Prediction of age based on DNA methylation levels at age-related CpG sites

BO-HAN, YAO
H24111057
h24111057@gs.ncku.edu.tw

Code for this project:https://github.com/curryhank08/analyze-data-from-GEO-in-python

Abstract

Epigenetic clock is a remarkable achivement in the research field of aging [2]. It not only shows, as biomarkers of aging, age-related CpG cites are more reliable than tolemere, but also introduce a new way to look into the mechanism of aging. In our study, we aims to investigate if only few CpG sites as features still maintain comparable performance to the Epigenetic clock from previous research. In short, we use gradient boosting random forest with 21 CpG sites (reduced-CpGs is after DBSCAN clustering) can have better performance than Horvath's epigenetic clock (Mulitiple regression with 353 CpG sites).

1. Introduction

Aging is associated with numerous diseases, including cardiovascular disease and Alzheimer's Disease, among others [7]. The impetus for this research stems from my previous investigation into the identification of age-related CpG sites during a project study.

Prior to model constructing, it is imperative to identify age-related CpG sites from among the approximately 480,000 CpG sites within the microarray platform known as Illumina HumanMethylation450 BeadChip. One of the initial challenges to address is acquiring a foundation in the domain knowledge of DNA methylation. A comprehensive understanding of the subject matter we intend to analyze is instrumental in guiding the feature engineering process.

2. System framework

Figure 1 illustrates the overarching framework of this research project. In the initial stage, we employ a multiple linear regression model for each CpG site. This enables us to effectively filter age-related CpG sites by establishing specific criteria, such as setting the p-value/adjusted p-value below a predetermined threshold or the R-squared/Adjusted R-squared above a specified threshold.

Moving forward, our subsequent step involves the constructing of a multiple linear regression model for predictive purposes. We utilize the β.values of previously identified age-related CpG sites as features, with the corresponding age of each sample serving as the target variable [2, 5, 7]. It is worth noting that we may explore the application of alternative approach, such as neural networks [4] or transformer, at a later stage.

Additionally, an integral aspect of this research involves the clustering of age-related CpG sites. This stems from the diverse patterns observed in the changing trends of β.values across these age-related CpG sites. To achieve this, we employ density-based clustering algorithms to effectively categorize these CpG sites.

In the final step of our research, we rigorously validate the performance of the constructed model using additional datasets sourced from the same microarray platform, Illumina HumanMethylation450 BeadChip. These datasets are readily accessible through the NCBI Gene Expression Omnibus.

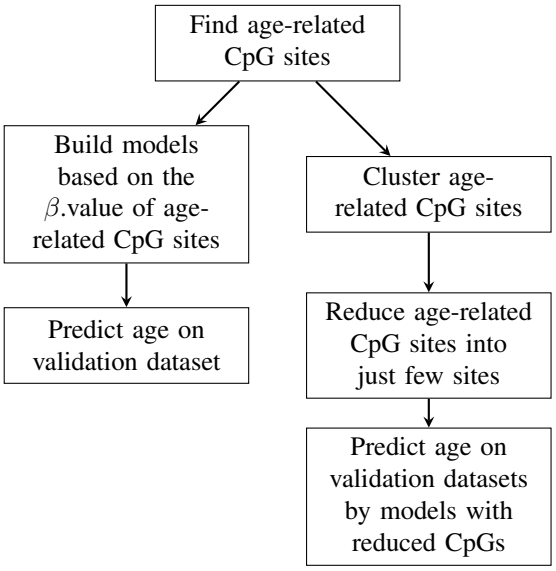


Figure 1. Flowchart of System Framework

3. Expected results

The datasets used for feature extraction and model constructing originate exclusively from the Illumina HumanMethylation450 BeadChip. This microarray chip comprises over 480,000 probes, each corresponding to a specific CpG site. The datasets, obtained from the NCBI Gene Expression Omnibus, have undergone pre-processing, resulting in the transformation of methylation levels into β.values. The β.value is calculated using the following equation:

$$\beta.value = \frac{M}{M + U + a} \tag{1}$$

where $M > 0$ and $U > 0$ denote the methylated and unmethylated signal intensities, respectively, measured by the Illumina HumanMethylation450 BeadChip microarray platform. The offset $a \geq 0$ is usually set equal to 100 and is added to $M + U$ to stabilize β.values when both M and U are small. [6]

Notably, former researchs revealed that the precision of age prediction on additional testing datasets significantly improves when considering a larger number of age-related CpG sites and a more extensive training sample size [2, 7]. Furthermore, the predicted age appears to closely align with biological age [5, 7].

