Improving and Interpreting Epigenetic Age Predictors

A Machine Learning Approach to Improving Epigenetic Age Predictors and Understanding How DNA Methylation Affects Aging

Elena Langens
Responsible Professor: Marcel Reinders[1]
Supervisors: Bram Pronk[1]**, Inez den Hond**[1]**, Gerard Bouland**[1]
[1] **EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Elena Langens
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Bram Pronk, Inez den Hond, Gerard Bouland, Kaitai Liang

Abstract

Understanding the mechanisms of aging can help us live longer and healthier lives. Epigenetic age predictors are machine learning models that use methylation levels at CpG sites to predict the biological age of the cell. Horvath's linear clock uses 353 CpGs with a median absolute error (MedAE) of 3.530, while the deep learning model AltumAge uses 20,318 CpGs to achieve a MedAE of 2.147. This study explores how to improve the accuracy of age predictors through model architecture selection, hyperparameter optimization, and feature selection. ElasticNet regression with recursive feature elimination achieved a MedAE of 2.820 using 341 CpGs, outperforming Horvath's clock. The two models shared 95 CpG sites, and gene enrichment analysis revealed that several associated genes are involved in stem cell regulation. Feature importance and model interpretation were performed using SHAP analysis, which indicated that age prediction cannot be captured by a small subset of CpG sites. It was concluded that epigenetics has an influence on stem cells, which was found to be a biomarker of aging. Aging remains a complex process that deep learning models may capture better.

1 Introduction

Gaining insight into how aging occurs within humans can help us age better and live longer and healthier lives. One reason why some people live longer is due to the influence of epigenetics on aging [1]. Epigenetics describes how other molecules in the cell interact with DNA [2]. A crucial part of epigenetics is DNA methylation, which involves the addition of a methyl group to the DNA. This chemical modification can alter the activity of a gene without changing the DNA sequence itself [3]. DNA methylation occurs at the so-called CpG sites, which are specific regions on a DNA strand where methyl groups can bind. These CpG sites are associated with certain genes and can influence their expression. Changes in methylation levels over time are closely associated with biological aging and serve as biomarkers for aging [4].

Machine learning models that predict the age of a cell based on methylation values are called epigenetic aging clocks. Horvath created one of the oldest clocks and predicts biological age based on methylation levels at 353 CpG sites. Horvath's clock is a linear model that uses data from multiple human tissues [5]. Hannum created another relevant linear model that uses only blood methylation values of 71 CpG sites [6]. With recent development in machine learning models, deep learning has offered new possibilities for aging clocks. DeepMAge [7] is a deep neural network based on around 1,000 CpG sites. While DeepMAge gets its methylation values from blood samples only, AltumAge [8] is a deep neural network based on multi-tissue data. AltumAge uses 20,318 CpGs and achieves better accuracy than the linear models and DeepMAge.

Although the above-mentioned research has contributed enormously toward developing better age predictors, some questions still remain unanswered. Both Hannum and DeepMAge do not use publicly available data and focus only on blood tissue. This makes their aging clocks less generalizable and limits their applicability to other types of tissue. Horvath and AltumAge, on the other hand, do focus on multi-tissue data and make use of publicly available data. These two models are therefore easier to reproduce, however, also have issues. Horvath's selected CpGs were chosen for their correlation with age, not because they are directly involved in the biological processes of aging. AltumAge suffers from a similar limitation, namely that a deep learning model is often seen as a black-box, where it is not clear which of the 20,318 CpG sites contribute the most to the prediction of age. Although both papers discuss how CpG sites are related with age, it remains unanswered which are the most relevant CpG sites and what they teach us about aging.

Additionally, the AltumAge paper explores the benefit of a higher number of CpG sites on deep learning models, but does not explore the effect of these CpG sites on other models in depth. In their work, they focus mainly on the deep learning model using different regularization techniques to improve it. However, they do not tune hyperparameters for other linear and non-linear models, failing to show their full potential. Machine learning models can also be improved by performing feature selection. When looking at AltumAge and Horvath's clock, they only use ElasticNet regression to perform feature selection, leaving other feature selection techniques unexplored.

It remains unanswered if we can further improve models by testing various architectures, using different regularization techniques, tuning hyperparameters, and performing feature selection. Using fewer features with better or similar performance can lead to new biological insights. Finding a feature set of 353 CpGs or less with better performance metrics than Horvath's clock would mean having a new aging clock which is more interpretable and predicts age better. Identifying overlapping CpG sites between Horvath's clock and a new aging clock and analyzing associated genes could lead to new insights into biomarkers of aging.

Due to the unexplored effect of certain machine learning techniques on age prediction, the question of how to reproduce or improve current age predictors based on epigenetic modification data arises. Understanding how methylation levels at specific CpG sites correlate with biological age requires identifying the sites most strongly linked to aging. Since the features of the machine learning models are the CpG sites, it raises the questions if we can interpret the most important features for prediction. After discovering the most relevant CpG sites, it is useful to find out what this

means and if we can use this knowledge to find biomarkers for aging.

A machine learning approach is taken to improve multi-tissue epigenetic age predictors. The effects of feature selection and different architectures on age prediction will be studied. The tuning of hyperparameters will be used as an optimization technique of different architectures. In order to understand how DNA methylation affects aging, we will analyze CpG sites with high importance shared between models in depth.

The Methodology, Section 2, describes how existing multi-tissue epigenetic age predictors are reproduced, improved, and interpreted. Section 3 discusses ethical considerations and reproducibility. Section 4 states the results of the reproduction, improvements, and feature analysis. The interpetation of age predictors and explanation of results will be given in Section 5. Section 6, summarizes the main findings of this research and proposes recommendations for further research.

