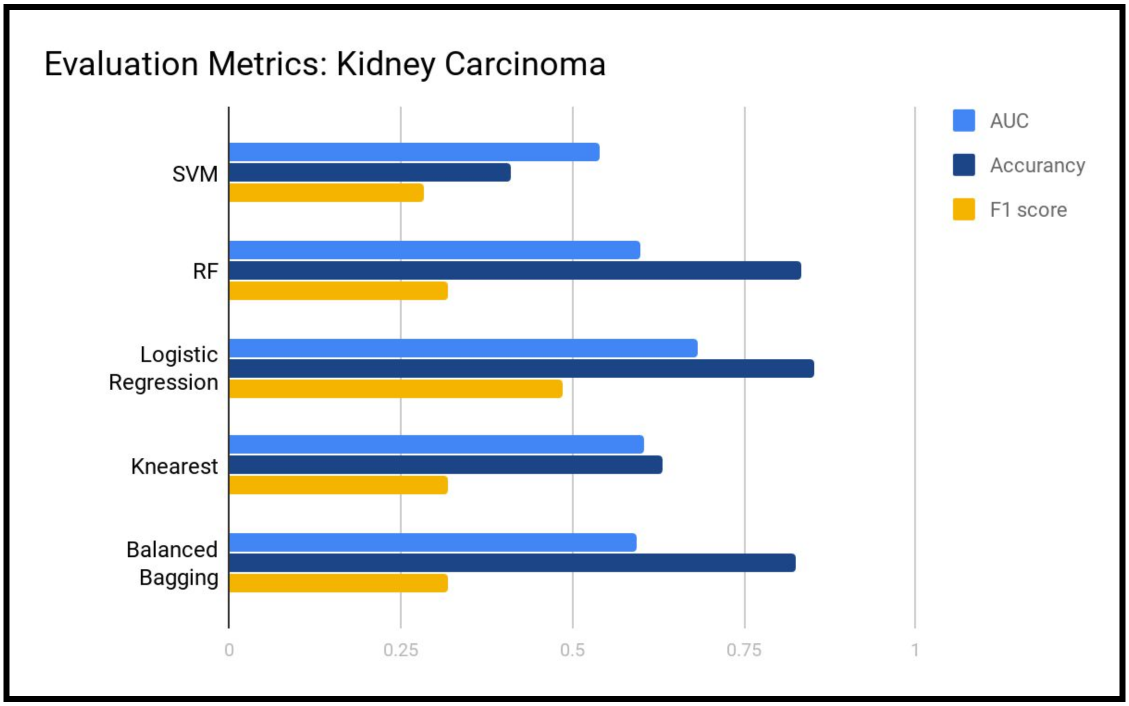
Selected Model: **Logistic Regression with F1 score 0.923**

Kidney Carcinoma:

Selected Model: **Logistic Regression with F1 score 0.48**



Kidney Carcinoma:



Survival Analysis

One of the problems posed by a computational pipeline to identify biomarkers for a disease is its verifiability. The CpG islands showing high statistical significance may not be biologically significant. Hence we performed survival analysis with consideration to top three CpG sites with a different dataset, according to our preprocessing tests and plotted the Kaplan Meier Curve in R. “survival” and “survminer” packages were used to generate the plots.

Dependent Variables

We considered the first three CpG islands (the most significant) and their beta values for each patient. For each patient, the island was either assigned a value of 1 if its degree of methylation was greater than 0.8, otherwise 0. Here, 1 corresponds to that the region was methylated, and 0 corresponds to that the region was not methylated.

The reason we considered this value as threshold is that this was the common value which showed a sharp peak in the bar graph for the beta value distribution. Further, small differences in beta values for threshold can be neglected as these can be overlooked when looking from biological perspective. Hence, we made sure to choose a high common value for threshold.

Event and Time

The event corresponds to the status of the patient. If the status of the patient was 0, it meant that the patient had died due to cancer. For these patients the number of days to death were given. For patients with the status 1, they were labelled - “censored” - the status was unknown. For such patients, the time of death was assigned the number of days corresponding to the last follow-up day. Here, the variable time refers to the number of days to death.

