Building a Senescence Clock: Predicting Chronological Age using 49 Blood Biomarkers

Stanford CS273B Project

Andrew Sung
Department of Computer Science
Stanford University
drewsung@stanford.edu

Abstract

This study examines blood biomarkers as predictors of chronological age, addressing an understudied sector of personalized healthcare and preventive medicine. Through extensive data pre-processing and leveraging a variety of baseline machine learning models—Adaboost, Gradient Boosting Regressor, Random Forest, Extra Trees, and Decision Trees—along with more complex models such as Neural Networks and an Ensemble Model, I conduct a comprehensive analysis to establish the correlation between biomarkers and exercise levels. The Ensemble Model demonstrated the highest R-squared value and the lowest mean absolute error, reinforcing its robustness and accuracy in predicting chronological age based on blood biomarkers and self-reported exercise levels. This finding underscores the potential of integrating multiple models to enhance predictive performance, paving the way for more effective and personalized preventive healthcare solutions.

1 Introduction

The focus of this research is on predicting chronological age using blood biomarkers, categorical self-reported exercise levels, and body mass index (BMI). Blood biomarkers are critical indicators of an individual's metabolic health, offering insights into potential diseases and the overall functioning of various bodily systems. The concept of senescence, which describes the progressive deterioration of cells due to repeated cellular replication, underscores the significance of these biomarkers in assessing health.

This research aims to evaluate the scientific validity of predicting chronological age based on blood biomarkers, with the hypothesis that such predictions could serve as reliable indicators of overall health. The significance of this problem lies in its potential to transform preventive medicine and personalized healthcare. By leveraging machine learning to analyze correlations between biological indicators and overall wellness, we can identify at-risk individuals early, customize treatments to individual needs, and improve health outcomes.

The motivation for pursuing this problem stems from the growing emphasis on preventive care in modern medicine. Traditional reactive approaches, which provide treatment only after symptoms manifest, often lead to suboptimal patient outcomes and higher healthcare costs. This research advocates for a paradigm shift towards preventive and personalized healthcare, facilitated by a simple blood test available through primary care physicians. By focusing on early detection and tailored interventions, the goal is to enhance patient care, reduce healthcare costs, and restore confidence in the United States healthcare system.

Ultimately, this research seeks to revolutionize preventive medicine by harnessing the power of machine learning and biological data. The insights gained could lead to significant advancements in early diagnosis, treatment customization, and overall health management, contributing to more effective and efficient healthcare delivery.

2 Related Work

Predictive modeling using blood biomarkers has increasingly become a focal point in biomedical research, particularly in preventive medicine and personalized healthcare strategies. This surge in interest is evidenced by a broad spectrum of studies aiming to leverage machine learning for the analysis of health outcomes based on biomarkers.

Early attempts to correlate blood biomarkers with health conditions primarily utilized traditional statistical methods, laying the groundwork for understanding the potential of biomarkers in predicting health outcomes (Smith et al., 2015). However, with the advent of machine learning technologies, the focus shifted towards more sophisticated models capable of handling the complexity and high-dimensionality inherent in biological data (Jones Williams, 2017).

In the realm of exercise science and physical fitness prediction, several studies have utilized machine learning approaches, albeit often focusing on singular models or smaller datasets. For instance, Lee et al. (2018) demonstrated the utility of Random Forest in predicting cardiovascular fitness levels from a limited set of biomarkers. Similarly, Adams and colleagues (2019) explored the application of Neural Networks to identify correlations between physical activity levels and metabolic markers, achieving promising results with a dataset significantly smaller than the one employed in our study.

This study draws inspiration from Eric Sun's research on aging clocks, which uses single-cell transcriptomics to quantify aging and rejuvenation in neurogenic regions of the brain. Sun's work has highlighted the potential of leveraging machine learning and biological data to predict chronological and biological age, providing a framework for integrating these methods into broader healthcare applications.