2 Methodology

Dataset

For model reproduction and training, publicly available human DNA methylation data is used. The preprocessed data is downloaded from the AltumAge [8] repository. The data consists of DNA methylation beta values (see Appendix A) from the GEO datasets from the 27K and 450K Illumina arrays [9], the 27K and 450K platform dataset from ArrayExpress [10] and the cancer-related datasets from TCGA [11]. 27K and 450K mean the amount of CpG sites, namely 27,000 and 450,000. A 60/40 train/test split is applied within the dataset, consistent with the AltumAge repository, enabling model comparison. All models are trained with the dedicated train dataset and evaluated using the separate test dataset. Data samples consist of tissue type, age, sex, and CpG sites with their methylation value. The different types of tissue are visualized in Figure 1.

Setup

To make the datasets compatible with each other, 20,318 overlapping CpG sites are selected. The methylation values are scaled with the RobustScalar class of the scikit-learn library in Python. The scalar normalizes the features and provides robustness to outliers. The ages of are transformed using Equation 1 and Equation 2 developed by Horvath [5].

$$\text{t\_age} = \begin{cases} \log(\text{age} + 1), & \text{if age} \leq 20 \\ \text{age} - 20 + \log(21), & \text{if age} > 20 \end{cases} \tag{1}$$

$$\text{age} = \begin{cases} \exp(\text{t\_age}) - 1, & \text{if t\_age} \leq \log(21) \\ \text{t\_age} - \log(21) + 20, & \text{if t\_age} > \log(21) \end{cases} \tag{2}$$

This age transformation stabilizes the variance of the ages and improves the performance of the regression

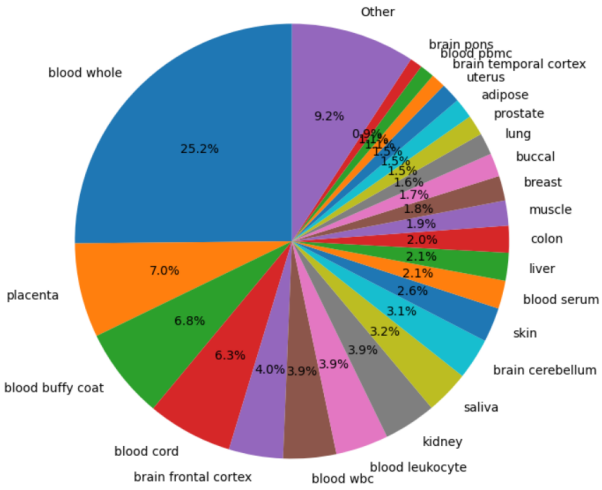

Figure 1: Tissue type distribution across both training and test data with grouped smaller values, named as other.

models. The ages are transformed to t_age before training the model and anti-transformed when using the test data to predict the ages.

Reproduction and evaluation

Reproduction of Horvath's model and AltumAge is done using the pyaging library[1] and the test data. Both models can be used directly from this Python library and run with the test data. Their performance is assessed with the Median Absolute Error (MedAE), Mean Squared Error (MSE), and Pearson's correlation coefficient (R). These metrics evaluate the accuracy, robustness, and predictive relationship of the models. After running the models, the accuracy is compared with the performances measured in the AltumAge paper [8]. They use the same test data making this comparison accurate and relevant. Other models trained in this research are also evaluated using the same performance metrics to provide fair comparisons.

Model Selection and Optimization

Linear regression and tree-based models are considered in this research. Linear models with regularization techniques have a low risk of overfitting, provide high interpretability, and have a fast training speed. Tree-based models have a higher risk of overfitting, and have lower interpretability, but are able to capture more complex, nonlinear interactions between CpG sites. Deep learning models are not considered due to the slow training speed and the high overfitting risks. Linear regression is analyzed with L1 (Lasso), L2 (Ridge), and combined regularization (ElasticNet). XGBoost was considered as a tree-based model for this research. It is known to have higher accuracy compared to other tree-based models, such as Decision Trees or Random Forest.

Lasso regression minimizes the objective function (Obj) shown in Equation 3, which adds an L1 regularization term to the ordinary least squares function. The strength of regularization is controlled by α, where

[1] https://pyaging.readthedocs.io/en/latest/

$\alpha = 0$ corresponds to standard linear regression, and higher values increase the degree of regularization. Lasso regression shrinks the coefficients (β_j) and sets some to 0, effectively performing feature selection [12]. This type of feature selection is called embedded feature selection, where the selection is integrated into the training of the model [13].

$$\text{Obj} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha\sum_{j=1}^{p}|\beta_j| \tag{3}$$

Ridge regression minimizes the objective function shown in Equation 4, adding an L2 regularization term to the ordinary least squares function. The regularization strength is again controlled by α, where higher values of α lead to a larger shrinkage of the coefficients (β_j). Ridge regression prevents overfitting and handles the correlation between features by penalizing large coefficients. This is done by squaring β_j instead of taking the absolute value Unlike Lasso regression, it does not set coefficients to zero, effectively not performing feature selection. [14].

$$\text{Obj} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha\sum_{j=1}^{p}\beta_j^2 \tag{4}$$

ElasticNet regression combines L1 and L2 regularization, as shown in Equation 5 [15]. It therefore performs both embedded feature selection and prevents overfitting. ElasticNet balances Lasso and Ridge using the hyperparameter λ (or $l1\_ratio$), where λ = 1 is equivalent to Lasso and λ = 0 equivalent to Ridge regression. The regularization strength is again controlled by α. The L1 and L2 regularization strength is calculated by multiplying α by λ and α by 1 - λ, respectively.

$$\text{Obj} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha\left[\lambda\sum_{j=1}^{p}|\beta_j| + (1-\lambda)\sum_{j=1}^{p}\beta_j^2\right] \tag{5}$$

The hyperparameters (α and $l1\_ratio$) are tuned using LassoCV, RidgeCV and ElasticNetCV. These Python classes perform automated grid search with k-fold cross-validation, exploring all combinations of hyperparameters and choosing the best performing one. A 5-fold cross-validation was chosen, meaning that the training data is split in 5 subsets, of which 4 are used for training the model and 1 for validating it, repeated 5 times. A k-fold of 5 offers a balance between the computational cost and the prediction error [16]. Table 1 displays the hyperparameters with which the grid search was performed.

