Biological Age Estimation Using DNA Methylation Sites

Sharanya Pathakota, Jay Katyan, Praneeth Uppari

December 7, 2023

Abstract

This project focuses on using DNA Methylation sites (the presence of these epigenetic markers) as the features to predict biological age. This paper aims to assess the effectiveness of various popular models in this field by evaluating their performance on a distinct data set, ultimately identifying the model with the highest efficacy. This paper focuses on elastic-net regression, random forest regression, support vector regression, and nu-support vector regression. Initially, we used a random sample of 20,000 CpG sites out of 485,512 sites as our features to be computationally viable. During this initial phase, we found support vector machine (SVM) to be the most effective. Through research, we were able to reduce the number of CpG sites used to a specific 353 sites that were found to be most indicative of biological age. After retraining our models, we observed across-the-board enhancements in mean absolute error values and correlation coefficients for both the validation and test sets. However, SVM still emerged as the most effective model. To demonstrate a practical application, we developed a classifier capable of predicting disease types based on each sample's methylation data. This model underscores the versatility of BA prediction and the potential for disease classification.

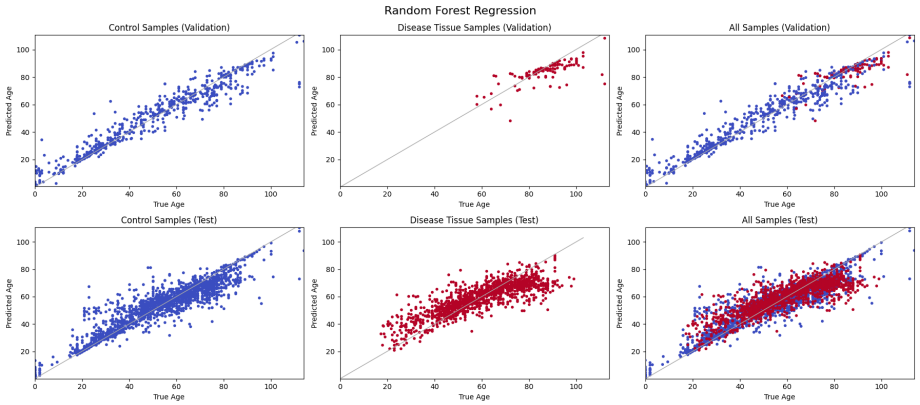
1 Statement of Problem

In assessing age-related health risks, chronological age has limitations due to its inability to capture the intricate influence of epigenetic factors. To address this gap, researchers are increasingly turning to biological age (BA) as a more nuanced metric. Unlike chronological age, BA captures behavioral and environmental factors, providing better predictions for age-related diseases such as cardiovascular diseases, Alzheimer's, and Parkinson's. Accurate BA prediction, crucial for estimating life expectancy and promoting healthier lifestyles, relies on DNA methylation. DNA Methylation regulates gene expression and the acceleration or change in the methylation at CpG sites offers a promising and established avenue for precise BA prediction. As we seek to develop a machine learning model for accurate biological age prediction, the potential impact on healthcare, mortality risk assessment, and disease classification is significant.

2 Results

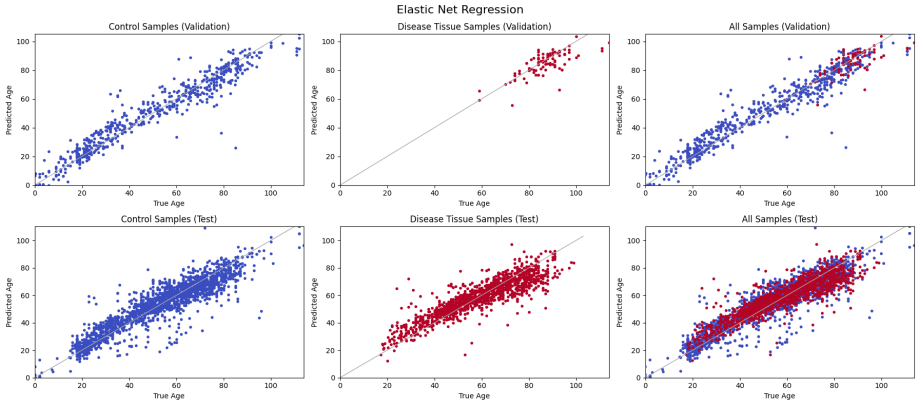
2.1 Random Forest Regression

Although the results of the Random Forest Regression were not as compelling as Support Vector Regression, it still had a very robust performance with MAE value of 4.71 years and R^2 values of 0.937 (validation) and 0.830 (test) for overall samples. Although the results indicate a strong fit for the overall samples, they also show the training error was much lower than the testing error for disease samples (shown in Table 1) potentially indicating an area of an over-fit model that can be further improved through more methylation sites and disease sample data.