By implementing clustering algorithms, we aim to classify these age-related CpG sites into distinct clusters based on their diverse trends in β.value changes.

Consequently, we plan to reconstruct a model using the reduced set of CpG sites to predict age on additional datasets, with the hope that the reduced model will maintain a level of performance comparable to that of the original model.

4. Dataset preparation

We combine the series datasets by a same microarray platform, Illumina HumanMethylation450 BeadChip, from NCBI Gene Expression Omnibus through flollowing accession id: GSE40279, GSE30870, GSE104812, GSE137503, GSE36064, GSE42861, GSE50660, GSE56581, GSE56105, GSE61496, GSE55763, GSE56064. Steps are followed as

1. Transform M-value to β.value if expression value is in M-value form.

2. Remove samples with age in NA.

3. Impute missing expression value by KNN.

The distribution of samples' chronological age from combined dataset is shown as Figure 2.

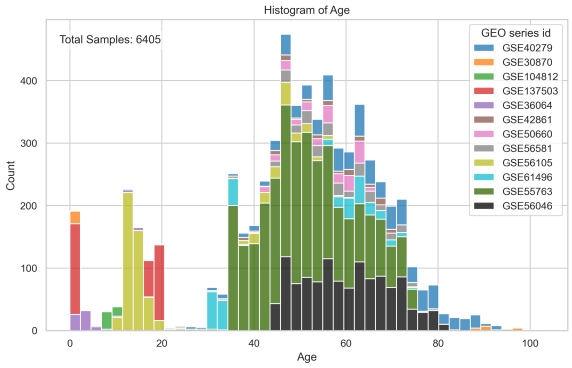


Figure 2. Histogram of samples' age

5. Feature Pre-selection

Before model traning, we have to find out these age-related CpG sites from approximately 480,000 CpG sites the microarray platform illumina HumanMethylation450 BeadChip sensored. In my study, I implement two different approaches, Regression and Random Forest, to select CpG sites. [2]

The difference between Regression and Random Forest is that CpG cites are filtered by correlation in Regression approach whereas by gini impurity in Random Forest approach.

5.1. Regression approach

Regression model is easy to adjust the effects of some counfounding factors since we can add these counfounding factors as independent variables to model. We will construct two-stage regression model to filter CpG cites as following steps:

1. Regression model for filtering age-related CpG cites.

2. Regression model for filtering other CpG cites relating to the age-related CpG cites (stage1-CpGs).

In stage 1, we build a regression model for each CpG sites, then filter the CpG cites with correlation > 0.7 as features. Model for each CpG cites is shown as formula 2. Y_i is ith sample's β.value of a specific CpG cite, while $X_{i,age}$ are ith sample's chronological age.

$$Y_i = \beta_0 + \beta_1 X_{i,age} + \varepsilon_i \tag{2}$$

By appling ordinary least squares method, we can minimize the sum of squared residuals $S = \Sigma_{i=1}^{n} \varepsilon_i^2$ to find the best estimations of the parameters for following fitted model 3.

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i,age} \tag{3}$$

In stage 2, we build a regression model 4 for each age-related CpG cites filtered by stage 1 on each CpG cites sensored by lllumina HumanMethylation450 BeadChip except the age-related CpG cites. Y_j is the β.value of the j-th age-related CpG cite while X_i is the β.value of the i-th CpG cite except age-related CpG cites.

$$Y_j = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{4}$$

The procedure of stage 2 will spend lots of time since there is approximate $N \times (480,000 - N)$ models have to constuct. N is the total amount of the age-related CpG cites.

After the two-stage regression modeling, we will set the age-related CpG cites filtered in stage 1 and the CpG cites filtered in stage 2 as the features for the predictive purpose models construced in section 7. Note: Stage1-CpG sites are denoted as filtered in 10a

5.2. Random Forest approach

The Random Forest approach provides an alternative methodology for feature selection. This ensemble learning technique builds multiple decision trees and aggregates their outputs. The importance scores of CpG sites are then used to identify relevant features. The Random Forest method is particularly beneficial when dealing with complex interactions within the data.