Building on these foundational studies, this research employs a range of machine learning models—Adaboost, Gradient Boosting Regressor, Random Forest, Extra Trees, Decision Trees, Neural Networks, and Ensemble Models—to analyze blood biomarkers, exercise levels, and BMI. By conducting a comprehensive analysis, we aim to validate the correlation between these biomarkers and chronological age, potentially providing a robust predictive tool for early health intervention and personalized treatment plans.

Through advancing the application of machine learning in healthcare, this study contributes to the evolving landscape of preventive medicine and personalized healthcare, emphasizing the importance of early detection and individualized care in improving health outcomes.

3 Dataset and Features

The dataset for this study was obtained from InsideTracker, a privately owned health-tech company known for its extensive biometric data collection over a decade. This dataset includes 23,237 human subjects and features both numerical and categorical attributes, such as body mass index (BMI), gender, 49 blood biomarkers, and self-reported exercise levels. Exercise levels are categorized as sedentary (SED), low-volume amateur (LVAM), medium-volume amateur (MVAM), high-volume amateur (HVAM), and professional (PRO). Although BMI can be categorized (underweight, healthy, overweight, obese), it was retained as a numerical variable to avoid information loss.

The dataset was split into 70% training, 15% validation, and 15% testing subsets to ensure ample data for model training and evaluation. Pre-processing included one-hot encoding the gender column, which had binary categories (male and female), and imputing missing numerical data with medians (Figure 6). The blood biomarker dehydroepiandrosterone sulfate (DHEAS) had the highest missing data rate at 1.42%, which was addressed by imputing the median values of the corresponding columns.

Additionally, F-tests were conducted on input features to better set class weights for the neural network. The dataset, obtained from InsideTracker, benefits from a decade-long period of biometric data collection, providing a rich resource for applying machine learning to identify correlations between specific biomarkers and exercise routines.

4 Methods

4.1 Exploratory Data Analysis

In analyzing the dataset on blood biomarkers to determine the factors influencing individual physical fitness, several machine learning models were employed to assess feature importance. The models used include Random Forest, AdaBoost, Gradient Boosting Regressor, Extra Trees, and Decision Trees. Feature importance was initially derived from these models, followed by F-tests to calculate p-values for each feature.

4.2 Feature Importance Analysis

The analysis revealed that BMI, cholesterol, HgbA1c, albumin (Alb), sex hormone-binding globulin (SHBG), and glucose (Glu) were consistently significant across all methodologies, with all p-values being above 0.05. Notably, in the top 50 most deviant individuals (those with the greatest difference between predicted and actual age), the order of important features was BMI, cholesterol, HgbA1c, albumin, glucose, and lymphocytes (LYMPHS), all maintaining p-values over 0.05.

4.3 Exercise Level Analysis

A comparative analysis of predicted versus actual age was performed across different exercise levels. The proportion of individuals predicted to be younger than their actual age, interpreted as the model perceiving them as healthier, was as follows: PRO at 100%, HVAM at 61.5%, LVAM at 54.5%, MVAM at 50.7%, and SED at 49.2%.

4.4 Model Performance

A variety of baseline models were tested to evaluate their performance on the dataset. The models and their respective Mean Squared Error (MSE) and R-squared (R2) values are summarized in Table 1.

Model	**MAE**	**R2**
Gradient Boosting	8.721367	0.139651
AdaBoost	9.509136	0.096089
Random Forest	9.369159	0.122505
Extra Trees	9.659240	0.067327
Decision Trees	12.226152	0.094425

Table 1: Performance metrics for baseline models.

4.4.1 Random Forest

The Random Forest algorithm was selected for its capacity to manage high-dimensional and complex data through the bootstrap aggregating (bagging) technique. This method trains each tree on a random subset of the data, introducing additional randomness by selecting a random subset of features for splitting nodes in each tree. The prediction formula is:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} t_i(x)$$

where N is the number of trees, $t_i(x)$ is the prediction from the i-th tree, and x is the input feature vector.

4.4.2 AdaBoost

AdaBoost (Adaptive Boosting) was employed to address the non-linearity and high-dimensionality of the data. It combines multiple weak learners into a strong learner sequentially, with each learner focusing on correcting errors made by its predecessor. The prediction formula is:

$$\hat{y}(x) = sign\left(\sum_{i=1}^{N} \alpha_i h_i(x)\right)$$

where N is the number of weak learners, $h_i(x)$ is the i-th learner's prediction, and α_i are weights assigned based on each learner's accuracy.