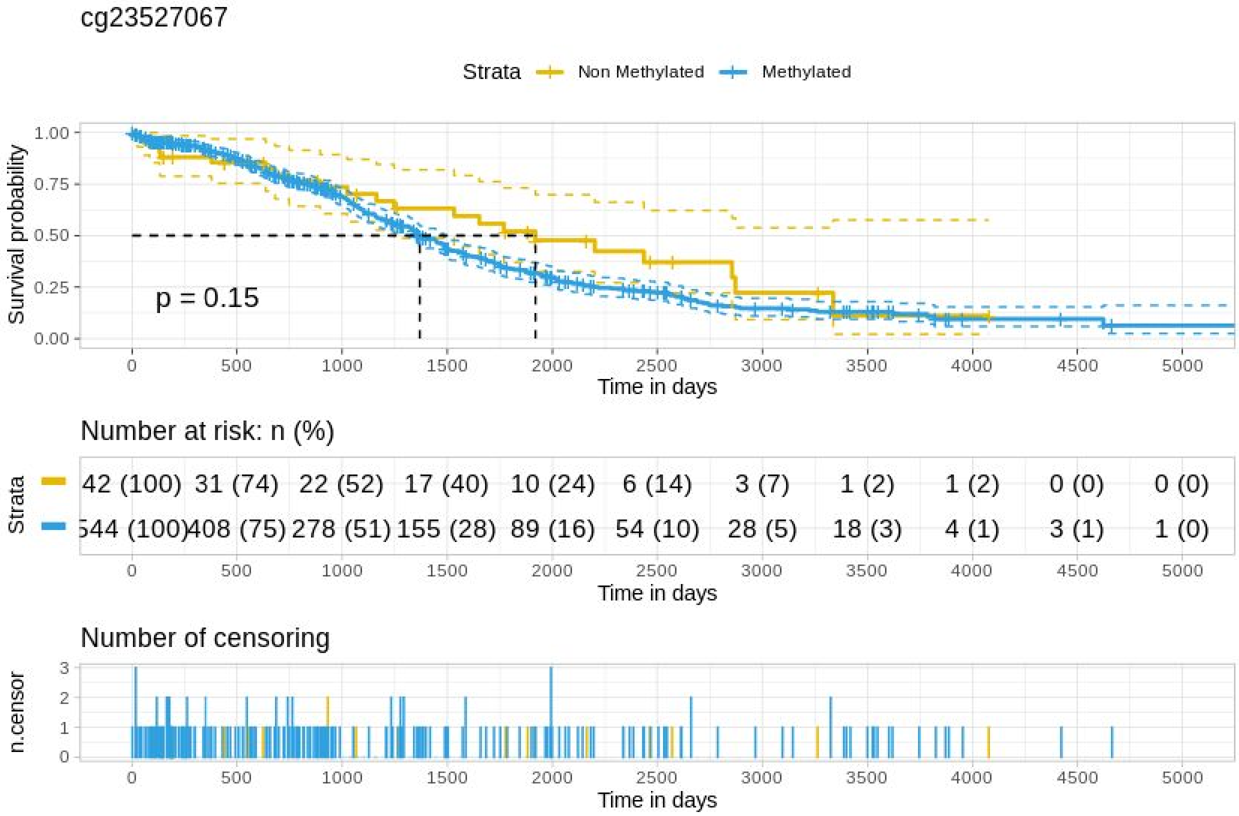
Method

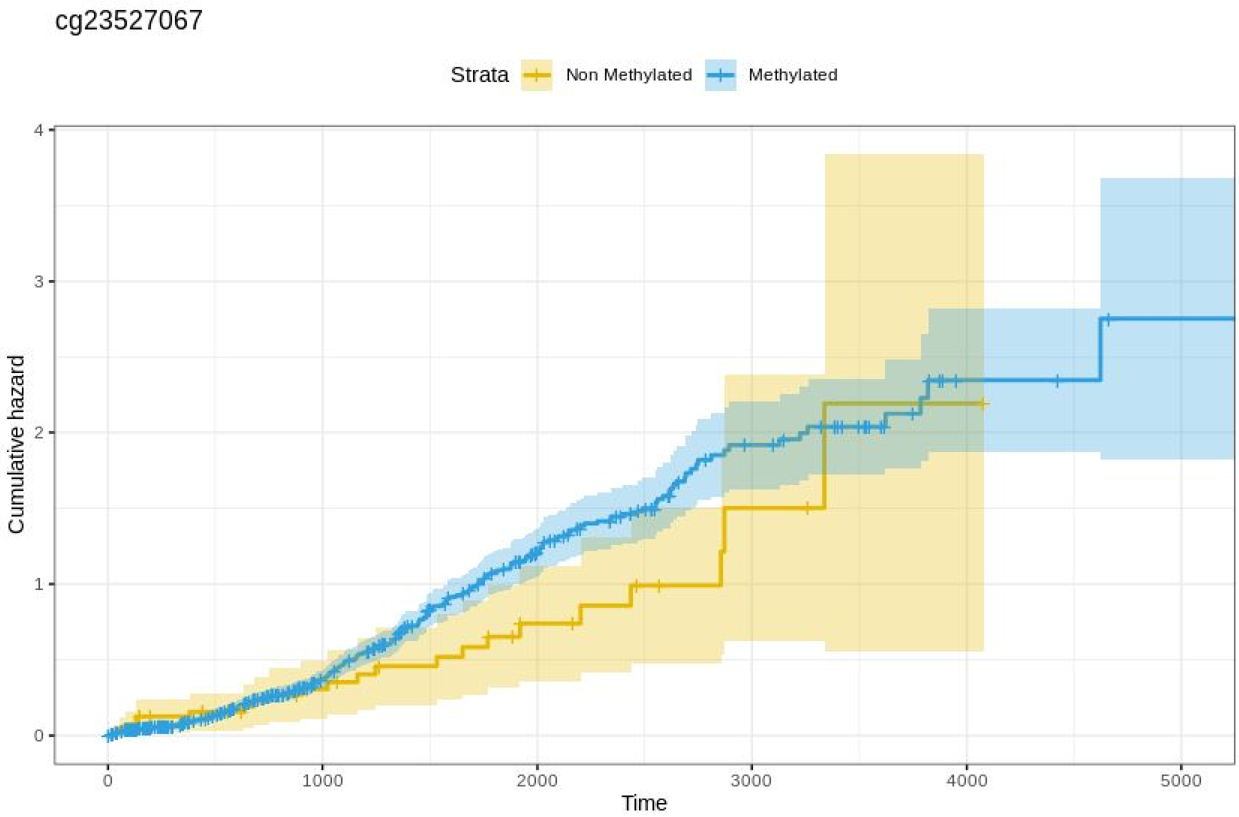
[Due to resource constraints (internet limitations and server errors), we were able to do survival analysis for only Ovarian Cancer. We used data from GDC-TCGA for our analysis. We used only data extracted by Illumina Human Methylation 27]

The first three CpG islands were identified. These were - cg23527067, cg16977035, cg23412777.

Initially, the plots for the first CpG island were done.