6. Clustering of age-related CpG cites

Based on DBSCAN algorithm, we're trying to cluster stage1-CpG sites due to their different patterns of β.value change trend during aging. The stage1-CpG sites are clustered into 5 cluters with total amount 26. Then we randomly choose only a CpG site from each cluster 0, 1, 2, 3, 4 and all CpG sites in cluster -1 as reduced CpG sites.

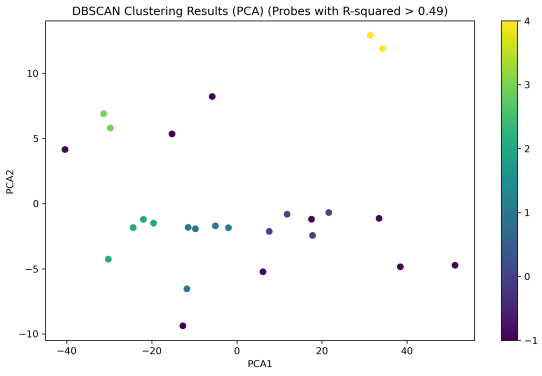


Figure 3. Visualization of DBSCAN result

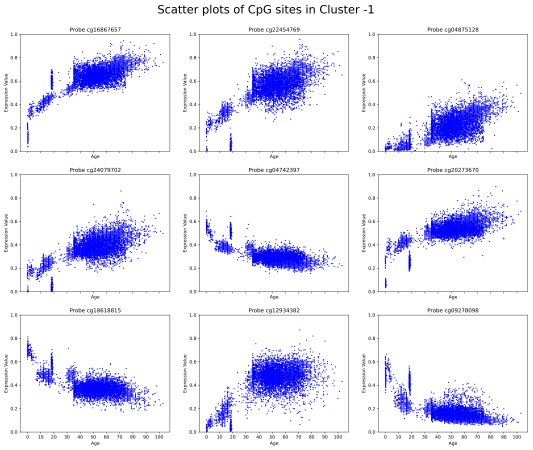


Figure 4. Scatter plos for the CpG sites not belongs to any cluster which all put into cluster -1

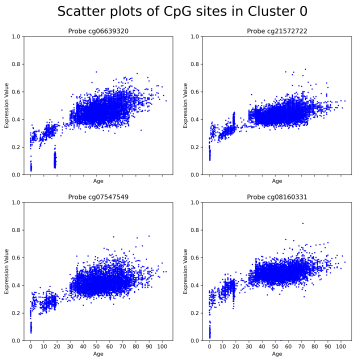


Figure 5. Scatter plos for the CpG sites belongs to cluster 0

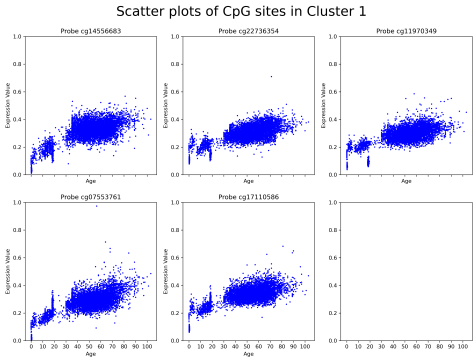
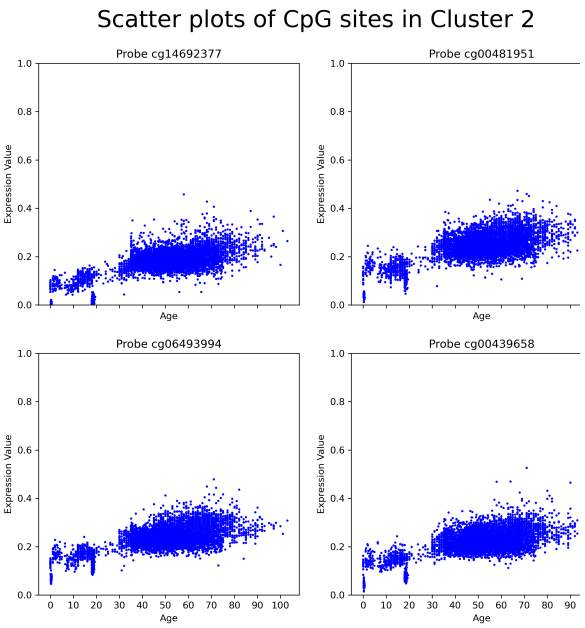
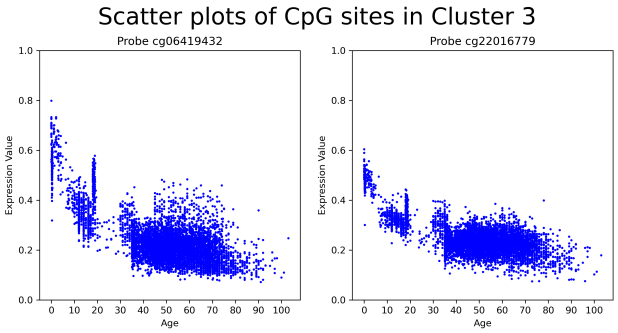