XGBoost is a gradient boosting algorithm that combines decision trees. This means that it trains weak decision trees sequentially, where each new tree tries to correct the errors made by the previous one and combines them to create a stronger predictive model [17]. XGBoost implicitly performs feature selection when training. It does so by measuring the frequency with which a feature is used in a split across all trees. Features with a frequency of 0 are therefore not used and are implicitly eliminated. XGBoost minimizes an objective function consisting of a loss term and a complexity penalty for each tree, as shown in Equation 6 [18]. The loss term measures the difference between the predicted and actual values and the regularization term penalizes overly complex trees.

Model	α	**l1_ratio**
Linear	–	–
Ridge	0.01, 0.1, 1, 10, 100	–
Lasso	0.001, 0.01, 0.1, 1, 10	–
ElasticNet	0.001, 0.01, 0.1	0.2, 0.5, 0.7

Table 1: Linear regression models and their hyperparameter values used in automated grid search cross-validation to find the optimal values.

$$\text{Obj} = \sum_{i=1}^{n} Loss(y_i, \hat{y}_i) + \sum_{k=1}^{K}\Omega(f_t)$$
$$\text{where} \quad \hat{y}_i = \sum_{k=1}^{K}\eta f_k(x_i) \tag{6}$$

Each tree f_t is added sequentially to correct for the error of the previous prediction. The final prediction $\hat{y}_i$ is the sum of all tree outputs, scaled by the learning rate η, as shown in Equation 6. Here, K (n_estimators) is the number of trees and η is the learning rate. K influences the complexity of the model and η controls how much each tree contributes. The maximal depth of each tree is determined by the hyperparameter max_depth, which influences the complexity of each tree. Overly complex trees are penalized by Ω.

Tuning these 3 hyperparameters is done with manual grid search without cross-validation. Cross-validation, when used with XGBoost, can significantly increase the duration of training. In manual grid search, a grid consisting of all possible combinations of the hyperparameters is looped through, after which the best performing hyperparameter combination is selected. The values used for the grid search are displayed in Table 2.

Model	**max_depth**	**n_estimators**	**learning_rate**
XGBoost	3, 5	100, 200, 500	0.05, 0.1

Table 2: XGBoost and its hyperparameter values used in manual grid search to find the optimal values.

Feature Selection

Feature selection was used to remove less relevant CpG sites and compare model performance with more versus fewer CpGs. The goal is to outperform Horvath's model using fewer CpGs. Two feature selection methods are explored, feature filtering and recursive feature elimination (RFE). Both of these methods are applied to ElasticNet and XGBoost with

optimized hyperparameters. RFE requires a model that identifies the most important features, so ElasticNet and XGBoost were chosen to compare linear and tree-based approaches. ElasticNet was preferred over Lasso as it combines L1 and L2 regularization.

Feature filtering is performed prior to model training using the Python class SelectKBest, a fast univariate method that helps reduce overfitting [13]. The feature space is reduced from 20,318 CpG sites to 350 using the f_regression scoring function. This function evaluates the linear relationships between each feature and the target. It calculates the F-statistic and p-value (see Appendices B and C) and ranks the features accordingly.

RFE removes the least important features and iteratively retrains the model with the smaller subset of the original feature set [19]. The Python class RFE was used for this and was given the value 350 for the amount of features to select. It uses a step of 0.1, which means that 10% of the features are removed each iteration until 350 remain from the original 20,318.

Feature Importance
SHAP values are computed for the model with the lowest error, highest correlation, and with an amount of features close to 353. These values were computed using the SHAP Python library. Depending on the architecture, either the Python class LinearExplainer or TreeExplainer from the SHAP library is used. These two classes compute SHAP values for linear and tree-based models, respectively. SHAP values indicate the contribution of each feature to the prediction and are calculated using a game-theoretic approach [20]. SHAP allows for the comparison of feature importance across different model architectures. Visualizing SHAP values is done using a beeswarm plot, which shows how each feature influences the prediction.

Biological context
The Illumina 450K annotation file[2] is used to map CpG sites to genes, on which a gene set enrichment analysis is performed. For the best-performing model with a feature set close to 353, an overlap of its features with Horvath's CpGs is found. The set of genes belonging to the overlapping CpG sites is entered into Enrichr [21] with all the genes in the Illumina 450K annotation file as background. Enrichr is a gene set enrichment analysis tool which compares an input gene list to a collection of annotated gene sets across approximately 200 gene set libraries. These libraries group genes based on shared biological functions or other annotations.

Enrichr evaluates whether the input gene set is statistically enriched for overlap with any of the annotated sets, indicating biological relevance [21]. Within Enrichr, the GO Biological Process 2025 library was selected for analysis. This library contains gene ontology terms that describe biological functions. The enrichment results were ordered by the p-value, which reflects the statistical likelihood that the observed overlap occurred by chance. P-values lower than 0.05 are often considered statistically significant.

