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, Gowri Priya Sunkara.

August 18, 2026

1 Introduction

Anthropometric measurements provide information about the size, shape, and composition of the human body. Advances in three-dimensional body scanning and machine learning have created opportunities to use these measurements for predicting physiological and body-composition characteristics without relying exclusively on traditional clinical measurement techniques.

Previous projects conducted through the Math Consultation Clinic investigated the prediction of specific anthropometric and body-composition variables from measurable biomarkers. Traditionally, quantities of interest were obtained using methods such as dual-energy X-ray absorptiometry (DXA). Previous work demonstrated that machine-learning models could predict some of these quantities like Appendicular Lean Mass(ALM), Bone Mineral Density(BMD), and Body Fat Percentage (BFP) with promising accuracy using measurements obtained from patient body scans. [1].

In these projects, a three-dimensional scan of a patient's body is obtained and used to extract a collection of anthropometric measurements. These measurements characterize different aspects of body geometry and composition and can subsequently be used as predictor variables in machine-learning models.

The success of these earlier studies motivated the present investigation, which shifts the prediction target from individual anthropometric quantities to biological age. Chronological age represents the amount of time that has elapsed since an individual's birth, but individuals of the same chronological age can differ substantially in their physical characteristics. These differences motivate the concept of biological age, which attempts to describe an individual's physiological or physical aging status rather than simply the number of years lived.

In this project, biological age is investigated from the perspective of measurable body characteristics. The central objective is to determine whether anthropometric biomarkers extracted from body measurements contain sufficient information to predict a person's chronological age. The resulting model prediction is then interpreted as a body-shape-based estimate of biological age.

For example, if a 23-year-old individual exhibits anthropometric characteristics that cause a predictive model to estimate an age closer to 30 years, the difference between chronological and predicted age may indicate that the individual's body measurements resemble those typically observed in an older population. Conversely, a predicted age below chronological age would indicate measurements more similar to those of younger individuals.

The primary goal of this study is therefore to develop and compare statistical and machine-learning models capable of predicting age from anthropometric biomarkers. Particular attention is given to identifying which biomarkers contribute most strongly to age prediction, determining whether nonlinear models improve predictive performance, and investigating how model performance changes when the population is separated by age or sex.

To address these questions, two anthropometric datasets are considered: the Pennington dataset and the CAESAR dataset. A range of regression approachesincluding linear regression methods, Random Forest, XGBoost, Support Vector Regression, Least Squares Support Vector Regression, Kernel Ridge Regression, and Gaussian Process Regressionare evaluated. Feature-selection and dimensionality-reduction techniques, including correlation analysis, SHAP feature importance, Principal Component Analysis, and clustering, are also investigated.

2 Data Description and Cleaning

Two anthropometric datasets were used in this project: the Pennington (Penn) dataset and the CAESAR dataset. The datasets differ substantially in both sample size and the number of available anthropometric biomarkers. Therefore, they were processed and analyzed separately.

2.1 Pennington Dataset

The Pennington dataset contains data from 515 participants with 44 anthropometric biomarkers. These biomarkers describe different aspects of body size, shape, and composition and were used as predictor variables for chronological age.

The Pennington dataset served as the initial dataset for investigating whether anthropometric measurements could be used to estimate age. Several statistical and machine-learning approaches were evaluated on this dataset, including linear regression-based methods, Random Forest, and XGBoost.

2.2 CAESAR Dataset

The CAESAR dataset provides a substantially larger collection of anthropometric measurements. The original dataset contained approximately 120 biomarkers and related data fields for each participant. These variables describe both individual anatomical regions and overall body characteristics, including measurements such as torso volume, total body volume, circumferences, lengths, surface areas, and other body-shape measurements.

Before model development, variables that were not required for the age-prediction analysis were removed. Redundant variables were also excluded to reduce unnecessary duplication of information. Following this preprocessing, 94 usable anthropometric biomarkers were retained for model development.

2.3 Missing Data and Participant Selection

The CAESAR dataset was further examined for missing measurements because incomplete observations can interfere with the training and evaluation of machine-learning models. Participants with incomplete data across the required biomarkers were removed before model development.

After the data-cleaning procedure, 2,164 participants remained for the subsequent age-prediction analysis.