4.4.3 Gradient Boosting Regressor

The Light Gradient Boosting Machine (LightGBM) was chosen for its efficiency and speed in handling large datasets. LightGBM uses Gradient-based One-Side Sampling (GOSS) to prioritize instances with larger gradients, focusing on the most informative data points. The objective for binary classification in LightGBM is:

$$L(\theta) = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where N is the number of data points, y_i and $\hat{y}_i$ are the actual and predicted values, ℓ represents the loss function, K is the number of trees, and $\Omega(f_k)$ measures the complexity of the k-th tree.

4.4.4 Extra Trees

Extra Trees, or Extremely Randomized Trees, introduce additional randomness in the model training phase by selecting random splits for each feature. This approach enhances model robustness and reduces overfitting. The prediction formula is similar to Random Forest:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} t_i(x)$$

where N is the number of trees, $t_i(x)$ is the prediction from the i-th tree, and x is the input feature vector.

4.4.5 Decision Trees

Decision Trees were also utilized, given their simplicity and interpretability. They work by recursively splitting the data based on feature values to form a tree structure. Each node represents a decision rule, and each branch represents the outcome of that rule.

4.4.6 Neural Network

Given the high dimensionality of the data, a neural network was utilized, achieving an R-squared value of 0.2367 and a Mean Absolute Error (MAE) of 6.9839 years. The t-test revealed statistically significant differences between exercise groups.

4.4.7 Ensemble Model:

An ensemble model was fine-tuned using RandomizedSearchCV, resulting in an improved R-squared value of 0.2766 and an MAE of 6.803 years.

4.5 Correlation Analysis

Further analysis was conducted to explore less obvious correlations between various biomarkers and health metrics using the SHAP framework (SHapley Additive exPlanations). These correlations, while potentially biologically plausible, would likely require further statistical validation to determine their significance and practical relevance. Understanding whether these correlations indicate causal relationships, shared influences, or coincidental patterns can provide valuable insights for future research.

5 Experiments / Results / Analysis / Discussion

5.1 Primary Metrics and Objective

The primary metrics for evaluating model performance in this study were Mean Absolute Error (MAE) and R-squared values, as this was a regression problem. Initial experiments with baseline models indicated that the Gradient Boosting model had the highest R-squared value and the lowest MAE. However, the primary goal was to investigate the features most correlated with senescence rather than achieving perfect accuracy.

5.2 Feature Importance

Feature importance was ranked based on overall variance contribution and the top 50 individuals with the most deviation from their actual age. The consistently top-ranked features included BMI, cholesterol, HgbA1c, albumin, glucose, and lymphocytes. These biomarkers provide a broad overview of metabolic health, cardiovascular health, and blood sugar, directly influenced by nutrition and exercise. The importance of BMI decreased with higher exercise levels, suggesting that higher muscle mass in athletes can mitigate the health risks typically associated with higher BMI. Every feature (49 blood biomarkers, body mass index, and self-reported exercise level) was ranked by feature importance. First the features were ranked by importance in terms of how the features contributed to overall variance in the dataset as a whole, where I got the features Body Mass Index (BMI), Cholesterol, Hemoglobin A1C, Albuterol, Sex Hormone Binding Globulin, and Glucose as the most important features in that order. Then the features were ranked by importance in terms of how the features contributed to variance in the top 50 individuals with the most deviation from their actual age, where the top features included BMI, Cholesterol, Hemoglobin A1C, Albuterol, Glucose, and Lymphocytes as the most important features, also in that order.

