Machine Learning and Statistical Analysis of Biological Data to Understand Mechanisms of Aging

Author: Klára Hirmanová (k.hirmanova@student.tudelft.nl)
Responsible Professor: Prof. dr. ir. Marcel Reinders
Supervisors: Bram Pronk, Inez den Hond, Gerard Bouland

BACKGROUND

- .Aging is a biological process characterized by gradual functional deterioration that ultimately compromises an organism's survival.
- Epigenetic alterations are considered a primary hallmark of aging and include DNA methylation, histone modification, and chromatin remodeling [1]
- DNA methylation levels at specific CpG sites are quantified using beta values, providing a normalized measure of methylation intensity
- .Aging clocks use machine learning to predict biological age from features like DNA methylation
- Biological age - age predicted from aging clock
- Age acceleration - discrepancy between biological and chronological age. Can indicate accelerated or resilient aging
- **Horvath2013** (the og)- first multi-tissue epigenetic aging clock, uses 353 CpG sites and Elastic Net regression [2]
- **AltumAge** (2022) - a deep learning multi-tissue epigenetic clock, improves prediction accuracy by modeling non-linear relationships between CpG sites, trained on more than 21,000 features [3]

RESEARCH QUESTION

Can we reproduce state-of-the-art age predictors based on epigenetic modification data and understand the model features specifically related to those predictions that differ significantly from the true age?

To answer this question, we created a set of sub-questions:

- Can we reproduce performance metrics of epigenetic age prediction models?
- Which methods can determine CpG feature importance and select key predictors for age estimation?
- How well do different classification algorithms detect age acceleration using selected CpG features?
- Which features drive acceleration predictions, and do they differ across age groups?

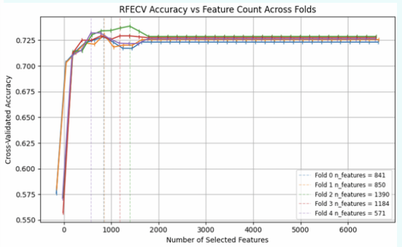
[1] Carlos Lo´ pez-Ot´ ın, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. Hallmarks of aging: An expanding universe. *Cell*, 186(2):243–278, 2023.
[2] S. Horvath. DNA methylation age of human tissues and cell types. *Genome Biol.*, 14(10), 2013.
[3] L. R. Lapierre L. P. de Lima Camillo and R. Singh. A pan-tissue DNA-methylation epigenetic clock based on deep learning. NPJ Aging, 8(1), 2022.

METHODOLOGY

- Data Acquisition
  - Collect DNA methylation data from GEO, ArrayExpress, and TCGA
- Preprocessing & Age Estimation
  - Aggregate probes and apply Horvath2013 and AltumAge clocks to predict biological age
- Residual Analysis
  - Compute age acceleration residuals (predicted age – chronological age) and apply residual thresholding
- Feature Selection
  - Apply variance filtering, RFECV
- Model Training & Interpretation
  - Use stratified 5-fold cross-validation to train and tune Logistic Regression, Random Forest and XGBoost classifiers
  - Perform SHAP analysis to identify most predictive features in positive and negative acceleration.
- Age groups Comparison
  - Identify distinct predictive CpG features across different age groups.

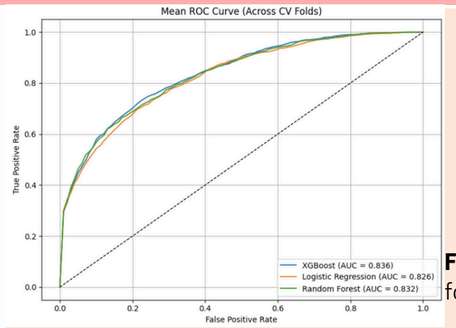
RESULTS

- .Horvath2013 was perfectly replicated across all metrics (MAE, MSE, R, median error = 1.000), confirming full reproducibility, AltumAge replication showed high but not perfect agreement, with normalized scores ranging from 0.632 to 0.993.
- Samples with residuals <1 year were excluded, leaving 4,119 samples.
- Variance filtering removed 14,888 low-variance CpGs, reducing to 6,481.



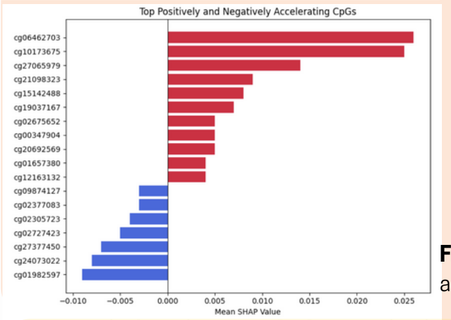
- RFECV reduced dimensionality, optimal feature counts ranged from 451 to 1,172 across folds

Figure 1: RFECV accuracy across five cross-validation folds, plotted against the number of selected features. Each colored line represents the accuracy curve for one fold.



- .XGBoost outperformed Logistic Regression and Random Forest, achieving the highest AUC (0.836) and F1 score (0.753).
- Random Forest achieved comparable performance with AUC score 0.832.

Figure 2: Mean ROC AUC curves from nested 5-fold cross- validation.



- .A stable subset of 312 CpGs was identified by intersecting SHAP and stability selection results.
- SHAP analysis highlighted 11 top positively and 7 top negatively contributing CpGs.

Figure 3: Bar plot of mean SHAP values for top accelerating CpG sites

- Positively Accelerating CpGs enriched in immune signaling , neurodevelopmental suppression, and prenatal growth pathways-inflammaging, mitochondrial stress.
- Negatively Accelerating CpGs associated with DNA repair, neurogenesis, and antiviral defense-genomic stability and cognitive resilience.
- Age-Associated Positively Shifting CpGs initially suppressed in young, these features increase with age and relate to oxidative stress response and amyloid-beta metabolism-transition from homeostasis to stress compensation.

CONCLUSION

- The study relied on preprocessed methylation data and a single clock's residuals.
- Biological age acceleration modeled as a binary classification, which may oversimplify the complexity of aging signals.
- XGBoost performed better than Logistic Regression and Random Forest.
- SHAP analysis and GO enrichment revealed biologically meaningful CpG features linked to immune response, DNA repair, and oxidative stress.
- This project offers a reproducible pipeline and novel insights into the biological mechanisms behind discrepancies in predicted biological age in healthy samples.
- Age prediction and may be in the future used in practical setting, interpreting their residuals will be key to understand why we age the way we do.