One variable required special consideration: Bust/Chest Circumference Under Bust (mm). A substantial number of observations were missing for this measurement because it was not collected for male participants. Therefore, the missing values in this variable were not treated in the same manner as unintentionally missing measurements.

Rather than discarding a large number of otherwise complete participants, different versions of the dataset were considered depending on whether this measurement was included or excluded. Sex-Specific Data Analysis

In addition to analyzing the combined CAESAR population, male and female participants were considered separately. This allowed the study to investigate whether sex-specific differences in anthropometric measurements affected the ability of the models to predict age.

The resulting analyses therefore considered three primary populations:

- Male participants,
- Female participants, and
- Combined male and female participants.

Where necessary, versions of these datasets were constructed with or without the Bust/Chest Circumference Under Bust measurement to avoid introducing missing-data problems into model training.

Separating the datasets in this manner also provided a framework for determining whether age-related anthropometric patterns were consistent across the entire population or whether prediction performance differed between male and female participants.

Overall, the Penn dataset provided an initial 515-participant, 44-biomarker dataset for model development, while the larger CAESAR dataset provided 2,164 usable participants and 94 biomarkers after preprocessing. The larger CAESAR dataset was subsequently used for more extensive model comparisons, feature-importance analyses, dimensionality reduction, and kernel-based regression methods.

3 Models on Pennington Dataset

Several regression and machine-learning models were applied to the Penn dataset for age prediction. The models considered in this part of the project were

- Linear Regression (LR),
- Ridge Regression (RR),
- Bayesian Ridge Regression (BRR),
- Random Forest (RF),
- XGBoost,
- Neural Network

3.1 Linear Regression Models

To establish a baseline for biological age prediction, three linear regression models were investigated: Linear Regression, Ridge Regression, and Bayesian Ridge Regression. These methods estimate biological age from body-shape biomarkers while differing in their treatment of correlated predictors and model complexity.

- Linear Regression (LR) seeks an optimal linear relationship between the input biomarkers and the chronological age of a patient.
- Ridge Regression (RR) extends linear regression by shrinking the regression coefficients to reduce overfitting, particularly when biomarkers are correlated. For example, measurements such as waist and narrow waist may be highly correlated. Ridge Regression reduces the influence of such correlated variables by shrinking their coefficients.
- Bayesian Ridge Regression (BRR) treats the regression coefficients as probability distributions. This allows uncertainty in the data to be incorporated into the regression model and can produce more stable predictions.

3.1.1 Feature Selection for Body Shape Biological Age

Feature selection was performed to reduce the number of biomarkers used for predicting Body Shape Biological Age (BSBA). The initial model contained 46 biomarkers. Ridge coefficient-based feature reduction was then used to identify a smaller collection of biomarkers.

This procedure reduced the feature set from 46 biomarkers to 21 biomarkers, while also improving the RMSE.

The final BSBA biomarkers were organized into two groups.

Morphometric biomarkers included:

- Waist and narrow waist measurements
- Thigh, calf, and leg measurements
- Inseam and outside leg length
- Arm length and arm volume
- Leg surface area

Body composition biomarkers included:

- Appendicular Lean Mass (ALM)
- Bone Mineral Density (BMD)
- Total Body Volume
- Torso Volume

Comparison of BSBA Models

The performance of Ridge Regression and Bayesian Ridge Regression was compared using both the original 46 biomarkers and the reduced set of 21 biomarkers.

Model	Biomarkers	Overall RMSE
Bayesian Ridge	46	13.278
Ridge	46	13.320
Bayesian Ridge	21	12.812
Ridge	21	12.867

Reducing the model from **46 biomarkers** to **21 biomarkers** improved overall RMSE and produced a simpler, more interpretable BSBA model. Among the models compared, Bayesian Ridge Regression using 21 biomarkers achieved the lowest overall RMSE of 12.812 years.

3.2 Random Forest

Random Forest is based on decision trees. At each node of a tree, a biomarker and a corresponding threshold are selected to reduce the impurity of the target. This procedure continues throughout the tree until the specified maximum depth is reached.

The primary Random Forest parameters considered were the number of estimators, maximum tree depth, and number of samples used per decision tree.

3.2.1 Random Forest Methodology

A random search algorithm was used to test different combinations of Random Forest parameters across a grid of values. The model was trained using 80% of the data.

