

Development of an Accurate Biological Aging Clock Using Machine Learning Models and DNA Methylation Patterns in Skeletal Muscle

RISHAB PERATI

MONTA VISTA HIGH SCHOOL, 21840 McClellan Rd, Cupertino, CA, 95014. 21840
McClellan Rd, Cupertino, CA 95014.

ABSTRACT:

Objective: A leading challenge in aging research is measuring age accurately, as monitoring healthy individuals over decades to assess the effects of interventions for the aging process is time and cost-prohibitive. The goal of this work was to develop an accurate epigenetic skeletal muscle-based aging clock using high-dimensional DNA methylation patterns to calculate chronological age.

Method: Genome-wide DNA methylation values from skeletal muscle tissue samples from 47 individuals with 200,000 methylation sites (CpGs) per sample were used for the study. Feature selection was used to narrow down CpGs. An Elastic Net regression model with a Leave-One-Sample-Out Cross-Validation (LO-SAMPLE-OCV) was used to evaluate feature relevance and model performance. I then implemented an ensemble approach where predictions from the LO-SAMPLE-OCV process served as inputs to another model using ElasticNet regression. This two-step approach leverages the strengths of ElasticNet for both feature selection and meta-learning, aiming to enhance predictive accuracy while also capturing the complex relationships of high-dimensional data.

Results & Conclusion: The model outperformed the Voisin et. al. method against the performance metrics of median absolute age differences (6.5 years using the Voisin method versus 1.9 years and 2.31 years using the LO-SAMPLE-OCV and ensemble methods respectively) and mean age differences (1.9 years using the Voisin method versus 0.26 years and -0.17 years using the LO-SAMPLE-OCV and ensemble methods respectively).

Impact: This tool can be used to accurately predict and understand aging to develop interventions for the prevention, early detection, diagnosis, and treatment of aging-related diseases, including cancer.

KEYWORDS: *Epigenetic Aging Clock, DNA Methylation Patterns, Aging-Related Diseases, Age Prediction Models, Chronological Age Calculations.*

1. INTRODUCTION

The human genome consists of 3.2 billion base pairs, of which we have 30 million methylation sites.¹ Methylation is a universal biochemical process that covalently adds methyl groups to a variety of molecular targets, including neurotransmitters, lipids, proteins, and DNA.² DNA methylation is essential for normal development, and plays an important role in key processes including genomic imprinting, X-chromosome inactivation, and suppression of repetitive element transcription and transposition. DNA methylation, when dysregulated, contributes to many diseases including cancer.³ Understanding aging is important to develop new interventions

for the prevention, early detection, diagnosis, and treatment of aging-related diseases, disorders, and disabilities. To develop appropriate interventions, understanding the causes and the factors that place people at increased risk of age-related diseases is essential. One of the leading challenges in the field of aging research is measuring age accurately. Monitoring healthy individuals for decades to assess whether an intervention affects aging is time and cost-prohibitive. As aging is strongly correlated with changes in DNA methylation and methylation changes across all organs at different speeds, the creation of epigenetic clock age predictors that use DNA methylation data has given researchers tools to measure the aging process quantitatively.⁴ Epigenetic clocks are mathematically derived age estimators that are based on combinations of methylation values that change with age at specific CpGs (CpG sites are the regions in the DNA sequence, which comprises a cytosine followed by guanine from 5' to 3' direction) in the genome.⁵ Epigenetic clocks have the potential of not only measuring aging but also guiding epigenetic interventions.⁶ Two of the most well-known age predictors are the ones developed by Hannum et al. and Horvath in 2013. Hannum et al. developed a blood-based epigenetic clock. Horvath then developed a more accurate epigenetic clock to predict age across tissues, developing a predictor with 353 CpG sites.⁷ Horvath's original pan-tissue epigenetic clock excluded skeletal muscle samples and exhibited a strong deviation between DNA methylation and chronological age.