Survival Probability Time Curve



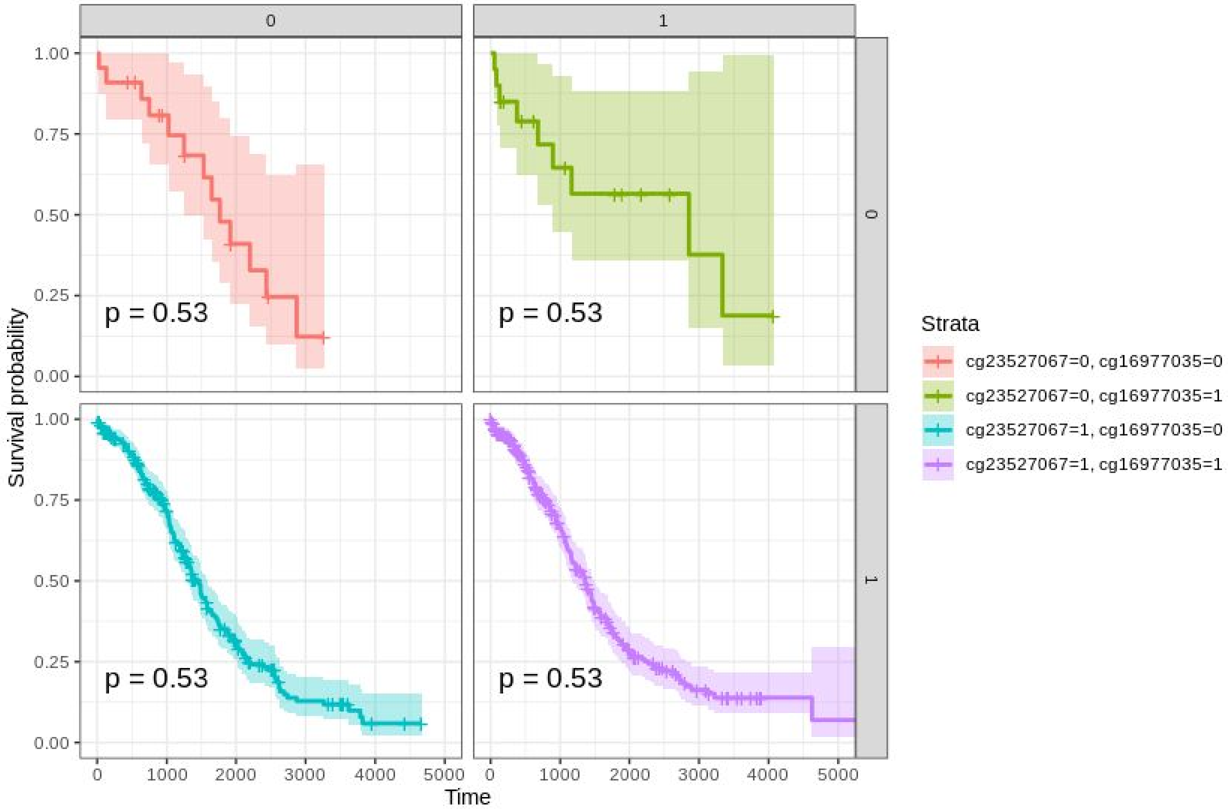


Cumulative Hazard Time Curve

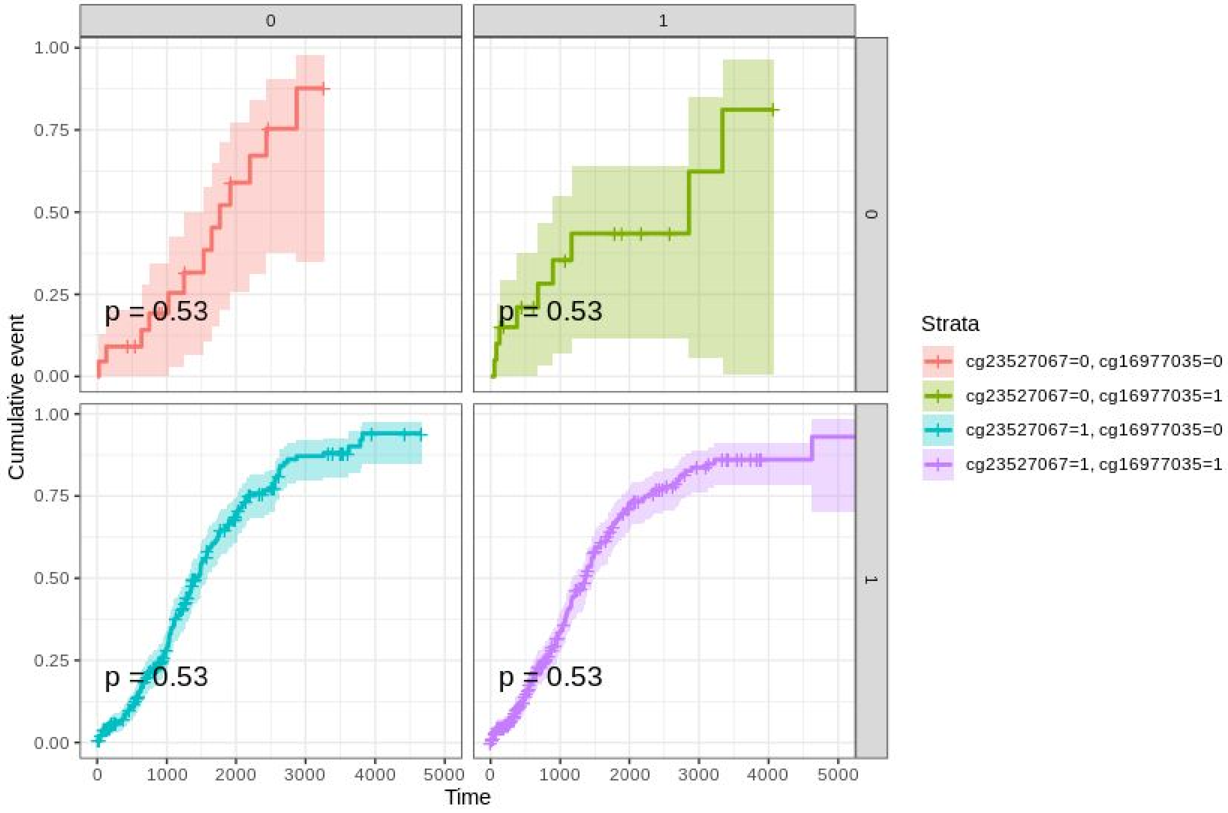
In this survival probability - time curve, we can observe that the median survival probability is higher for the methylated than the non-methylated. As the survival probability decreases, we can see that the distinction between the curves becomes more and more decreased.