The biomarkers were also separated into high-, medium-, and low-importance groups based on their importance values. Additional experiments considered interactions among biomarkers. The Random Forest experiments produced the following results:

Model	Error
Random Forest with 11 age brackets	11.6 years
Random Forest Tuned Hyper-parameters	10.7 years
Random Forest Bracketed markers	13.3 years
Random Forest Bracketed Markers with interactions	12.42 years

Table 1: Results from RF tests.

3.3 Random Forest Results

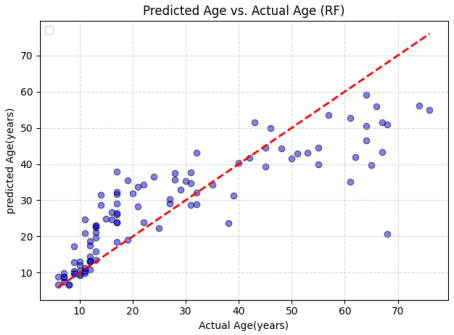


Figure 1: RF with tuned parameters and 11 age brackets

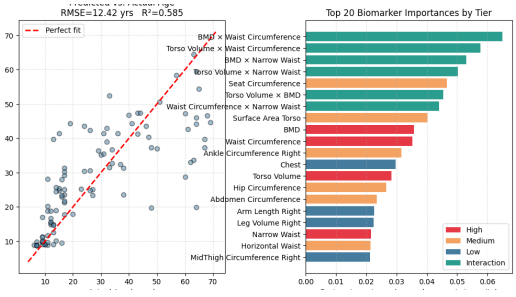


Figure 2: RF with tuned parameters and interacting tiered markers

Among these Random Forest experiments, the tuned hyperparameter model achieved the lowest error of 10.7 years.

3.3.1 Random Forest Limitations

Random Forest assigns importance values to biomarkers while training its decision trees. Therefore, imposing rigid importance values on the biomarkers can limit the model's ability to treat the measurements as continuous variables and construct a smooth model.

Across the Random Forest experiments, the RMSE values remained relatively high, exceeding 10 years. The model also consistently overestimated ages in approximately the 1040 year range and underestimated ages for participants approximately 45 years and older.

3.4 XGBoost

Extreme Gradient Boosting (XGBoost) is also an ensemble learning algorithm. In contrast to the bagging approach associated with Random Forest, XGBoost uses boosting, in which models are constructed sequentially.

The XGBoost model produced an overall RMSE of 12.88 years. The model performed best for the densely sampled 512 age group, where the RMSE was approximately 7.3 years. Larger prediction errors occurred at sparsely sampled age extremes, particularly among teenagers and older adults. The older-adult group contained only 13 patients and had an RMSE greater than 24 years. These sparsely sampled age groups contributed disproportionately to the overall RMSE.

4 Caesar Dataset

4.1 Models on Caesar Dataset

The models applied to the CAESAR dataset were:

- Multiple Linear Regression (MLR) with Principal Component Analysis (PCA)
- Support Vector Regression (SVR)
- Least Squares Support Vector Regression (LSSVR)
- Extreme Gradient Boosting (XGBoost)
- Comparison based on Groups with Kernel-based regression models

Additional analyses included correlation analysis, variance inflation factor analysis, SHAP feature importance, K-means clustering, and Random Forest.

4.2 Multiple Linear Regression and Principal Component Analysis

Previous models applied to the Pennington dataset used Ridge Regression and Bayesian Ridge Regression for feature selection in predicting Body Shape Biological Age. These methods were intended to reduce the number of biomarkers while improving prediction accuracy.

However, a single global regression model may not capture age-dependent relationships among biomarkers. Therefore, segmented Multiple Linear Regression with Principal Component Analysis was applied to the CAESAR dataset.

4.2.1 Methodology

The objective was to predict chronological age from anthropometric biomarkers using segmented Multiple Linear Regression, correlation analysis, and Principal Component Analysis.

The model pipeline consisted of the following steps:

- Use the cleaned anthropometric biomarker dataset.
- Perform correlation and variance inflation factor analysis to examine multicollinearity.
- Standardize the biomarker measurements.
- Apply PCA independently within each age group.
- Retain principal components explaining approximately 95% of the total variance.
- Train a separate Multiple Linear Regression model for each age group.
- Evaluate performance using MAE, RMSE, and R^2.