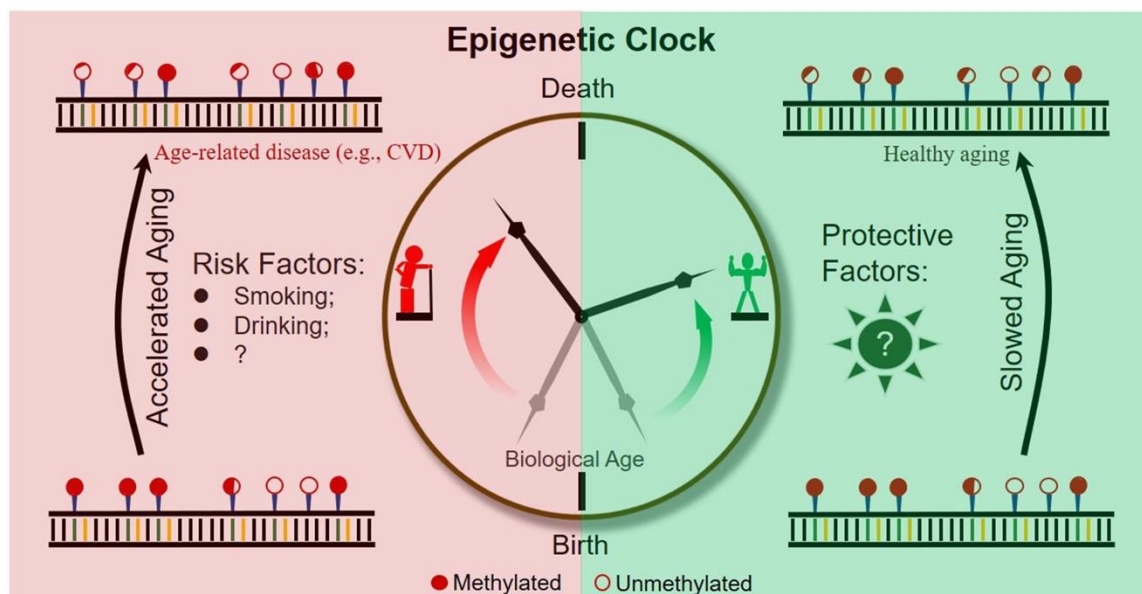


Figure 1: Epigenetic Clock Across a Human Lifetime⁹

As the basic function of muscle is to generate energy, and as life depends on energy generation, it is important to include the rate of DNA methylation in the muscle when calculating chronological age and its impact on disease progression.⁸ Voisin et.al used a leave-one-out-cross-validation (LO-SAMPLE-OCV) approach to create a muscle-based epigenetic clock using over 17,000 CpG sites from 650 samples (across 8 data sources) but it lacked the desired precision.

My study aims to develop a more accurate muscle-based epigenetic aging clock to calculate chronological age.

2. METHODOLOGY AND FRAMEWORK

Human skeletal muscle DNA methylation data from the open-access Gene Expression Omnibus (GEO) platform was used for the study. The dataset from this platform will have data from the Genotype-Tissue Expression (GTEx) project database. This dataset's epigenome-wide DNA methylation values contain skeletal muscle tissue samples from 47 individuals with 200,000 methylation sites (CpGs) per sample. The samples were from individuals in 5 different age ranges, 6.5% individuals in the 30-39 age group, 23% in the 40-49 age group, 17% in the 50-59 age group, 45% in the 60-69 age group and 8.5% in the 70-79 age group. To account for our high-dimensional, low-sample-size data, I implemented a custom technique of Elastic Net models trained on randomized feature subspaces. This method is similar to Random Forest's split-variable randomization techniques. Finally, I aggregated model predictions from each subspace in our ensemble approach. The code was entirely done in Python, using open-source machine-learning libraries including Pandas, NumPy, and Scikit-Learn. I also used Seaborn, and Matplotlib for data visualization. A split of 60% train data and 40% test data was used for the models. The models were taught using the train data, applied to the test data, and modified to achieve the highest accuracy. The work in this study was an extension of the Voisin et.al leave-one-out-cross-validation (LO-SAMPLE-OCV) approach.