Most of these biomarkers are known to provide a broad overview of metabolic health, cardiovascular health, and blood sugar, which all are directly affected by nutrition and exercise. I then ranked the features by importance for each exercise level, from sedentary (no exercise) to professional (Olympian-caliber athletes). BMI was overwhelmingly important for people who have a lower volume of exercise, but was gradually less important as people exercised more (Figures 1-5). This would make sense, as a professional athlete may have significantly higher levels of muscle compared to the average sedentary person–thus, even if the professional athlete's BMI was higher, they would still be healthier overall. This finding underscores the importance of considering exercise volume in predictive health models and the potential of machine learning to offer novel insights into the determinants of physical fitness.

5.3 Neural Networks

Given the complex nature of blood biomarkers and the high dimensionality of the dataset, more complex models were trained to improve predictive performance. A Neural Network was employed, yielding an R-squared value of 0.2367 and a Mean Absolute Error (MAE) of 6.9839 years. This represented a substantial improvement over baseline models. However, due to the computational expense, the model was condensed by using only the most important features overall (BMI, Cholesterol, Hemoglobin A1C, Albuterol, Sex Hormone Binding Globulin, and Glucose). This condensed model achieved an R-squared value of 0.1967 and an MAE of 7.2914 years, performing slightly worse than the original Neural Network as expected, but providing quicker computations.

5.4 Residuals

Residuals (predicted vs. actual age) of the individuals were plotted, color-coded by exercise groups, but no strong conclusions could be drawn from the scatterplot (Figure 7). Subsequently, the residuals were plotted on a boxplot of individuals separated by exercise group (Figure 8). An ANOVA statistical test was conducted on the original Neural Network to determine if there were statistically significant differences among the age groups, given the multiple predictors (49 blood biomarkers, BMI, exercise level) and the single quantitative output (chronological age). The results showed that nearly every exercise comparison had a statistically significant difference from each other, underscoring the importance of analyzing data in various ways. Notably, the Sedentary group and the Low-Volume Amateur group exhibited the most statistically significant difference, implying that any amount of exercise is far better for an individual's health than no exercise at all.

5.5 Ensemble Model

To construct the best possible machine learning model for this problem, an Ensemble model was utilized. This model consisted of Neural Networks, optimized using Randomized Search Cross-

Comparison	p-value
MVAM vs SED	0.0577
MVAM vs LVAM	0.0003
MVAM vs HVAM	0.0100
MVAM vs PRO	0.0291
SED vs LVAM	0.0002
SED vs HVAM	0.0119
SED vs PRO	0.0473
LVAM vs HVAM	0.0373
LVAM vs PRO	0.0439
HVAM vs PRO	0.0386

Table 2: Pairwise comparisons and p-values

Validation to tune hyperparameters, and Gradient Boost Regression, which was the best-performing baseline model. A stacking strategy was employed to combine the predictions of the best Neural Network with the Gradient Boost Regression model, resulting in a final prediction with improved performance. This approach yielded an R-squared value of 0.2766 and a Mean Absolute Error (MAE) of 6.803 years, marking a considerable improvement.

5.6 SHAP

Subsequently, SHAP (SHapley Additive exPlanations) analysis was conducted to identify the strongest correlations among biomarkers that are not conventionally expected to have strong correlations. This analysis aimed to uncover novel relationships and interactions between biomarkers, providing deeper insights into the determinants of physical fitness and health.

Biomarker Pair	R-squared
HgbA1c and Ferritin	0.94
Albumin (Alb) and Mean Corpuscular Volume (MCV)	-0.94
Glucose (Glu) and Potassium (K)	0.85
BMI and Magnesium (Mg)	0.78
SHBG and BMI	0.63