[2]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL13534

3 Responsible Research

Doing research comes with various ethical considerations. This research makes use of publicly available biological data from GEO, ArrayExpress and TCGA [9; 10; 11]. Using genomic data can introduce risks if the person it is from can be identified. The sources have therefore made sure that personal information is removed to protect the identity of individuals. The participants also gave their informed consent to have their genomic data taken and used. It should be noted that these public datasets lack representation of diverse populations. Due to this under-representation, there is a potential bias of the age predictors trained in this paper.

Reproducibility is ensured by giving a step-by-step transparent instruction on how the results were reached. These instructions can be found in section 2 as well as in the pyaging documentation[1]. The methodology describes in detail where the data can be found, how it was preprocessed, and how the different models were trained. Following the methodology should therefore result in the same performance metrics, ensuring reproducibility and replicability. The performance of the models is comparable and reliable since they were all assessed with the same unseen test set. The genes used for the gene enrichment analysis can be found in Appendix F, which makes our results easier to validate. The code made for this research is publicly available on GitHub [3], again ensuring transparency and reproducibility.

4 Results

The 60/40 split of the data resulted in a training set of 8050 samples and a test set of 5455 samples. The age distribution of the training data shows that most samples had a age of around 0, as shown in Figure 2. A linear model with a smaller error, higher correlation, and fewer features than Horvath's model was found.

Reproduction
We successfully reproduced the results of Horvath's clock and AltumAge using the test data. As shown in Table 3, AltumAge outperformed Horvath's model, achieving a lower median absolute error and mean squared error, while maintaining a higher correlation.

Model	CpGs	MedAE	MSE	R
Horvath's clock	353	3.530	71.030	0.951
AltumAge	20,318	2.147	29.077	0.980

Table 3: Evaluation metrics and amount of CpG sites used of Horvath's clock and AltumAge using the test set and pyaging library.

[3]https://github.com/elena011/rp-delft/tree/dev

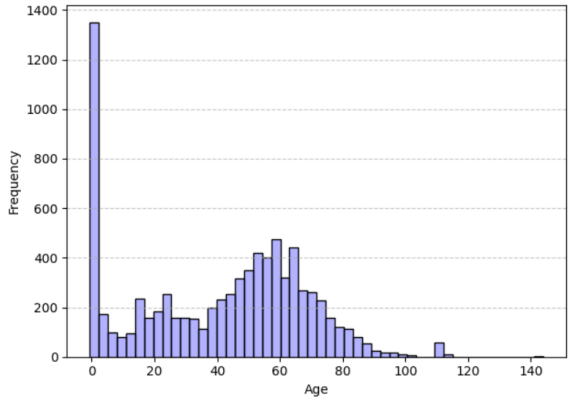

Figure 2: Age distribution of training data with the x-axis displaying the age and y-axis displaying the amount of samples with this age, the frequency.

Model selection and optimization

The training of linear architectures with 20,318 CpGs resulted in the following performance metrics and optimal hyperparameters. RidgeCV chose α = 200, LassoCV chose α = 0.001 and ElasticNetCV chose α = 0.01 and an l1_ratio of 0.2. The computed metrics of their performance can be found in Table 4. ElasticNet regression showed the lowest error and the highest correlation (MedAE = 2.620, MSE = 38.401, R = 0.974) and is highlighted in bold. After ElasticNet, Lasso regression had the lowest error, followed by Ridge regression. The linear model without regularization had the highest error of the four. Compared to Horvath's clock, there is improvement in performance metrics (ElasticNet regression with MedAE = 2.620 versus Horvath with MedAE = 3.530). None of the trained linear models have a lower error than AltumAge (MedAE = 2.147).

Since Lasso and ElasticNet regression perform embedded feature selection, their selected number of features is also displayed in Table 4. In this table, we can see that ElasticNet selected fewer features than Lasso (2,551 compared to 3,967). ElasticNet and Lasso regression both perform L1 regularization, but have regularization strengths due to different hyperparameters. ElasticNet has a stronger L1 regularization strength, namely 0.01 $\times$ 0.2 = 0.002, compared to 0.001 for Lasso regression. The number of selected CpG sites exceeds Horvath's 353 CpG sites, showing possibilities for prior feature selection.

The tuning of XGBoost its hyperparameters resulted in the best performance metrics when max_depth was equal to 5, n_estimators equal to 500 and learning_rate equal to 0.1. Of the 20,318 features it used 8,444 and the rest of the features had a used frequency of 0. Its results can be viewed in table 4.

XGBoost performed better than the Linear regression model without regularization but underperformed in comparison to Ridge, Lasso and ElasticNet regression. XGBoost also selected more features than Lasso or ElasticNet regression (8,444 compared to 3,967 and 2,551). XGBoost did not outperform AltumAge with its MedAE of 2.147. However, XGBoost did have a lower error than Horvath's clock in performance metrics (MedAE = 2.996 versus MedAE = 3.530), but when comparing the 8,444 features to Horvath's 353 features, XGBoost also has the possibility to improve with prior feature selection.

Model	CpGs	MedAE	MSE	R
Linear	20,318	3.455	61.511	0.959
Ridge	20,318	2.960	47.470	0.968
Lasso	3,967	2.753	42.767	0.971
ElasticNet	2,551	**2.620**	**38.401**	**0.974**
XGBoost	8,444	2.996	47.426	0.968
ElasticNet(SKB)	262	3.485	66.173	0.955
ElasticNet(RFE)	341	**2.820**	**44.085**	**0.970**
XGBoost(SKB)	350	3.094	48.965	0.967
XGBoost(RFE)	350	3.029	45.307	0.969

Table 4: Evaluation metrics and amount of CpG sites used to predict age of linear and tree-based architectures. Above the horizontal line are the models trained without prior feature selection, below are the models trained with prior feature selection, where SKB stands for SelectKBest and RFE for Recursive Feature Elimination. The best performance metrics for models with and without feature selection are highlighted in bold.