The complete procedure can be summarized as

$$\text{Biomarkers} \rightarrow \text{Age Segmentation} \rightarrow \text{PCA} \rightarrow \text{Multiple Linear Regression} \rightarrow \text{Predicted Age}.$$

4.2.2 Correlation and Variance Inflation Factor Analysis

Correlation analysis was used to examine the strength and direction of the relationships among variables. In regression models, high correlation among independent variables may lead to multicollinearity.

Variance inflation factor was used to determine how much multicollinearity increased the variance of the regression coefficients.

The highest finite-valued VIF features included:

- Volume,
- Total Surface Area,
- Torso Volume,
- Torso Surface Area,
- Right and left leg surface areas,
- Right and left leg volumes,
- Right and left arm surface areas,
- Right and left arm volumes,
- Thumb Tip Reach, and
- Right and left standing Infraorbitale Height.

4.2.3 Principal Component Analysis

The correlation and VIF analyses showed that several biomarkers measured similar anatomical characteristics. This introduced multicollinearity into the regression model.

Principal Component Analysis was used to transform the original biomarkers into orthogonal principal components. The biomarker measurements were first standardized, and only the components explaining approximately 95% of the total variance were retained.

A separate Multiple Linear Regression model was then trained using the retained principal components within each age group.

The procedure was

Biomarkers → PCA → Linearly Independent Principal Components → Segmented MLR Model.

4.2.4 Results

The segmented MLR with PCA produced the following results.

Age Group	Samples	No. of PCs	Variance	RMSE
0–20	58	17	95.36%	0.56
20–30	539	27	95.20%	2.69
30–40	612	26	95.01%	2.88
40–50	639	27	95.11%	2.87
50–60	396	28	95.05%	2.86
60–70	142	25	95.29%	1.85

The overall results were MAE = 2.331 years, RMSE = 2.746 years, MSE = 7.538 years^2, and $R^2 = 0.9486$.

PCA reduced the dimensionality of the biomarker space while retaining approximately 95% of the variance within each age group. The segmented MLR with PCA achieved an overall RMSE of approximately 2.75 years.

Overall Performance and Findings

MAE = 2.331 years
RMSE = 2.746 years
MSE = 7.538 years
R^2 = 0.9486

- PCA reduced the dimensionality of the biomarker space while retaining approximately 95% of the variance within each age group.
- This pipeline achieved an overall RMSE of 2.75 years and an R^2 of 0.9486, substantially improving upon the initial performance.

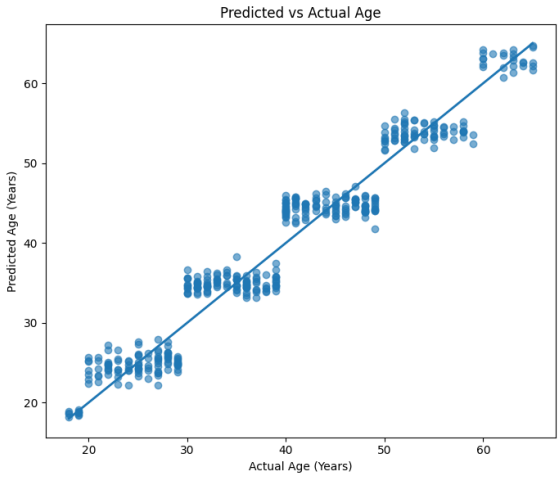


Figure 3: Segmented MLR w/ PCA Model; RMSE = 2.75 years

4.3 XGBoost

XGBoost was applied separately to male and female participants in the CAESAR dataset.

The presentation identifies three reasons for the improvement over the original global XGBoost RMSE of 12.88 years:

- The CAESAR dataset contained approximately four times as many participants.
- Separating male and female participants reduced physical differences unrelated to age.
- Hyperparameters were tuned using five-fold cross-validation.

The resulting RMSE values were:

Model	Error
XGBoost w/ tuned hyperparameters (Males)	**8.24 years**
XGBoost w/ tuned hyperparameters (Females)	**8.66 years**

Table 2: Results from XGBoost tests.

4.4 Support Vector Regression

Support Vector Regression predicts continuous values by fitting a function to the data while controlling model complexity.

SVR uses an (ϵ)-tube around the fitted function. The width of this tube is controlled by the parameter (ϵ). Data points located on the boundary or outside the tube are called support vectors and are used to construct the regression model.