Log rank test was done in order to check if the two groups are producing any significant statistical difference. The p value comes out to be 0.2, proposing that there might not be any significant statistical difference between the groups.

We then performed the analysis by testing the dependence on both the first and the second CpG markers, and then eventually adding the third one as well.



Survival Probability Time curve for top 2 CpG islands



Cumulative Time curve for top 2 CpG islands

Conclusion :

- Initial analysis shows that methylation beta values are highly correlated to cancer stages.
- We found Ovarian Cancer to be best predicted by beta values (**92% F1 score**). But the other cancers were not predicted very badly either.
- Further analysis showed that our results could be the result of bad data sampling and one or many other factors listed below.
- Survival Analysis showed good correlation but p-values suggested otherwise. P-values showed that our results could just be coincidental.

Limitations:

- The data that we received was not very well balanced. There was a lot of bias. For eg: Ovarian Cancer had 4x late-stage cancer patients than early stage. This could have made our model skewed. This might explain why we have very high accuracy (more than initially expected). We have tried to use SMOTE to overcome this.
- The survival analysis data has a large p-value, meaning our results could have been random. This might be due to a lot of reasons. As mentioned above skewed data might be the reason. TCGA data, the one which we have used, is known to have problems with survival analysis. The other most important reason might be that methylation data just might not be that good at predicting survival chances. (Although this opposes our high accuracy on models).
- Some ways to improve our survival analysis would be to use more biomarkers, since biomarkers together may control cancer phenotypes. Also, it is to be noted that most survival analyses use not just beta values, but also other criteria to judge like age, sex, ethnicity, etc.
- The biomarkers we found were found using very basic statistical tests. These might have given us a wrong list of markers. This is one thing which can be easily improved by using various other techniques to find biomarkers like Univariate Selection and Feature Importance (with methods like ExtraTreeClassifiers).
- Our ROC was consistently low, even though our accuracy and F1 scores were very good. Better validation and testing might help overcome this, though we tried many other things like Random Sampling.
- We have used only Early and Late Stage for distinction. We could have tried to get results for all the stages separately. [We actually tried this and got a significantly lower accuracy. That has to also be blamed on the fact that the data was very skewed(Stage 1 for Ovarian cancer had 17 samples, whereas stage 3

had 300+ samples)]. We got an accuracy of 72% for the best case and an average of 63%

Links:

Google Drive Link - https://drive.google.com/drive/folders/106LcWGiVI7ZIGyRvzTQk_drndGLBmFjZ?usp=sharing