2.1 Preprocessing of the Data

The data was preprocessed to remove missing or inconsistent data values resulting from human or computer error, which can improve the accuracy and quality of a dataset, making it more reliable. Standard Scaler was used to standardize the input data to ensure that the data points had a balanced scale, which is crucial for machine learning algorithms, especially those that are sensitive to differences in feature scales.

2.2 Randomized Feature Selection

Random samples of 20 subsets each consisting of 10,000 CpGs, totaling 200,000 CpGs were used. The following methods were independently applied to each subset until the ensemble step, where they were aggregated. CpG sample until aggregated in the

2.3 Feature selection

Feature selection techniques were employed to reduce the initial dataset of 200,000 CpG sites to those indicative of age. Specifically, an Elastic Net regression model ($L1_ratio = 0.5$, $\alpha = 0.01$) was utilized, which selected approximately 233 CpG sites per 10,000 CpGs.

2.4 Leave-one-sample-out cross-validation

The dataset comprised 47 skeletal muscle samples and 200,000 CpG sites. A leave-one-sample-out cross-validation approach was employed, where 46 samples were used for training and the remaining sample for testing, across all samples. The Elastic Net regression model was configured with both the L1_ratio and α parameters set to 0.01.

2.5 Ensemble Method on the DNA Methylation Sites

Median age predictions and CpG sample age predictions were used as input features for the ensemble method. A split of 60% train data and 40% test data was used for the model. The model was taught using the train data and then applied to the test data to achieve the highest accuracy