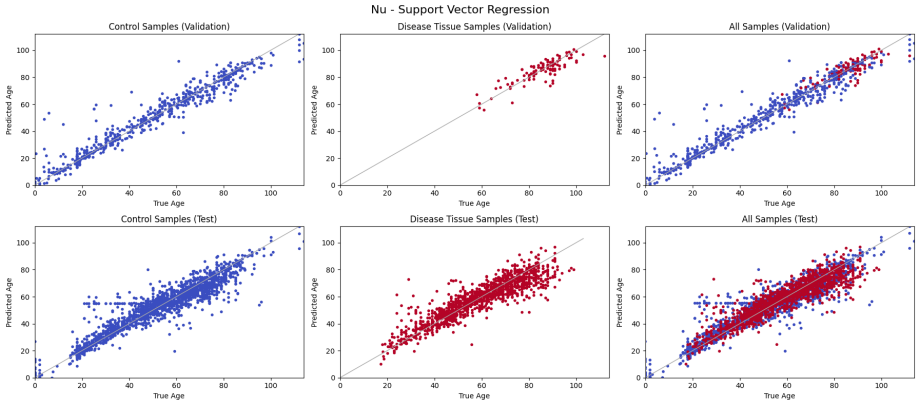
2.2 Elastic-Net Regression

The elastic net regression model demonstrated significantly improved accuracy, with a decreased MAE of about 4.60 years and a validation R^2 of about 0.94. The test set exhibits a MAE of about 5.19 and a test R^2 of about 0.84.



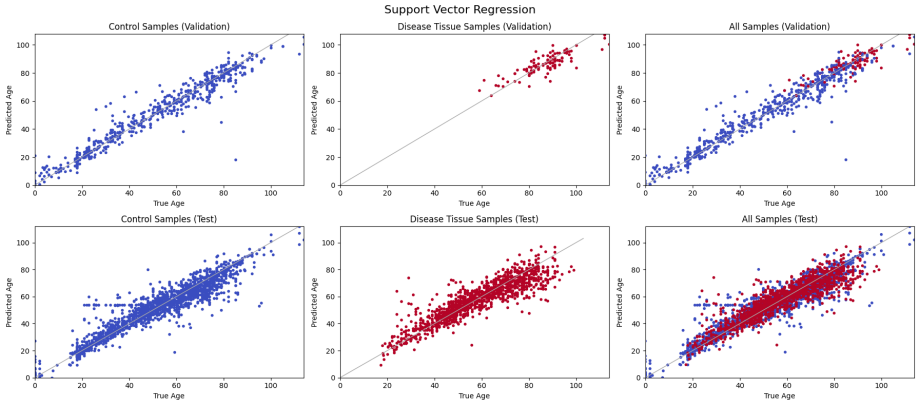
2.3 Nu-Support Vector Regression

The Nu-Support Vector was similar to the SVR model in that the values for these models was significantly better than the others. However, comparing this model with the performance of SVR, it can be seen that the MAE values for Nu-Support Vector were lower while the R^2 values were higher for SVR.



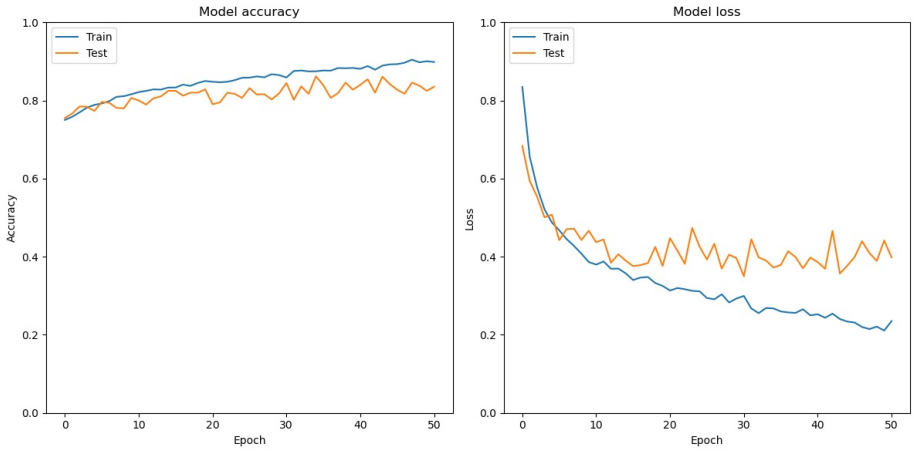
2.4 Support Vector Regression

The Support Vector Regression (SVR) model had the highest R^2 values compared to the other models, indicating a better fit for the dataset. The model also had the lowest MAE values compared to the others, indicating a better prediction accuracy. When looking at the different MAE values and the determination coefficients (R^2), it can be determined that the model did not have over-fitting and is well trained.



2.5 Feedforward Neural Network

Using a Feedforward Neural Network, a disease classification model was trained using data received from biological age prediction and the disease types. To decrease the effects of over-fitting, hyper-parameter tuning was used, specifically L2 regularization. ReLU was used as the activation function to introduce nonlinearity to the model in order to detect complex patterns. To increase performance, multiple iterations were ran while testing out the early-stopping's patience levels. The final testing accuracy was 83.21% which can be improved through further training.