Feature Selection

Table 4 compares the two feature selection methods on ElasticNet (α = 0.01, l1_ratio = 0.2) and XGBoost (max_depth = 5, n_estimators = 500, learning_rate = 0.1). The SelectKBest method followed by ElasticNet (ElasticNet(SKB)) selected 262 CpG sites, indicating that ElasticNet reduced the 350 CpG sites to 262. RFE method combined with ElasticNet (ElasticNet(RFE)) resulted in 341 CpG sites. Using SelectKBest and RFE with XGBoost (XGBoost(SKB) and XGBoost(RFE)) both resulted in 350 features being selected.

Of the four models trained with feature selection, ElasticNet(RFE) showed the lowest error and the highest correlation (MedAE = 2.820, MSE = 44.085, R = 0.970). The two XGBoost models had a larger error and lower correlation but had very similar performance to each other. XGBoost(RFE) had a slightly lower MedAE than XGBoost(SKB) (MedAE = 3.029 compared to 3.094). ElasticNet(SKB) showed a larger error and lower correlation than both the XGBoost models and ElasticNet(RFE). Unlike the XGBoost models, there was more difference between ElasticNet(SKB) and ElasticNet(RFE). ElasticNet(SKB) showed a MedAE of 3.485, compared to the MedAE of ElasticNet(RFE), 2.820.

ElasticNet and XGboost's performance did not benefit from feature selection, while they did improve compared to Horvath's clock. ElasticNet(RFE) had a higher MedAE than ElasticNet, namely 2.820 compared to 2.620. XGBoost(RFE) had a MedAE higher than XGBoost, 3.029 compared to 2.996. ElasticNet(RFE) showed higher errors than Ridge, Linear, and XGBoost but showed lower errors compared to Lasso Regression. Additionally, ElasticNet(RFE) showed a MedAE lower than Horvath's clock (2.820 versus 3.530) while also using fewer features (341

versus 353). There were 95 overlapping CpG sites between Horvath's 353 and ElasticNet(RFE) 341 sites, as shown in Figure 3.

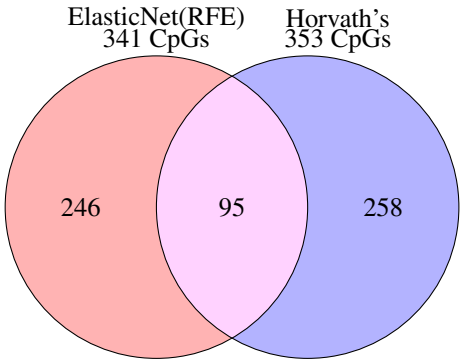

Figure 3: Venn diagram showing overlap between CpG sites selected by ElasticNet(RFE) (341 sites) and Horvath's epigenetic clock (353 sites). A total of 95 CpG sites are common to both models.

Biological context

The associated genes of the 95 overlapping CpG sites between Horvath's 353 and ElasticNet(RFE) are listed in Appendix F. From the enrichment analysis of the 95 genes, one biological process (Positive Regulation of Stem Cell Differentiation) had an adjusted p-value close to 0.05. Its adjusted p-value was 0.06773, the p-values of the other biological process were all 0.2365 or higher. The complete enrichment analysis of the 95 genes against the GO Biological Process 2025 library can be found in Appendix D.

The genes overlapping with this process were TCF15, HOXB4 and LTBP3. TCF15 encodes proteins that are involved in the regulation of the mesoderm, a germ layer found in embryos [22]. HOXB4 plays a role in embryo development, but is also involved in adult stem cell renewal and expansion [23]. LTBP3 is known to regulate TGF-beta, a signaling protein involved in cell growth and immune responses [24]. Mapping these genes back to their associated CpG sites using Appendix F gives the following sites: cg22449114 (TCF15), cg21460081 (HOXB4) and cg08965235 (LTBP3).

SHAP analysis

The beeswarm plot of the CpG sites corresponding to the three genes involved in the enriched ontology term can be seen in Figure 6. Each dot in the beeswarm plot represents a sample of the data set. The x-axis indicates the SHAP value, and the y-axis indicates the different CpG sites, showing the SHAP value for each sample for the specific CpG site. The red dots represent high methylation values, while the blue dots represent low methylation values.

For the first two CpG sites (cg08965235 and cg21460081), we can see that a high feature value (red) contributes to a low SHAP value (left of x-axis), while a low feature value (blue) contributes to a high SHAP value (right of x-axis). For the third CpG site,

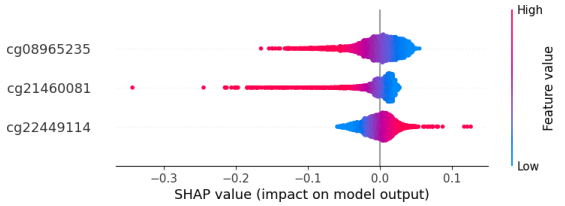

Figure 4: Beeswarm plot of the 3 CpG sites ordered from highest to lowest mean SHAP value. These CpG sites are associated to the three genes which are involved in the biological process called Positive Regulation of Stem Cell Differentiation, found in the GO Biological Process 2025 library.

cg22449114, a high feature value contributes to a high SHAP value and a low feature value contributes to a low SHAP value. This is seen in the figure with the second site having blue values right of the x-axis and red values left of the x-axis. Based on this information, we can say that the CpG sites cg08965235 and cg21460081 correlate negatively with age, while cg22449114 correlates positively with age.