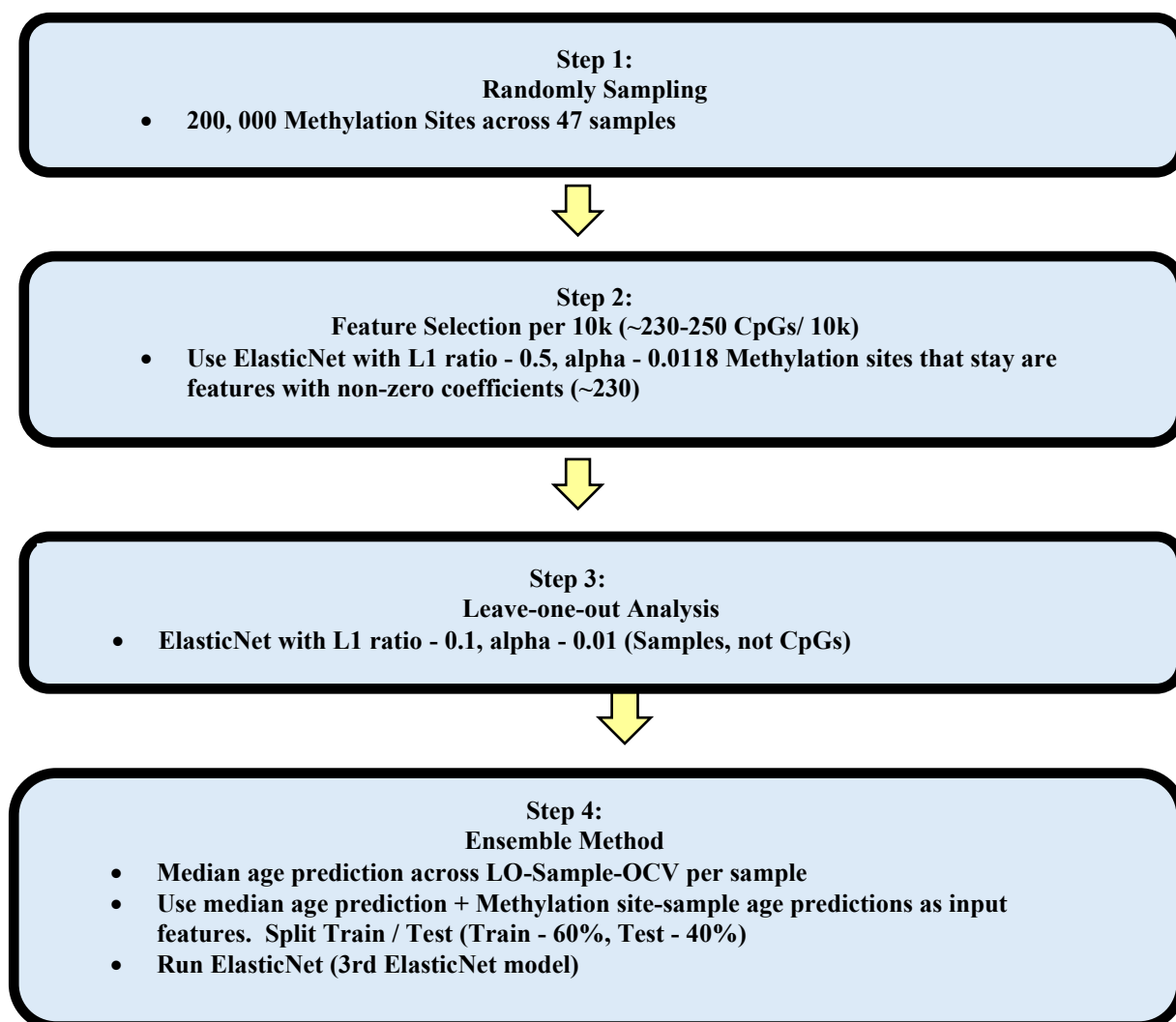


Figure 2: Flow chart of the methodology

3. RESULTS

3.1 Leave-one-out cross-validation

In the Leave-One-Out Cross-Validation (LO-SAMPLE-OCV) procedure, one dataset is designated as the test set, while the remaining datasets are used as training sets to develop the age predictor. The CpG profiles from the training sets were input into an elastic net regression model to estimate the age of the test set. The predicted ages were then compared with the actual ages by calculating the correlation coefficient

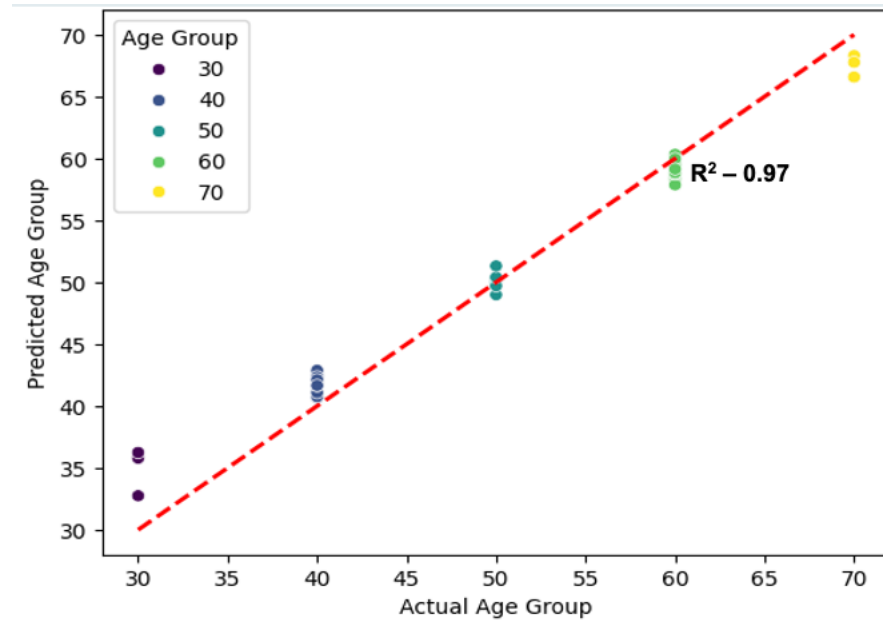


Figure 3: Linear Regression for the Leave-one-out cross-validation

Figure 3 shows that the 30s and 70s age groups show worse prediction accuracy than the 40s, 50s, and 60s age groups. The correlation coefficient of the LO-SAMPLE-OCV, linear regression was 0.97.