The parameter (C) controls the penalty assigned to prediction errors and determines how strongly the support vectors influence the fitted model.

4.4.1 SVR Methodology

The SVR methodology consisted of the following steps:

- Use a single global model to obtain predictions.
- Determine important and unimportant biomarkers.
- Remove unimportant biomarkers based on SHAP values.
- Optimize (C) and (ϵ) using random search.
- Pair SVR with K-means clustering to compare representative clusters with patients in similar age groups.
- Correct regression dilution using an uncompressed prediction based on the training set.
- Center the corrected prediction at zero rather than at the model mean.
- Conduct centile analysis to determine the atypicality of each test participant and their biomarkers.

SVR Results

The reported SVR results were:

Input	Result
Whole data set split into 10 year age columns	7.9 years
SHAP tuned model (0.02)	2.4 years
SHAP tuned model with tuned parameters	1.94 years
SHAP tuned model with tuned parameters and tuned SHAP thresholds	1.92 years
New data set with SHAP and K-Means on one age bracket	8.54 years

4.4.2 SVR Results

Highlighted below are some of the plots from the recent SVR runs

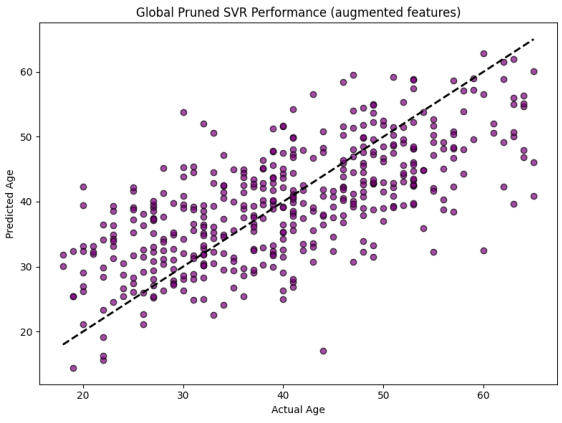


Figure 4: SVR SHAP with K-Means; RMSE=8.54 years

The model also generated participant-level outputs containing participant ID, actual age, corrected age gap, and centile deviation. Examples included:

Table 3: Results from SVR run output to CSV

PPT ID	Actual Age	Corrected Gap	Centile Deviation
478	42 y/o	-1.5 years	-13.9
1745	51 y/o	-11.3 years	-24.8
1851	64 y/o	-9.1 years	-11.9

4.4.3 SVR Results

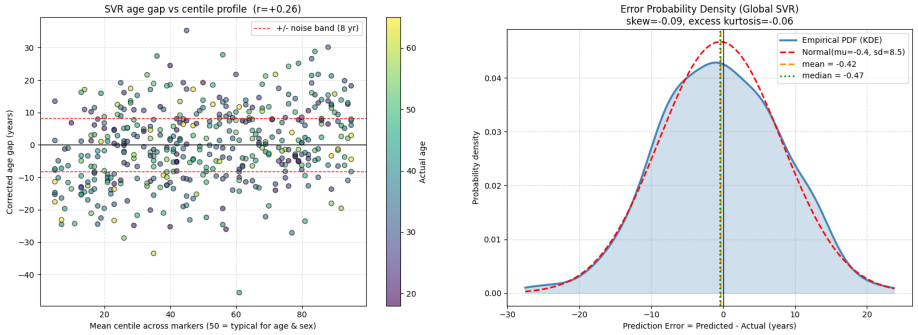


Figure 5: Age gap Vs. centile deviation (Left), Probability density of Age Gap(Right)

4.4.4 SVR Limitations

- Interpreting outputs
  - in the pursuit of lower RMSE it continues to be difficult to interpret the outputs on a per patient status. Finding what error is model noise and what is a true biological indicator
- Dataset Limitations
  - In current literature, most Bio-age projects have an output variable in their dataset. Something like mortality or morbidity to fit a rate of true aging to the patients.

4.5 Least Squares Support Vector Regression

Least Squares Support Vector Regression (LSSVR) is similar to Support Vector Regression in that they have the same ϵ tube around a hyperplane in your space and C parameter. In SVR, we determine the placement of the tube by looking at the distances of the support vectors outside of the tube and minimizing. In LSSVR, we look at the least squared distance of the support vectors outside the tube to minimize.