Cg08965235 had the highest mean SHAP value of the three, approximately 0.022 followed by cg21460081 and cg21460081 with 0.016. The highest SHAP value found for the model was approximately 0.087. The contribution of the 95 overlapping genes was 35.33% of the total SHAP value, while 95 of the 341 CpG sites is approximately 28%. The top 10 CpG sites explain 10.03% of the prediction of the model and a beeswarm plot of them can be found in Appendix E. 80% of the SHAP values were explained by 195 of the 341 features, of which 5 contributed more than 1% to the total importance.

5 Discussion

Improving current multi-tissue epigenetic age predictors (Horvath's clock and AltumAge) was done by developing an age predictor with better performance metrics and less CpG sites. We found that ElasticNet combined with recursive feature elimination outperformed Horvath's model with a MedAE of 2.820, compared to 3.350 and used less CpG sites (341 versus 353). Between their CpG sites, there was an overlap of 95 features that contributed approximately 35% to the total SHAP value. However, we were unable to develop a model that performed better than AltumAge (MedAE = 2.147).

For linear models, it was found that a combination of L1 and L2 regularization resulted in the best performance. Lasso outperformed Ridge regression, possibly due to its ability to perform feature selection via L1 regularization. This could indicate that smaller subsets of CpG sites are more effective for linear age predictors, while having more CpG's introduces more noise. ElasticNet then improved Lasso regression with L2 regularization and evenly distributing weights among coefficients. In this way, it prevents overfitting while using only CpGs which have a high contribution toward age prediction.

The degree of regularization also gives us insight into how methylation levels at CpG sites contribute to age prediction. The high value for α (200) in Ridge regression suggests that strong regularization was necessary to avoid overfitting. This could be due to having a too complex model, again indicating that linear age predictors benefit from a smaller set of features. In contradiction to this, the low value of α value (0.001) for Lasso regression indicates that linear age predictors do not benefit from an overly sparse model. The hyperparameters of ElasticNet also support this claim, a higher L2 ratio was preferred over a higher L1 ratio ($l1\_ratio$ = 0.2 $\times$ 0.01 = 0.02, $l2\_ratio$ = 0.8 $\times$ 0.01 = 0.08). Linear age prediction seems to benefit from some feature reduction, but not from a very small subset of features. This proposes the idea that aging is influenced by a varied set of epigenetic changes, rather than a small number of isolated CpG sites.

XGBoost outperformed linear regression without regularization, but failed to outperform the other linear architectures. However, it still outperformed the baseline linear model, namely Horvath's clock. This shows that tree-based models like XGBoost provide competitive performance in epigenetic age prediction tasks. XGBoost performed the best with deeper trees (max_depth = 5), a larger number of estimators (n_estimators = 500), and a standard learning rate (learning_rate = 0.1). The larger depth and number of estimators again suggest that the relationship between methylation levels and age is complex and involves nonlinear interactions among CpG sites.

Recursive feature selection performed better than statistical feature filtering for linear and tree-based models. RFE selects features which are actually important for the model's prediction, accounting for feature interaction, while feature filtering fails to account for this. CpG sites can appear near each other, introducing a correlation between their methylation values. A statistical filtering approach can miss these interactions between CpG sites, which could explain its poorer performance.

The enrichment analysis identified one ontology term with an adjusted p-value of approximately 0.06, indicating weak evidence of enrichment. Although not statistically significant, this result may still suggest a potential biological signal. The ontology term was called Positive Regulation of Stem Cell Differentiation and had 3 related genes. Stem cells are cells that have the ability to renew themselves and can differentiate into different types of cells [25]. Stem cell self-renewal can be disrupted during aging [26], indicating stem cells can serve as biomarkers for aging. There also exist explicit studies on how epigenetics, especially DNA methylation, influence the regulation of stem cell differentiation [27].

The study on the influence of epigenetics on the regulation of stem cell differentiation highlights the function of DNA methylation in embryo development. The embryo undergoes an epigenetic reprogramming that starts with demethylation, after which a rebuilding of methylation patterns is started. If this reprogramming is incomplete, the embryo does not develop properly or could lead to premature death [27], ultimately influencing the age of the embryo. The three associated genes are all involved in embryol development or cell growth, which explains their contribution to age prediction across models. Incomplete reprogramming could mean a hypomethylated CpG site, which explains why HOXB4 and LTBP3 are negatively corrolated with age, as derived from the beeswarm plot in Figure 6. The age distribution in Figure 2 further explains the enrichment of this process, since it consists primarily of ages around 0. This is due to the use of placental and blood-cord data. Different biomarkers of aging could have been found if the age of the training data was not dominated by ages around 0.

The SHAP analysis showed us that the contribution of the 95 overlapping genes explained a relatively high amount of the total SHAP values. Combining this with the fact that these CpG sites are found in Horvath's clock indicates that they offer some robustness in age prediction. However, the SHAP values also indicate that the age prediction cannot be captured by a small set of features since there was no small subset of features that explained a large percentage of the SHAP values. It should be noted that the 10 most important features explained 10% of the prediction of the model, but there were only 5 CpG sites that contributed more than 1% to the importance. This leads to the belief that aging is a process too complex to be captured by only a small subset of CpG sites, which explains why deep learning models such as AltumAge, trained with a large number of CpG sites, perform better at age prediction.

6 Conclusions and Future Work

We developed a multi-tissue DNA methylation-based aging clock using ElasticNet regression with recursive feature elimination, achieving a MedAE of 2.820 with 341 CpG sites. This clock improved Horvath's clock (MedAE 3.530, 353 sites), and the two models shared 95 CpG sites. Although ElasticNet(RFE) did not outperform the deep learning model AltumAge, it offered a compact and interpretable feature set. Several overlapping features were associated with genes involved in stem cell regulation. Epigenetics and its influence on stem cells was found as biomarker for aging. However, deep learning models offer the greatest potential to fully capture the complexity of aging.