Prediction error was calculated as age acceleration (AA): the difference, between predicted and actual age (AAdiff) and the residual from a linear regression of predicted age against actual age (AAresid). AAdiff is sensitive to the mean age of the dataset and the pre-processing of the DNA methylation dataset; AAdiff can be biased depending on how the dataset was normalized and depending on the mean age and age variance of the dataset. In contrast, AAresid is insensitive to the mean age of the dataset and is robust against different pre-processing methods.



Figure 4: Heat map showing the difference in median age predicted per age group in years using the Leave-one-out cross-validation method

Figure 4 shows that the median age prediction difference using the LO-SAMPLE-OCV model for the 30s age group is 5.79 years away from the actual age and the median age prediction difference for the 70s age groups is 2.8 years away from the actual age. This is because the least number of data points were available for these age groups.

3.2 Ensemble Method

The dataset was divided into two parts: a training set and a test set. In this case, a split of 60% train data and 40% test data was used. The features used for training the model include the median age prediction obtained from the LO-SAMPLE-OCV, along with methylation site-sample age predictions. These features are utilized to capture different aspects of age predictions. The chosen model for this task was ElasticNet, specifically the third ElasticNet model. ElasticNet is a linear regression model that combines L1 and L2 regularization techniques. Once the ElasticNet model is trained, it is applied to the test data to make predictions. The predicted age values are compared with the actual ages in the test set to assess the model's performance. Overall, this process aims to utilize ensemble methods along with ElasticNet regression to predict age based on median age predictions and methylation site-sample age predictions, with a focus on achieving high accuracy on the test data while avoiding overfitting.

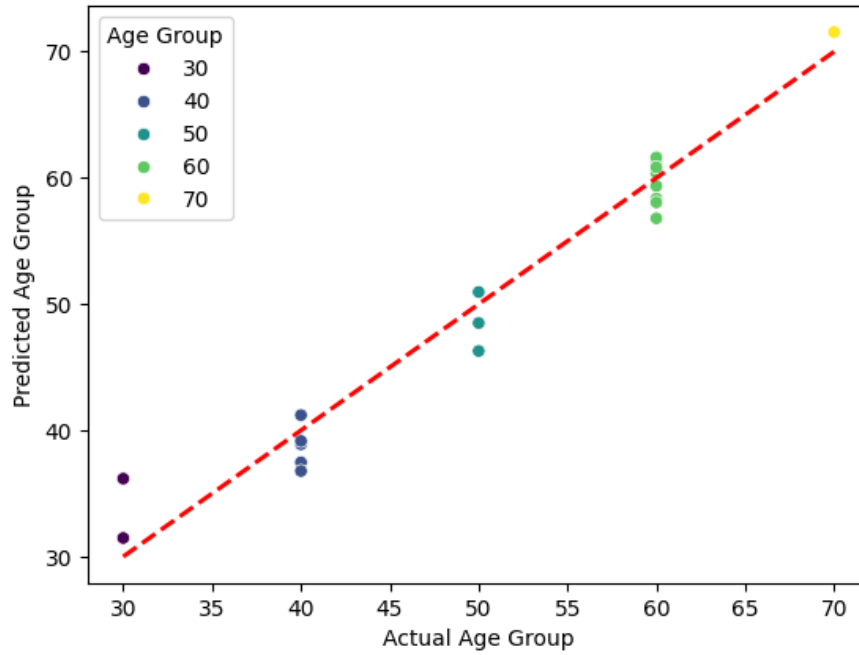


Figure 5: Linear Regression Analysis for the Ensemble Method

Figure 5 shows that the 30s and 70s age groups show worse prediction accuracy than the 40s, 50s, and 60s age groups using the ensemble method as well. The correlation coefficient of the Ensemble method was 0.96.