4.5.1 Results and Methods for LSSVR

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

Table 4: Initial attempts and results for the LSSVR with the Caesar dataset compared to the Pennington Dataset.

4.6 Kernel-Based Regression Models

4.6.1 Comparision Analysis

The objective of this study is to compare five regression models for biological age prediction using anthropometric measurements from the CAESAR dataset:

- RBF Support Vector Regression (SVR)
- RBF Kernel Ridge Regression (KRR)
- Gaussian Process Regression (GPR)

A balanced representation of the age distribution during SHAP background is constructed using age-stratified K-means representatives to obtain interpretable feature importance across the entire age range of 1869 years.The age range is divided into five equal-width buckets:18-27, 28-37, 38-47, 48-57, 58-69.

Before model training, all predictor variables are standardized using where

$$x_i^* = \frac{x_i - \mu}{\sigma}$$

μ denotes the training mean, σ denotes the training standard deviation.

Similarly, the response variable (Age) is standardized before regression and transformed back to years after prediction. Rather than performing one global random train/test split, the data are partitioned independently within each age bucket.For each bucket, 80% of the observations are assigned to the training set, and 20% are assigned to the testing set.This guarantees that every age interval contributes training and testing observations while preventing age imbalance across the evaluation sets. Under training with SHAP background, for each bucket applies 7 K-means clusters, creating 35 observations.

Support Vector Regression: Support Vector Regression (SVR) is a kernel-based supervised learning algorithm used for nonlinear regression. The objective is to estimate a function that accurately predicts the target variable while maintaining good generalization to unseen observations. In this work, the Radial Basis Function (RBF) kernel is employed to model nonlinear relationships between anthropometric measurements and biological age. The regression function is

$$f(\mathbf{x}) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b, \tag{1}$$

where α_i, α_i^* are the support vector coefficients, b is the bias term, $K(\cdot, \cdot)$ denotes the kernel function.

The RBF kernel is

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right), \tag{2}$$

where γ controls the influence of neighboring observations. The implementation uses $C = 5.0, \epsilon = 0.1, \gamma =$ `scale`, with all predictor variables standardized before training.

Kernel Ridge Regression (KRR) Kernel Ridge Regression combines ridge regression with kernel methods to model nonlinear relationships. Instead of explicitly estimating regression coefficients in the original feature space, KRR performs regression in an implicit high-dimensional feature space using the kernel trick. The prediction function is

$$\hat{y}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x}), \tag{3}$$

where α_i are the kernel coefficients, $K(\mathbf{x}_i, \mathbf{x})$ is the kernel function. The coefficients are obtained from

$$\alpha = (K + \lambda I)^{-1} \mathbf{y}, \tag{4}$$

where K is the kernel matrix, λ is the ridge regularization parameter. The implementation employs the RBF kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right), \tag{5}$$

with $\lambda = \alpha = 1.0$.

The kernel bandwidth uses the default scikit-learn setting because `gamma=None` in the implementation.

Gaussian Process Regression (GPR) Gaussian Process Regression is a Bayesian nonparametric regression model in which the unknown regression function is modeled as a Gaussian process. Instead of producing only a point prediction, GPR estimates a predictive probability distribution, allowing both the predicted age and the associated uncertainty to be computed.

The regression function is modeled as

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{6}$$

where $m(\mathbf{x})$ is the mean function, $k(\mathbf{x}, \mathbf{x}')$ is the covariance function. The implementation assumes a zero mean function, $m(\mathbf{x}) = 0$.

The covariance kernel used is

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2l^2}\right) + \sigma_n^2, \tag{7}$$

which corresponds to a Constant Kernel multiplied by an RBF kernel together with a White Kernel accounting for observation noise.

For a new observation,

$$f_* \mid X, \mathbf{y}, \mathbf{x}_* \sim \mathcal{N}\left(\mu_*, \sigma_*^2\right), \tag{8}$$

where μ_* denotes the predicted age, σ_* denotes the predictive standard deviation.

4.6.2 Correlation and Shap Analysis

To investigate the relationship between anthropometric biomarkers and biological age, Pearson correlation analysis is performed. Pearson correlation measures the strength and direction of the linear association between a biomarker X and age Y, and is computed as

$$r = \frac{\sum_{i=1}^{n}(x_1 - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2 n}}$$

where n is the number of observations, and $\overline{x}$ and $\overline{y}$ denote the sample means of the biomarker and age, respectively. The correlation coefficient satisfies $-1 \leq r \leq 1$, where r=1 represents a perfect positive linear relationship, r=1 represents a perfect negative linear relationship, and r=0 indicates no linear association.

