Predicting Age based on Bio-measurements

Peyton Ardoin, Khalil El-abbassi, Aditya Baisakh, Marilin Guerrero Laos, Tre' Henson, Matthew Lemoine, Austin Louque, Efren Mesino Espinosa, Aduragbemi Oyerinde, Shalini Shalini, and Gowri Priya Sunkara

Department of Mathematics, Louisiana State University, Baton Rouge, Louisiana
Faculty Advisors: Dr. Nadejda Drenska and Dr. Peter R. Wolenski

Introduction

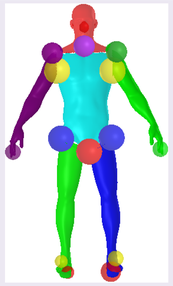

Figure 1. Example of an avatar from which body measurements are obtained.

- In previous projects, the Math Consultation Clinic has predicted anthropometric and body-composition values from biomarkers. These biomarkers include height, weight, waist circumference, and left and right thigh circumference.
- In this project, we are predicting the age of a given patient. Knowing a patient's biological age can provide a great insight into their relative health, when compared to other patients with similar chronological age (true age).
- In the first half of the summer, we worked with a dataset provided by the Pennington Biomedical Research Center. In the second half of the summer, we worked with a larger dataset that we call the Caesar Dataset.
- The main difference between these two datasets lies in the size of the datasets. The Caesar Dataset has more patients and more measurements than the Pennington Dataset. This gives better results and can provide more insight into which measurements are most correlated with age.

Correlation- and SHAP-Weighted RBF Models

Three kernel regression methods were evaluated for age prediction:

- **Support Vector Regression (SVR):** fits a nonlinear regression function using support vectors and an ϵ-insensitive loss.
- **Kernel Ridge Regression (KRR):** combines ridge regularization with a nonlinear kernel to control overfitting.
- **Gaussian Process Regression (GPR):** uses a kernel-defined probability distribution over possible regression functions.

All three methods use the Radial Basis Function kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \sum_{m=1}^{p} w_m (x_{im} - x_{jm})^2\right),$$

where w_m is the importance assigned to biomarker m.

Correlation and SHAP Feature Weights

Pearson Correlation
Measures the linear relationship between each biomarker and age:

$$r_m = \frac{\sum_{i=1}^{n}(x_{im} - \bar{x}_m)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{im} - \bar{x}_m)^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

The correlation weights satisfy

$$w_m^{\mathrm{Corr}} \propto |r_m|.$$

SHAP measures each biomarker's contribution to the prediction:

$$f(\mathbf{x}) = \phi_0 + \sum_{m=1}^{p} \phi_m.$$

Global SHAP importance is

$$I_m = \frac{1}{N}\sum_{i=1}^{N} |\phi_{im}|, \qquad w_m^{\mathrm{SHAP}} \propto I_m.$$

Feature-Importance Comparison

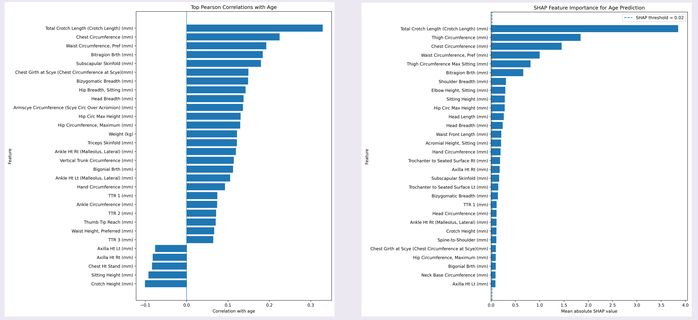

Figure 2. Pearson correlation coefficients (left) and mean absolute SHAP importance scores (right). Correlation measures direct linear biomarker–age relationships, whereas SHAP captures nonlinear effects and feature interactions.

RMSE Comparison of Weighted-RBF Regression Models

Table 1. Root Mean Square Error (RMSE, years) of the six regression models across a single global group. The models compare correlation-weighted (Corr) and SHAP-weighted (SHAP) radial basis function (RBF) kernels for Support Vector Regression (SVR), Kernel Ridge Regression (KRR), and Gaussian Process Regression (GPR). The lowest RMSE within each age group is shown in bold.