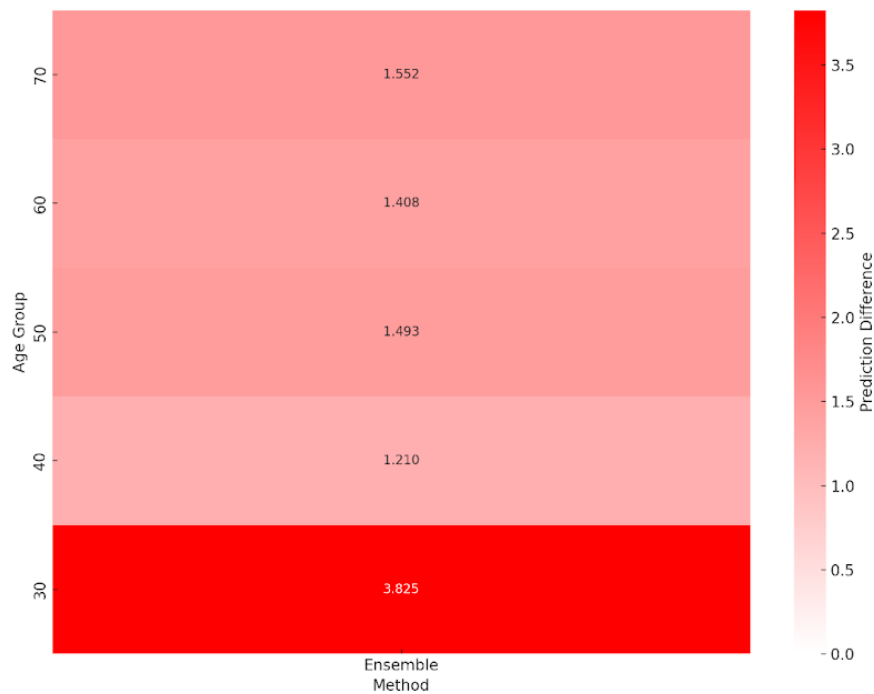


Figure 6: Heat map showing the difference in median age predicted per age group in years using the Ensemble method

Figure 6 shows that the median age prediction difference using the Ensemble model for the 30's age group is 3.8 years away from the actual age and the median age prediction difference for the 70s age groups is 1.8 years away from the actual age.

Method	Correlation coefficient (R ²)	Median Absolute Difference (Median AAdiff in Yrs)	Mean Difference (Mean AAdiff in Yrs)	Mean Squared Error	Mean Absolute Error
Voisin et al.	0.90	6.5 Yrs	-1.9 Yrs	N/A	N/A
Leave-one-out cross validation	0.97	1.9 Yrs	0.26 Yrs	3.88	3.81
Ensemble Method	0.96	2.31 Yrs	-0.17 Yrs	5.36	1.88

Table 1: Comparison of Voisin's Method, the Leave-one-out Cross-Validation Method, and the Ensemble Method

The Leave-One-Out Cross-Validation method outperforms the Voisin et al. method for all metrics, showing higher correlation coefficients, and smaller median and mean absolute differences. This indicates that the Leave-One-Out Cross-Validation method provides more accurate age predictions.

The Ensemble Method also shows better performance than the Voisin et al. method across most metrics, with a slightly lower correlation coefficient but still higher than 0.90. It also has a smaller median absolute difference and mean absolute error, indicating improved accuracy in age prediction compared to the Voisin et al. Method.

However, when comparing the LO-Sample-OCV method and the Ensemble Method, the former performs slightly better in terms of correlation coefficient and mean squared error, while the latter has a lower median absolute difference and mean absolute error. The choice between these two methods might depend on the specific priorities of the application, such as whether minimizing errors or maximizing correlation is more important.

4. CONCLUSION

- For individuals in the 30s age group, the ensemble model has a median prediction difference of 3.8 years. This suggests that, on average, the model's predictions for this age group were 3.8 years away from the actual age. The model was least accurate for this age group compared to the others.