Random Forest Model To model the relationship between anthropometric measurements and age, a Random Forest regression model was trained. A Random Forest consists of an ensemble of decision trees, where each tree predicts the target independently, and the final prediction is obtained by averaging the predictions of all trees:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x})$$

where T is the number of trees, $f_t(\mathbf{x})$ denotes the prediction from the t^{th} decision tree, and $\hat{y}$ is the final predicted age. In this study, the Random Forest model consisted of 500 decision trees.Model performance was evaluated using the root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2).

SHAP Feature Importance: To interpret the Random Forest model, SHapley Additive exPlanations (SHAP) were computed using the TreeExplainer algorithm. For an individual prediction, SHAP decomposes the model output into a baseline prediction and feature contributions,

$$f(\mathbf{x}) = \phi_{\mathbf{0}} + \sum_{\mathbf{i=1}}^{\mathbf{M}} \phi_{\mathbf{i}}$$

where

- ϕ_0 is the expected model prediction
- ϕ_i is the SHAP value corresponding to the i^{th} biomarker
- M is the total number of biomarkers

The feature importance used in this study is the mean absolute SHAP value

$$I_i = \frac{1}{N} \sum_{j=1} N|\phi_{ij}|$$

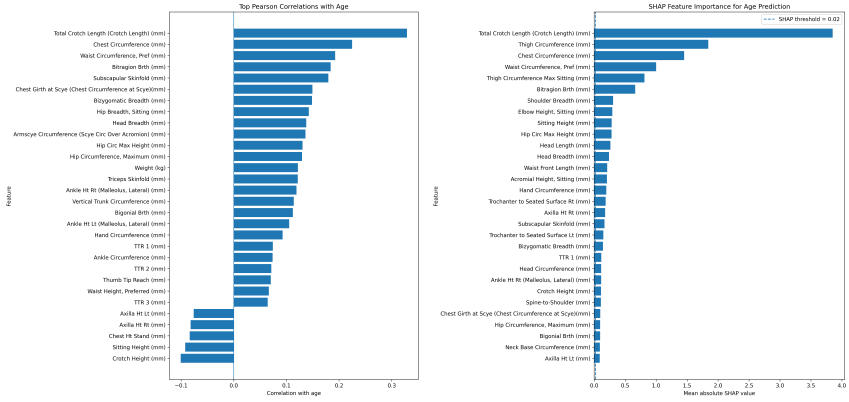
where N is the number of evaluated samples and ϕ_{ij} is the SHAP value of feature i for sample j. Larger values indicate biomarkers that contribute more strongly to the prediction of age. The biomarkers with

$$I_i \geq 0.02$$

were considered important and retained for further interpretation.

Linear and Nonlinear Relationships: Pearson correlation evaluates only the linear relationship between an individual biomarker and age. Consequently, biomarkers that influence age prediction only through interactions with other variables or through nonlinear patterns may exhibit small Pearson correlation coefficients.

In contrast, the Random Forest model partitions the predictor space using multiple decision trees, enabling it to model nonlinear relationships, threshold effects, and complex interactions among anthropometric biomarkers. SHAP therefore measures the contribution of each biomarker within this nonlinear predictive model rather than its direct linear association with age. Consequently, biomarkers with high SHAP values but relatively low Pearson correlations indicate nonlinear predictive effects that cannot be identified using correlation analysis alone.



Comparison of Pearson Correlation and SHAP: The combined analysis provides complementary information regarding the importance of anthropometric biomarkers. Pearson correlation identifies biomarkers that exhibit strong linear associations with age, whereas SHAP quantifies the contribution of each biomarker to the nonlinear Random Forest prediction model. Biomarkers that rank highly in both analyses, such as **Total Crotch Length, Chest Circumference,**

Waist Circumference (Preferred), and Bitragion Breadth, demonstrate both strong linear relationships with age and high predictive importance within the machine learning model. These biomarkers may therefore be considered robust indicators of age.