2.6 Model Performance Comparison

For validation samples, all models achieved an R^2 value of 0.93. Notably, the support vector regression model outperformed others, attaining an R^2 value of 0.954. In test samples, models reached R^2 values of 0.83 across the board, while support vector regression obtained an R^2 value of 0.879.

Results for mean absolute error (MAE) values demonstrated consistency and reliability across all models, with values ranging from 3.359 to 5.216 for validation samples. For test samples, the range extended from 4.102 to 5.958. Once again, support vector regression demonstrated the best predictive accuracy for both control and disease samples.

Model	Metric	Val. MAE	Test MAE	Val. R^2	Test R^2
Random Forest	Overall	4.714	5.261	0.937	0.830
	Control	4.597	5.100		
	Disease	4.832	5.421		
Elastic-Net	Overall	4.604	5.197	0.936	0.844
	Control	5.216	5.958		
	Disease	3.993	4.436		
Nu-Support Vector	Overall	3.427	4.306	0.949	0.878
	Control	3.951	4.511		
	Disease	2.903	4.102		
Support Vector	Overall	3.359	4.340	0.954	0.879
	Control	3.757	4.506		
	Disease	2.961	4.174		

Table 1: MAE and R^2 Scores for Optimized Regression Models

3 Interpretation of results

Four distinct ML models have been developed through iterative optimization processes, and the evaluation of the model performances revealed commendable results across all models. In accordance with the methodology proposed by Horvath in 2013, 353 CpG sites were chosen utilizing a penalized regression model with elastic net regularization. This regression model selected the specific DNA methylation sites that are most predictive of biological age across a wide spectrum of tissues and cell types. Following feature selection, hyperparameters for each model were iteratively optimized and selected based on optimal performance.

Based on results across all models, support vector regression achieved the highest predictive accuracy based on R^2 and MAE metrics. Notably, all models demonstrated an ability to accurately predict biological age, with fairly similar R^2 and MAE values across all sample types.

As highlighted above, feature selection enabled a significant decrease in our features underscoring the role certain CpG sites play in biological aging. Incorporating these selected CpG sites, alongside meticulous hyperparameter tuning, not only resulted in a marked reduction in result variance but also set the stage for future explorations. Moving forward, there is an opportunity to delve deeper into the study, examining whether certain CpG sites exhibit enhanced effectiveness when applied to specific tissue types, providing valuable insights into the tissue-specific nature of biological aging. Additionally, due to computational limitations, we were not able to utilize a larger data set to test and train our model which could be a further point of improvement.

References

[1] H. Fan et al., "Chronological age prediction: Developmental evaluation of DNA methylation-based machine learning models," Frontiers, https://www.frontiersin.org/articles/10.3389/fbioe.2021.819991/full (accessed Oct. 25, 2023).

[2] Petkovich, D. A., Podolskiy, D. I., Lobanov, A. V., Lee, S.-G., Miller, R. A., amp; Gladyshev, V. N. "Using DNA methylation profiling to evaluate biological age and longevity interventions," Cell metabolism (accessed Oct. 24, 2023).

[3] S. Horvath, "DNA methylation age of human tissues and cell types," Genome biology, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4015143/ (accessed Dec. 7, 2023).

[4] M. E. Levine et al., "An epigenetic biomarker of aging for lifespan and healthspan," Aging, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5940111/ (accessed Dec. 7, 2023).

Contributions

3.1 Sharanya Pathakota

I worked on training and testing the Support Vector Regression model using the dataset. I also obtained the MAE and R^2 values for the validation set and test set while also graphing the results. Along with this, I worked on the SVR section and the Further Investigation section in the milestone report. I worked on creating and tuning the feedforward neural network and helped with the feature selection process. During this process, I read through different research papers to find the specific CpG sites that are optimal and helped map these results to our dataset. I also helped run the different models with the new CpG sites and helped with the report.

3.2 Jay Katyan

I worked on training and testing the Elastic-Net model and the Linear Regression model using the dataset. I also obtained the MAE and R^2 values for the validation set and test set while also graphing the results. Along with this, I worked on the Elastic-Net section, Linear Regression section, and the Model Performance Comparison section in the milestone report. I worked on creating and tuning the feedforward neural network and helped with the feature selection process. During this process, I read through different research papers to find the specific CpG sites that are optimal and helped map these results to our dataset. I also helped run the different models with the new CpG sites and helped with the report.

3.3 Praneeth Uppari

I worked on training and testing the Random Forest model using the dataset. I also obtained the MAE and R^2 values for the validation set and test set while also graphing the results. Along with this, I worked on the Abstract section and the Random Forest section in the milestone report. I worked on creating and tuning the feedforward neural network and helped with the feature selection process. During this process, I read

through different research papers to find the specific CpG sites that are optimal and helped map these results to our dataset. I also helped run the different models with the new CpG sites and helped with the report.