- The ensemble model performed similarly for predicting individuals between the ages of 40 and 70. The model has a median prediction difference of approximately 1.5 years (or less).
- Even though the LO-SAMPLE-OCV method had a slightly higher R^2 value and lower Mean Squared Error value than the ensemble method, it was not as consistent and suggests that the data may have been underfitted.
- LO-SAMPLE-OCV model on average predicted within 1.95 years of the actual.
- The ensemble model on average predicted within 2.31 years of the actual.
- This new clock significantly outperforms the previous muscle clock developed by Voisin et al. against the performance metrics of median absolute age differences (6.5 years using the Voisin method versus 1.9 years and 2.31 years using the LO-SAMPLE-OCV and ensemble methods respectively) and mean age differences (1.9 years using the Voisin method versus 0.26 years and -0.17 years using the LO-SAMPLE-OCV and ensemble methods respectively).

5. IMPACT

- This tool can be used to accurately predict and understand aging to develop interventions (environmental factors, exercise, diet, etc.) for the prevention, early detection, diagnosis, and treatment of aging-related diseases, disorders, and disabilities.
- Understanding epigenetic aging is important as people with epigenetic age greater than their chronological age are at higher risk of age-related diseases, including cancer.
- The muscle clock developed in this study can be applied to genome-wide methylation patterns to study genetic variations related to aging and an increased risk of cancer. The muscle clock can be used to search the genome of individuals to predict their cancer risk and allow for early interventions and screening.

6. FUTURE WORK

- Develop this method into a Python library like the Muscle Epigenetic Age Test (MEAT) package in R.
- Apply the method to other datasets with larger sample sizes to check for the reproducibility of the model.
- Use additional datasets for a leave-one-out dataset analysis approach.

- Implement and compare the beta-mixture quantile normalization method for preprocessing the data and see if that improves the accuracy of the model.

7. REFERENCES

1. Adiv, A., Johnson, K., Alkman, Stuart, R.G., Calimport, Daniel, Wulttke, Alexandra Stolzing, and Joao, R.D., Magalhaes. The Role of DNA Methylation in Aging, Rejuvenation, and Age-Related Disease. *Rejuvenation Res*; 15 (5):483-94 (2012).
2. Zemach A. McDaniel IE. Silva P. Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*; 328:916–919 (2010).
3. Voisin S., Harvey N., Haupt L. An epigenetic clock for human skeletal muscle. *Journal of Cachexia, Sarcopenia and Muscle*; 11:887-898 (2020).
4. Benayoun BA, Pollina EA, Brunet A. Epigenetic regulation of aging: linking environmental inputs to genomic stability. *Nat Rev Mol Cell Biol*; 16:593–610 (2015).
5. Camillo L, Lapierre L, Singh R, A pan-tissue DNA-methylation epigenetic clock based on deep learning. *NPJ Aging*; 8: Article Number 4 (2022).
6. Horvath S, DNA methylation age of human tissues and cell types. *Genome Biol*; 14:3156 (2013).
7. Zykovich A, Hubbard A, Flynn JM, Tarnopolsky M, Fraga MF, Kerksick C, et al. Genome-wide DNA methylation changes with age in disease-free human skeletal muscle aging. *Cell*;13: 360–366 (2014).
8. Sehl ME, Henry JE, Storniolo AM, Ganz PA, Horvath S. DNA methylation age is elevated in breast tissue of healthy women. *Breast Cancer Res Treat*;164: 209–219 (2017).
9. Sailani M.R., Hailing J.F., Henrik D.M., Lee H., Plomgaard P., Pilegaard H., Snyder M.P., Regenberg B. Lifelong physical activity is associated with promoter hypomethylation of genes involved in metabolism, myogenesis, contractile properties and oxidative stress resistance in aged human skeletal muscle. *Scientific Reports*; 9: 3272 (2019).