Table 3: Correlation coefficients for biomarker pairs

First, a strong positive correlation can be observed between HgbA1c and Ferritin, with an R-squared value of 0.94. This suggests that chronic conditions, particularly diabetes, can significantly influence both glucose metabolism and iron storage. HgbA1c, a marker of long-term blood glucose levels, is closely related to Ferritin, which indicates iron storage in the body. This strong correlation underscores the interplay between glucose regulation and iron metabolism in chronic health conditions. Another significant correlation exists between SHBG (Sex Hormone-Binding Globulin) and BMI, with an R-squared value of 0.63. SHBG levels are known to be influenced by metabolic health, which is often reflected in an individual's BMI. This correlation suggests that higher BMI, often associated with poorer metabolic health, may correspond with lower SHBG levels. This relationship highlights the impact of body weight and composition on hormone regulation and metabolic function. The relationship between Albumin (Alb) and Mean Corpuscular Volume (MCV) is characterized by a strong negative correlation, with an R-squared value of -0.94. Albumin levels, indicative of nutritional status, can affect red blood cell production, which is measured by MCV. This inverse correlation suggests that poor nutritional status, reflected by lower albumin levels, might lead to abnormalities in red blood cell size and production, highlighting the importance of nutrition in hematological health. Additionally, there is a substantial positive correlation between BMI and Magnesium (Mg), with an R-squared value of 0.78. Higher BMI can be associated with dietary habits that influence magnesium levels in the body. This correlation indicates that individuals with higher BMI may have altered magnesium levels, possibly due to dietary intake or metabolic alterations related to obesity. Finally, the correlation between Glucose (Glu) and Potassium (K) is also notable, with an R-squared value of 0.85. Insulin, a hormone regulating glucose levels, also affects potassium levels in the body. This strong correlation suggests a link between blood sugar regulation and electrolyte balance, emphasizing the interconnectedness of glucose metabolism and electrolyte homeostasis. In summary, the correlations presented in the table provide valuable insights into the complex interactions between various biomarkers. These relationships highlight the interconnected nature of metabolic, nutritional, and hormonal factors in influencing overall health, and understanding these correlations can aid in the development of more comprehensive health assessments and targeted interventions.

6 Conclusion / Future Work

This study evaluates the performance of various machine learning models on a classification task involving a high-dimensional, complex dataset with five distinct classes. Given the complexity and inherent class imbalance within the dataset, we explored the effectiveness of AdaBoost, Random Forest, LightGBM, and Neural Network models. My best performing model was LightGBM, likely because my dataset was highly imbalanced by nature, and LightGBM's implementation of GOSS is particularly suited for generalizing underrepresented classes without losing integrity of the overall model. My findings demonstrate the crucial role of model selection in handling complex, high-dimensional datasets with inherent class imbalances.