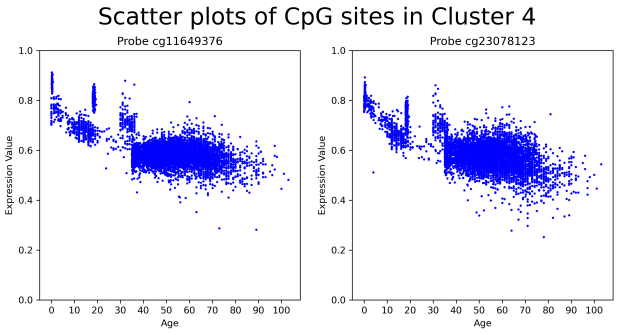
Figure 6. Scatter plos for the CpG sites belongs to cluster 1



(a) Scatter plots for the CpG sites belonging to cluster 2



(b) Scatter plots for the CpG sites belonging to cluster 3



(c) Scatter plots for the CpG sites belonging to cluster 4

Figure 7. Scatter plots for cluster 2, 3, 4

7. Modeling

7.1. Regression model

The formula of multiple regression model for prediction is defined as

$$Y_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_i X_i + \varepsilon_j \qquad (5)$$

Where Y_j is the age of jth sample, X_i is β.value of ith age-related CpG cite and ε_j is the residual of jth sample.

After minimizing the sum of squared residuals $S = \Sigma_{j=1}^{n} \varepsilon_j^2$, then the fitted model with the best estimations of the parameters is defined as

$$\hat{Y}_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i \qquad (6)$$

Where $\hat{Y}_j$ is the predicted age of jth sample, X_i is β.value of ith age-related CpG cite.

7.2. Random Forest model

We build a Random Forest model based on ensembling 100 decision trees with boostrap aggregating. Also, setting the minimum number of samples required to split an internal node as 4 and the minimum number of samples required to be at a leaf node as 3.

7.3. MLP model

The model structure is shown as Figure 8, and the loss fuction is Mean Squared Error (MSE) for the training process, then we train the model by 150 epoch and full batch size.

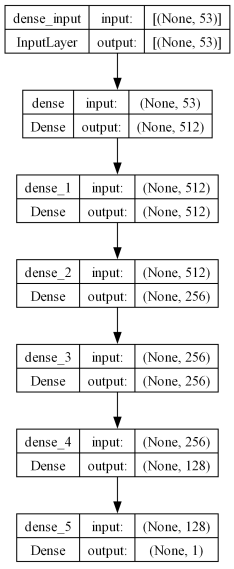


Figure 8. MLP model structure

7.4. Gradient Boosting model

We apply gradient boosing algorithm on decision tree model and random forest model which is shortly denoted as gbdt and gbrf in section 8.

7.5. Convolutional neural network model

My custom CNN model 9 is with dropout regularization for a one-dimensional input shape. The architecture is designed for tasks such as time-series analysis or sequence-based data. The model begins with a 1D convolutional layer with 64 filters and a kernel size of 3, followed by max-pooling to capture essential features. A Layer Normalization step enhances the network's stability by normalizing the activations. [1]

The flattened output is then passed through a series of densely connected layers, incorporating dropout regularization with a specified rate of 0.05 to prevent overfitting. The dense layers progressively reduce the dimensionality of the representation, employing rectified linear unit (ReLU) activation functions for non-linearity. The final layer produces a single-unit output with a linear activation, making the model suitable for regression tasks.