Aging clocks can still be improved in many ways. Most of the training data had an age of 0, coming from embryos, which leads to age predictors that work well mainly for embryo data. Future epigenetic age predictors could be based on data with a more evenly spread age distribution. This would lead to

the finding of different biomarkers of aging and more generalizable age predictors. It would be worth testing the developed age predictor on data sets grouped by age to assess how well it generalizes between different age groups.

Improving on deep learning architectures such as AltumAge could be done using either more features, different regularization techniques, or a different data set. The models trained in this research could also still be improved. The performance of XGBoost could still benefit by tuning more hyperparameters such as the min_split_loss or min_child_weight and using cross-validation to further assess the performance of the hyperparameters. Biologically informed feature selection prior to training models could also lead to new insights and possibly better age predictors.

A Beta Values

Beta values represent the ratio of methylated probe intensity to the sum of methylated and unmethylated probe intensities. They range from 1 (fully methylated) to 0 (fully unmethylated) [28].

B F-statistic

The F-statistc measures the strength of the linear relationship between the feature and the target.

C p-value

The p-value tests the null hypothesis that the feature is not linearly related to the target.

D Gene enrichment analysis

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Positive Regulation of Stem Cell Differentiation (GO:2000738)	0.00009901	0.06773	39.94	368.22
2	Positive Regulation of Ossification (GO:0045778)	0.0008067	0.2365	18.24	129.91
3	Protein Autoprocessing (GO:0016540)	0.001225	0.2365	46.87	314.27
4	Regulation of Execution Phase of Apoptosis (GO:1900117)	0.002310	0.2365	32.44	196.96
5	Organelle Organization (GO:0006996)	0.002561	0.2365	4.05	24.17
6	Retrograde Protein Transport, ER to Cytosol (GO:0030970)	0.002631	0.2365	30.13	178.95
7	Superoxide Anion Generation (GO:0042554)	0.003334	0.2365	26.36	150.33
8	Negative Regulation of Apoptotic Process (GO:0043066)	0.005855	0.2365	3.46	17.79
9	Regulation of Cholesterol Metabolic Process (GO:0090181)	0.005902	0.2365	19.16	98.36
10	Sphingolipid Catabolic Process (GO:0030149)	0.008548	0.2365	15.61	74.34

Figure 5: Top 10 enriched GO Biological Process terms identified by Enrichr using genes mapped from the 95 overlapping CpG sites. Terms are ranked by p-value. Positive Regulation of Stem Cell Differentiation shows the strongest enrichment (adjusted p-value = 0.0677)

E Beeswarm plot

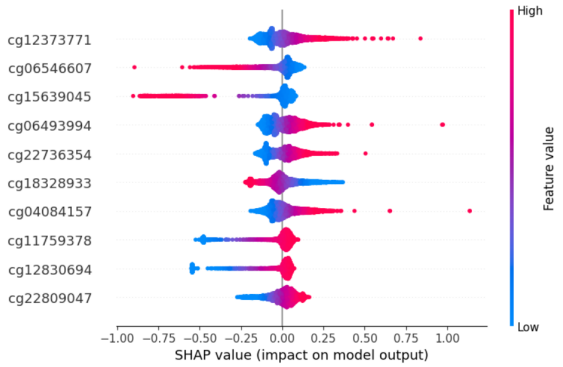

Figure 6: Beeswarm plot of the 10 CpG sites with the highest mean SHAP value for ElasticNet(RFE), ordered from highest to lowest mean SHAP value.

F Overlapping CpG sites between ElasticNet(RFE) and Horvath's model

Table 5: The 95 overlapping CpG sites between ElasticNet(RFE) and Horvath's model and their associated gene(s).

CpG site	Gene
cg00075967	STRA6
cg00091693	KRT20
cg00374717	ARSG
cg01262913	DSCR9
cg01459453	SELP
cg01968178	REEP1
cg02154074	AUP1, HTRA2
cg02335441	ASTE1, NEK11
cg02479575	C19orf30, MIR7-3
cg03019000	TEX264
cg03167275	CXADR
cg03330058	ABTB1
cg03760483	ALOX12
cg04084157	VGF
cg04126866	C10orf99
cg04474832	ABHD14A, ABHD14B
cg04528819	KLF14
cg04836038	DOCK9
cg04999691	C7orf29, LRRC61
cg05250458	ZNF177
cg05294243	KLK13
cg05442902	MGC16703, P2RX6
cg06144905	PIPOX
cg06493994	SCGN
cg06993413	DPP8
cg07158339	FXN
cg07388493	NDUFS5
cg07408456	PGLYRP2
cg07730301	ALDH3B1
cg08030082	POMC
cg08186124	LZTFL1
cg08370996	NR2F2
cg08413469	DEPDC1
cg08965235	LTBP3
cg09118625	DIRAS3
cg09722555	CCL27
cg09809672	EDARADD
cg10523019	RHBDD1
cg10865119	C6orf122, C6orf208
cg11299964	MAPKAP1
cg12373771	CECR6
cg12830694	PPP1R14A
cg12941369	PDCD6IP
cg13302154	MGP
cg13460409	DSCR6