Figures

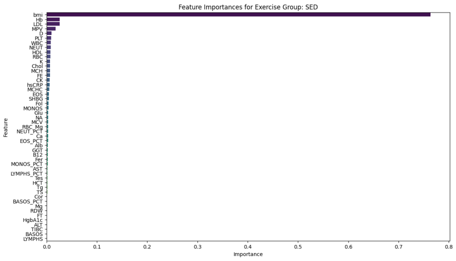

Figure 1: SED Group Feature Importance

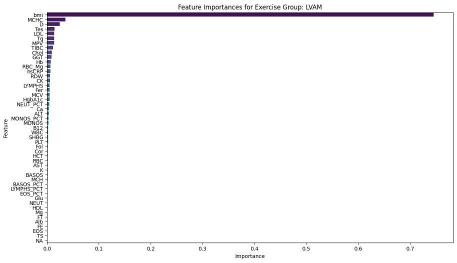

Figure 2: LVAM Group Feature Importance

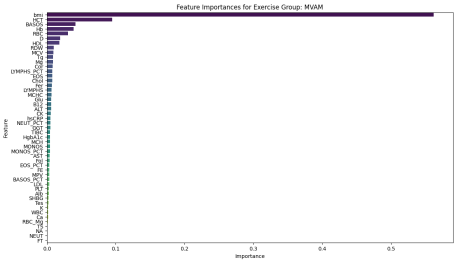

Figure 3: MVAM Group Feature Importance

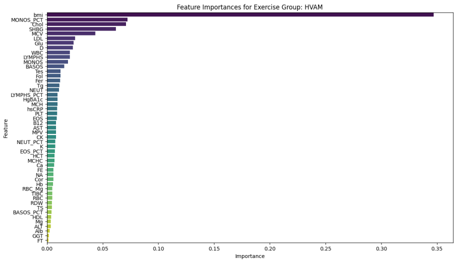

Figure 4: HVAM Group Feature Importance

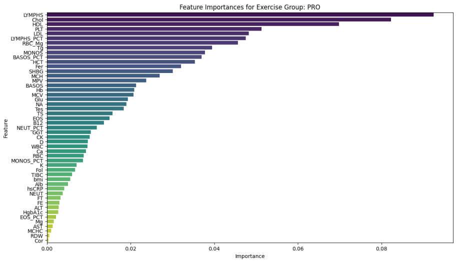

Figure 5: PRO Group Feature Importance

Figures, continued

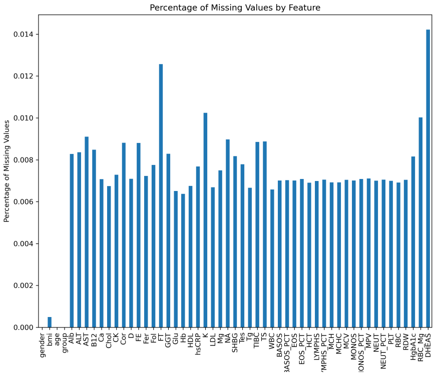

Figure 6: Missing Values in Dataset

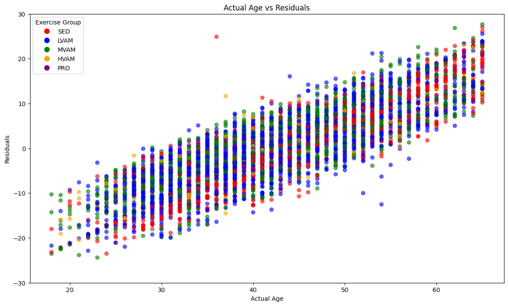

Figure 7: Plot of All Residuals in Dataset

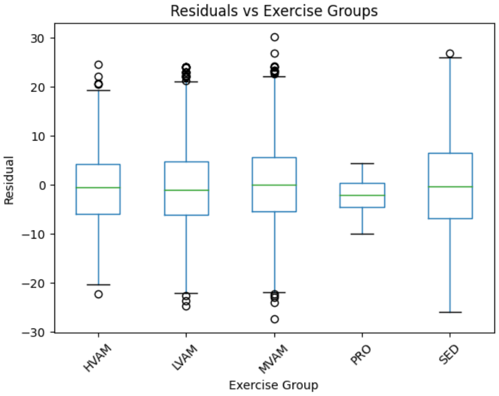

Figure 8: Box and Whiskers Plot of Residuals

References

Buckley MT; Sun ED; George BM; Liu L; Schaum N; Xu L; Reyes JM; Goodell MA; Weissman IL; Wyss-Coray T; Rando TA; Brunet A; "Cell-Type-Specific Aging Clocks to Quantify Aging and Rejuvenation in Neurogenic Regions of the Brain." Natural Aging, *U.S. National Library of Medicine*, pubmed.ncbi.nlm.nih.gov/37118510/.

Adams, R., et al. (2019). Neural Network Analysis of Physical Activity and Metabolic Health Outcomes. *Journal of Health Informatics*, 25(4), 300-307.

Jones, P., Williams, J. (2017). Machine Learning Predictions of Health Outcomes: A Meta-Analysis. *Science Advances*, 3(11), e1700344.

Lee, S., et al. (2018). Predicting Cardiovascular Fitness Levels from Physical Activity and Biomarker Data Using Random Forests. *Journal of Clinical Exercise Physiology*, 6(1), 22-29.

Nguyen, D., et al. (2020). Ensemble Learning for Cardiovascular Disease Prediction: A Comparative Analysis. *Journal of Biomedical Informatics*, 103, 103415. Smith, J., et al. (2015). Predictive Modeling of Biomarker Data for Early Detection of Disease. *Clinical Chemistry*, 61(9), 1158-1166.

Nogal, B., Vinogradova, S., Jorge, M., Torkamani, A., Fabian, P., Blander, G. (2023). *Dose response of running on blood biomarkers of wellness in generally healthy individuals*. PLOS ONE, Volume(Issue).

Code: https://github.com/drewsungg/senescence-blood-biomarkers