In essence, the CNN architecture aims to learn hierarchical features from the input data, and the inclusion of dropout helps improve generalization by reducing over-reliance on specific neurons. This model strikes a balance between complexity and regularization, making it a versatile choice for my one-dimensional DNA methylation data.

8. Cross-Validation

In order to show generalizabilty, we apply 10-Fold Cross Validation to compare reliabilities and performace between models. The evaluation metrics are by Mean Sqaured Error (MSE) and R-square due to our age prediction task belongs to regression problem.

Figure 10 shows performace between models under same features (filtered or reduced CpG sites) while figure 11 shows performace of our model with stage1-CpG sites or reduced-CpG sites are better than Horvath's epigenetic clock. [3].

9. Future works

9.1. IV Estimation

Instrumental Variable (IV) estimation is a statistical technique to address endogeneity issues in regression analysis. Endogeneity arises when an explanatory variable is correlated with the error term as Figure 12b, which means the model violate the casual relationship as Figure 12a shows, leading to biased and inconsistent parameter estimates. IV estimation, particularly the Two-Stage Least Squares (2SLS) method, is designed to mitigate this problem.

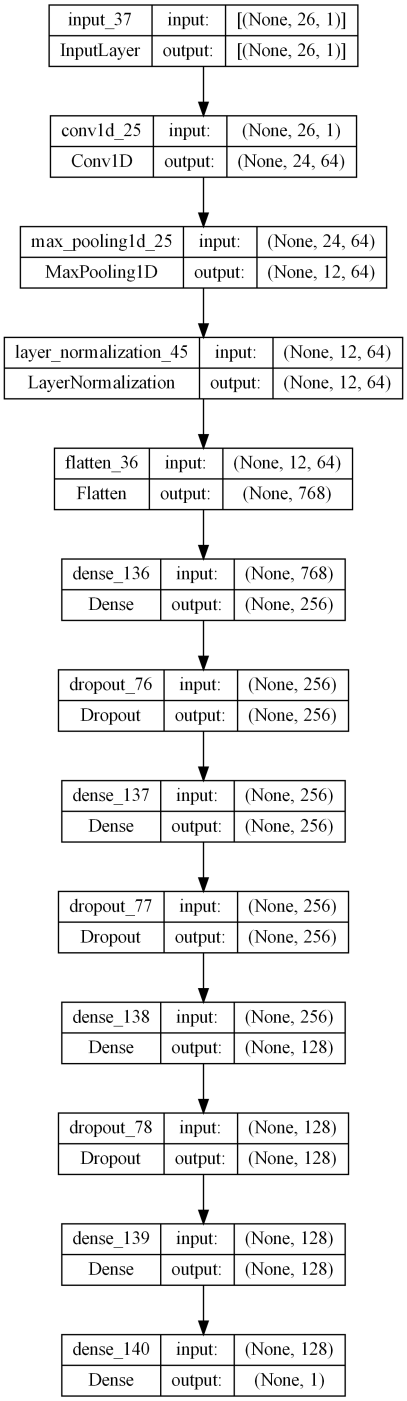


Figure 9. CNN model structure

In situations where a variable of interest is affected by unobserved factors, instrumental variables serve as proxies or instruments that are correlated with the endogenous variable but not directly related to the error term. The 2SLS

method involves two stages of regression:

1. In this stage, the endogenous variable is regressed on the instrumental variables. The predicted values from this regression, often referred to as the "fitted values," are then used as the instrument in the second stage.

2. The fitted values from the first stage are substituted for the endogenous variable in the main regression equation. This eliminates the endogeneity issue, allowing for unbiased and consistent estimation of the parameters.

Z is instrumental variable, X is independent variable(predictor or regressor), Y is dependent variable, u is error term.

We may introduce IV Estimation if Endogeneity existed at later stage of this project.

9.2. Causal Mediation Effect Analysis

Next stage application of our age prediction model, introducing casual mediation model, we can analyze the average casual mediation effects and average casual direct effects of mediator that accelerates aging speed such like smoking status of samples.