Age (Sample Size)	Corr-SVR	SHAP-SVR	Corr-KRR	SHAP-KRR	Corr-GPR	SHAP-GPR
Male (1029)	8.25	**7.80**	8.18	7.84	11.48	11.48
Female (1135)	8.87	8.66	8.87	**8.53**	12.07	12.07
Combined (2164)	8.51	8.45	8.34	8.12	**7.98**	7.99

Least-Squares Support Vector Regression

In addition to the above models, we used the Least-Squares Support Vector Regression to predict age. We used 5-fold cross validation for each of the following datasets. When compared to the Pennington Data from the first half of the summer, we can see that the error has gone down. This is due to the number of entries in the Caesar dataset compared to the Pennington Dataset and the additional measurements taken in the Caesar Dataset.

Input	Results (years)
Caesar (Female)	8.37
Caesar (Male)	8.06
Caesar (whole)	8.20
Caesar (whole; no gender)	8.21
Whole Dataset (Pennington)	11.73

Table 2. Results for the LSSVR with the Caesar dataset compared to the Pennington Dataset.

Random Forest

- Random Forest is an ensemble learning method that combines multiple decision trees, where each tree learns decision rules by selecting the best feature splits.
- Hyperparameters, such as the number of trees and maximum depth, are optimized using random search, and the model is trained on 80% of the data and evaluated on the remaining 20%.
- The trained model is combined with K-Means clustering, TreeSHAP feature importance, and centile deviation analysis to group subjects, interpret predictions, and identify atypical cases.

Below are the results obtained from testing Random Forest.

Model	Error (yrs)
11 age brackets	11.6
Tuned hyper-parameters	10.7
Bracketed markers	13.3
Bracketed markers + interactions	12.42
Single model, TreeSHAP + K-means	**9.09**

Table 3. Results from RF tests.

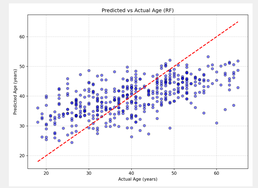

Figure 3. RF with K-Means on single age bracket: RMSE = 9.09 years

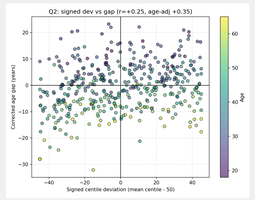

Figure 4. Signed deviation Vs. Age gap

Moving Forward

- Develop a Graphical User Interface (GUI) that allows users to input OBJ 3d scans and receive anthropometric measurements and predicted biological age.
- Compare the proposed GUI with existing biological age prediction tools developed by other research institutions[1].

Acknowledgements

- We would like to thank Dr. Steven B. Heymsfield from Pennington Biomedical Research Center for giving us the datasets and supporting the evaluation of the models.
- We would like to thank Prof. Peter Wolenski and Dr. Nadejda Drenska for their help with this project and their guidance.

References

[1] C. A. Fermín-Martínez, A. Márquez-Salinas, E. C. Guerra, L. Zavala-Romero, N. E. Antonio-Villa, L. Fernández-Chirino, E. Sandoval-Colin, D. A. Barquera-Guevara, A. Campos Muñoz, A. Vargas-Vázquez, C. D. Paz-Cabrera, D. Ramírez-García, L. M. Gutiérrez-Robledo, and O. Y. Bello-Chavolla, "Anthropoage, a novel approach to integrate body composition into the estimation of biological age," Aging Cell, vol. 22, no. 1, p. e13756, 2023.

[2] G. Agrahari, K. Bist, M. Pandey, J. Kapita, Z. James, J. Knox, S. Heymsfield, S. Ramirez, P. Wolenski, and N. Drenska, "Predicting anthropometric body composition variables using 3d optical imaging and machine learning," Frontiers in Bioinformatics, vol. 6, p. 1722578, 2026.

[3] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273–297, 1995.

[4] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," Neural Processing Letters, vol. 9, pp. 293–300, 1999.