Conversely, biomarkers such as **Thigh Circumference and Shoulder Breadth** exhibit relatively higher SHAP importance than Pearson correlation, indicating that their influence on age prediction arises primarily through nonlinear relationships or interactions with other anthropometric measurements. On the other hand, biomarkers showing relatively high Pearson correlation but lower SHAP importance contribute linearly to age but provide limited additional predictive information once correlated biomarkers are already included in the Random Forest model.

Overall, Pearson correlation provides a statistical assessment of individual linear relationships, whereas SHAP explains the contribution of each biomarker within the complete nonlinear prediction model. The agreement between the two methods identifies stable biomarkers associated with aging, while differences between the methods reveal nonlinear relationships and feature interactions captured only by the machine learning model.

Table 5: Biomarkers showing strong agreement between Pearson correlation and SHAP feature importance for age prediction.

Biomarker	**Pearson Rank**	**SHAP Rank**	**Interpretation**
Total Crotch Length (Crotch Length)	#1	#1	Strongest predictor overall.
Chest Circumference	#2	#3	Strong linear and model importance.
Waist Circumference (Preferred)	#3	#4	Consistently important.
Bitragion Breadth	#4	#6	Important cranial measurement.
Subscapular Skinfold	#5	#18	Linear relationship with moderate nonlinear contribution.
Head Breadth	#10	#12	Consistent anthropometric marker.
Hip Circ Max Height	#11	#10	Good agreement.
Hand Circumference	#19	#16	Moderate importance in both analyses.

4.6.3 RMSE Comparison of Weighted-RBF Regression Models

Table 6: Root Mean Square Error (RMSE, years) of the six regression models across a single global group. The models compare correlation-weighted and SHAP-weighted 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	**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

4.7 Random Forest Results

- Use a random search algorithm to test combinations of the many parameters used in RF across a grid of different values selected by the developer
- Train the RF model on 80% of the data to give it enough information to make reliable decisions.
- Pair with K-Means clustering on new data
- Use TreeShap to get marker importance to see how they affect the output of the model
- Run a centile deviation analysis to obtain atypicality ratings for each patient just as in the SVR.

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 7: Results from RF tests.

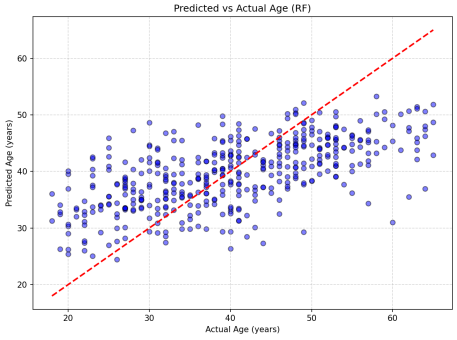


Figure 6: RF with K-Means on single age bracket: RMSE = 9.09 years

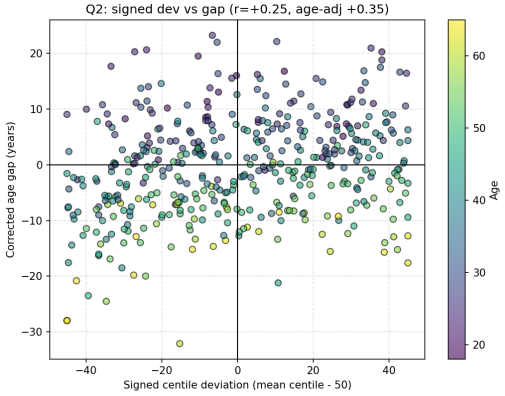


Figure 7: Signed deviation Vs. Age gap

Random Forest Limitations

- Random Forest still underperformed compared to other models even on the larger dataset.

- The piecewise nature of the RF handles other data more elegantly as it is built to withstand harsh interactions and breakpoints. Our data favors more of a linear nature with the continuous trend of aging, indicating the use of Kernel based methods work better with our specific data.

4.8 Goals moving forward

- Creating a sublevel under Age grouping and understanding the important biomarkers that impact Biological Age.
- Fine tuning the models to be work well with the caesar dataset. (trying different loss functions or different parameters)
- Focusing in on male vs. female predictions.

5 Acknowledgments

We would like to thank

- Prof. Peter Wolenski and Dr. Nadejda Drenska for guiding and supporting us.
- Dr. Steven B. Heymsfield from Pennington Biomedical Research Center for giving us the datasets and supporting the evaluation of the models.

References

[1] 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.

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

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

[4] 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.