10. Conclusion

In conclusion, our study focused on predicting age based on DNA methylation levels at age-related CpG sites. We utilized regression models, random forest, gradient boosting, neural networks, and clustering algorithms to enhance our understanding of the aging process. Notably, we introduced a reduced set of CpG sites through clustering, aiming for getting more less features while still maintain model performance and interpretability.

Our results demonstrated that the gradient boosting random forest with the reduced set of CpG sites outperformed Horvath's epigenetic clock, showcasing the efficacy of our approach. The clustering of age-related CpG sites provided valuable insights into the diverse patterns of methylation changes during aging.

Moving forward, our future work includes exploring instrumental variable estimation to address potential endogeneity issues and applying causal mediation analysis to understand the impact of mediators on the aging process. These advanced statistical techniques can enhance the robustness of our age prediction models and contribute to a deeper understanding of the factors influencing biological aging.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 5

[2] Steve Horvath. Dna methylation age of human tissues and cell types. *Genome Biology*, 14(10):3156, 2013. 1, 2

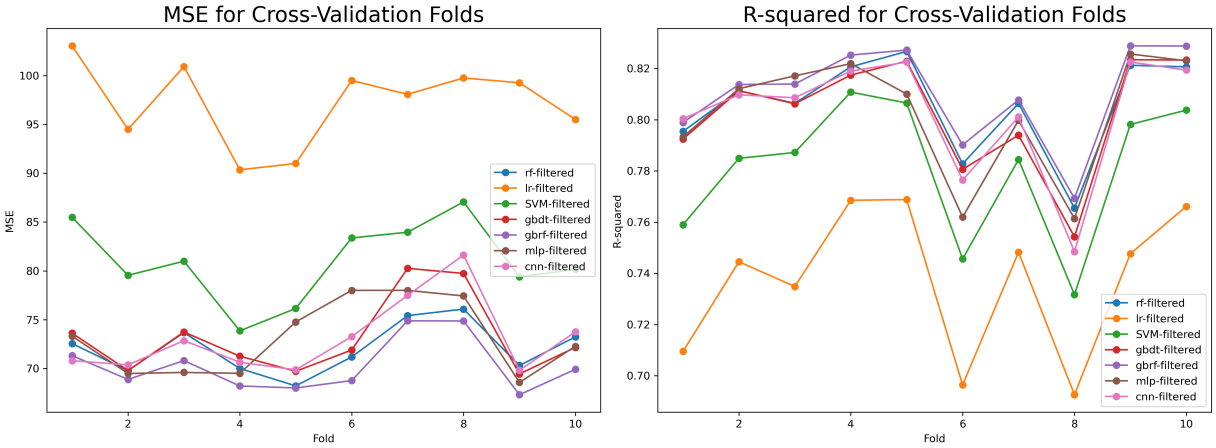
[3] Steve Horvath and Kenneth Raj. Dna methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews Genetics*, 19(6):371–384, 2018. 5

[4] Lechuan Li, Chonghao Zhang, Shiyu Liu, Hannah Guan, and Yu Zhang. Age prediction by dna methylation in neural networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(3):1393–1402, 2022. 1