cg13836627	TJP1
cg13931228	MPP6
cg14060828	PTH2
cg14175438	FAM3C
cg14424579	AGBL5
cg14894144	LAMA3
cg15547534	C7orf47
cg15804973	MAP3K5
cg15988232	CSPG5
cg16408394	RXRA
cg16547529	KLHL35
cg16744741	PRKG2
cg17063929	NOX4
cg17274064	ERG
cg17338403	SLCO3A1
cg17589341	SLC14A1
cg17655614	CDH1
cg17729667	NINL
cg18328933	ABHD14A, ABHD14B
cg18440048	ZNF70
cg18573383	KCNC2
cg19724470	CD274
cg19761273	CSNK1D
cg20692569	FZD9
cg20761322	CIB2
cg21460081	HOXB4
cg21801378	BRUNOL6
cg22197830	TXNDC15
cg22449114	TCF15
cg22736354	NHLRC1
cg22809047	RPL31
cg22947000	BCMO1
cg23124451	CBX7
cg23517605	TUBB2B
cg23941599	FEM1C
cg24058132	GALC
cg24081819	EPHX2
cg25101936	ZBTB16
cg25148589	GRIA2
cg25505610	EIF3M
cg25564800	KPNA1
cg25771195	C16orf80
cg25781123	THUMPD3
cg25809905	ITGA2B
cg26372517	TFAP2E
cg26394940	C22orf26, LOC150381
cg26614073	SCAP
cg27015931	C16orf65
cg27169020	BNC1
cg27544190	C21orf63

References

[1] S. Pal and J. K. Tyler, "Epigenetics and aging," *Science Advances*, vol. 2, no. 7, 2016.

[2] R. Feil and M. F. Fraga, "Epigenetics and the environment: emerging patterns and implications," *Trends in Biotechnology*, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168952507001862

[3] L. D. Moore, T. Le, and G. Fan, "Dna methylation and its basic function," *Neuropsychopharmacology*, vol. 38, no. 1, pp. 23–38, 2013.

[4] A. E. Field, N. A. Robertson, T. Wang, A. Havas, T. Ideker, and P. D. Adams, "Dna methylation clocks in aging: Categories, causes, and consequences," *Molecular Cell*, vol. 71, no. 6, pp. 882–895, Sep. 2018.

[5] S. Horvath, "Dna methylation age of human tissues and cell types," *Genome Biology*, vol. 14, no. 10, p. R115, 2013, erratum in: Genome Biol. 2015 May 13;16:96. doi: 10.1186/s13059-015-0649-6.

[6] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J. B. Fan, Y. Gao, R. Deconde, M. Chen, I. Rajapakse, S. Friend, T. Ideker, and K. Zhang, "Genome-wide methylation profiles reveal quantitative views of human aging rates," *Molecular Cell*, vol. 49, no. 2, pp. 359–367, Jan. 2013.

[7] F. Galkin, P. Mamoshina, K. Kochetov, D. Sidorenko, and A. Zhavoronkov, "Deepmage: A methylation aging clock developed with deep learning," *Aging and Disease*, vol. 12, no. 5, pp. 1252–1262, Aug. 2021.

[8] L. P. de Lima Camillo, L. R. Lapierre, and R. Singh, "A pan-tissue dna-methylation epigenetic clock based on deep learning," *NPJ Aging*, vol. 8, no. 4, 2022. [Online]. Available: https://doi.org/10.1038/s41514-022-00085-y

[9] NCBI Gene Expression Omnibus, "Geo: Gene expression omnibus," https://www.ncbi.nlm.nih.gov/geo/, accessed: Jun. 6, 2025.

[10] ArrayExpress Database, "Arrayexpress - functional genomics data," https://www.ebi.ac.uk/biostudies/arrayexpress, accessed: Jun. 6, 2025.

[11] The Cancer Genome Atlas Program, "The cancer genome atlas (tcga)," https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga, accessed: Jun. 6, 2025.

[12] IBM. (2024) What is lasso regression? IBM. [Online]. Available: https://www.ibm.com/think/topics/lasso-regression#:~:text=Lasso%20regression%20is%20a%20regularization,regularization%20for%20linear%20regression%20models.

[13] e. a. Nicholas Pudjihartono, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, p. 927312, 2022.

[14] E. K. Jacob Murel. (2023) What is ridge regression? IBM. [Online]. Available: https://www.ibm.com/think/topics/ridge-regression

[15] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[16] T. Fushiki, "Estimation of prediction error by using k-fold cross-validation," *Statistics and Computing*, vol. 21, no. 2, pp. 137–146, 2011.

[17] IBM, "XGBoost," https://www.ibm.com/think/topics/xgboost, accessed: 2025-06-14.

[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794.

[19] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[20] S. Lundberg, "An introduction to explainable ai with shapley values," https://shap.readthedocs.io/en/stable/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html, 2023.

[21] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Research*, vol. 44, no. W1, pp. W90–W97, Jul 2016.

[22] National Center for Biotechnology Information, "Tcf15 transcription factor 15 [homo sapiens (human)]," https://www.ncbi.nlm.nih.gov/gene/6939, 2025.

[23] ——, "Hoxb4 homeobox b4 [homo sapiens (human)]," https://www.ncbi.nlm.nih.gov/gene/3214, 2025.

[24] ——, "Ltbp3 latent transforming growth factor beta binding protein 3 [homo sapiens (human)]," https://www.ncbi.nlm.nih.gov/gene/4054, 2025.

[25] National Institutes of Health, "Stem cell basics," https://stemcells.nih.gov/info/basics/stc-basics, 2023, accessed: 2025-06-21.

[26] A. S. Ahmed, M. H. Sheng, S. Wasnik, D. J. Baylink, and K. W. Lau, "Effect of aging on stem cells," *World Journal of Experimental Medicine*, vol. 7, no. 1, pp. 1–10, 2017.

[27] H. Wu and Y. Sun, "Epigenetic regulation of stem cell differentiation," *Pediatric Research*, vol. 59, no. Suppl 4, pp. 21–25, 2006.

[28] P. Du, X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin, "Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis," *BMC Bioinformatics*, vol. 11, p. 587, 2010. [Online]. Available: https://doi.org/10.1186/1471-2105-11-587