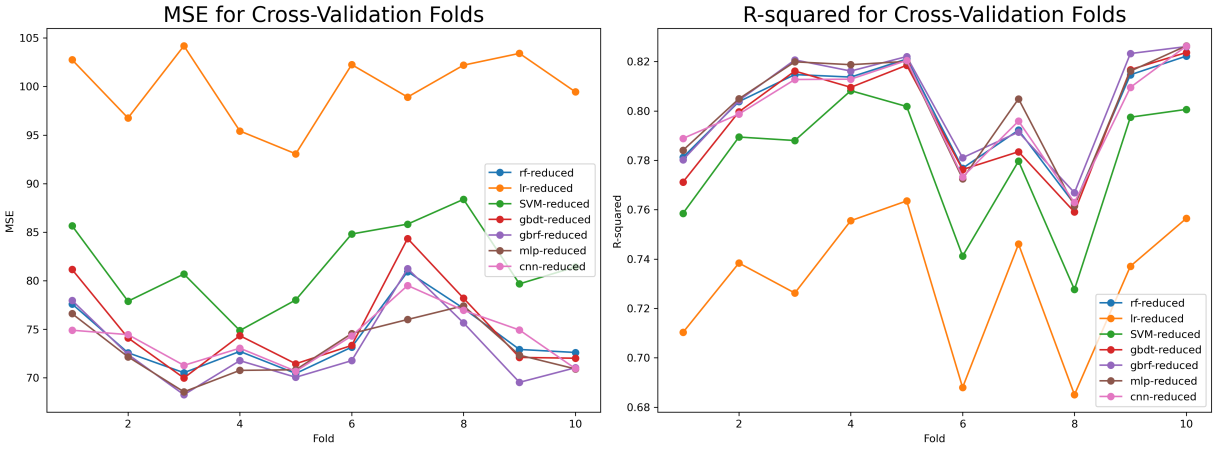
[5] Costa IG Marioni RE Ferreira MR Deary IJ Wagner W Lin Q, Weidner CI. Dna methylation levels at individual age-associated cpg sites can be indicative for life expectancy. *Aging (Albany NY)*, (2):394–401, 2016. 1, 2

[6] Pechlivanis S Hoffmann P Schmid M Weinhold L, Wahl S. A statistical model for the analysis of beta values in dna methylation studies. *BMC Bioinformatics*, (1):480, 2016. 2

[7] Suderman M. Langdon R. et al. Yousefi, P.D. Dna methylation-based predictors of health: applications and statistical considerations. *nature reviews genetics*, pages 369–383, 2022. 1, 2



(a) Cross-Validation for all models with stage1-CpGs



(b) Cross-Validation for all models with reduced-CpGs

Figure 10. 10-Fold Cross-Validation result of models under stage1-CpGs and reduced-CpGs (from stage1-CpGs after DBSCAN clustering)

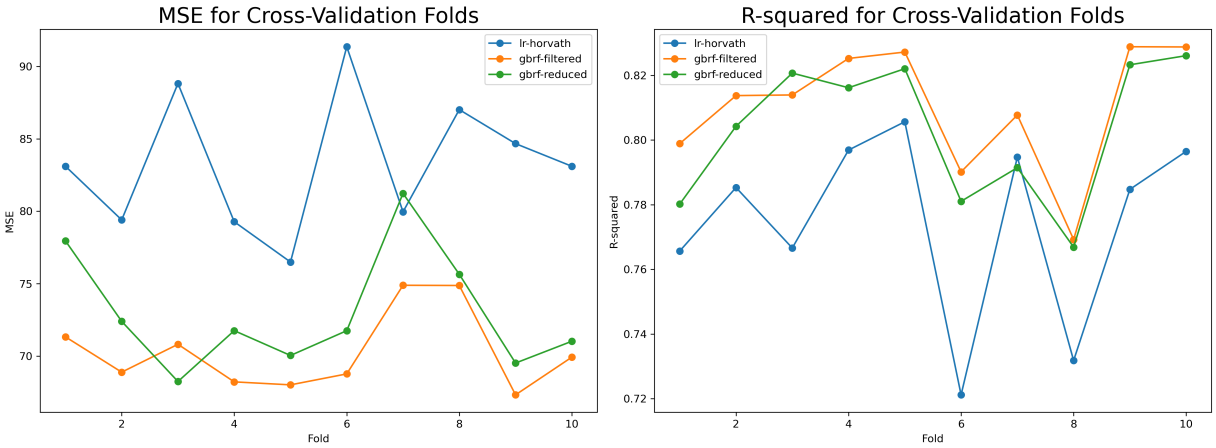
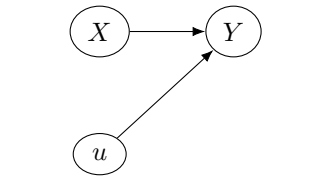
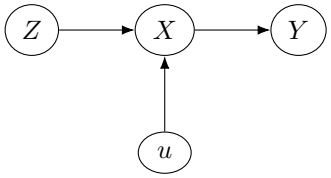


Figure 11. 10-Fold Cross-Validation result of our gbrf-filtered, gbrf-reduced and lr-horvath



(a) Causal Diagram 1



(b) Causal Diagram 2

Figure 12. Causal Diagrams