A Project Module Report for M. Sc. in Computational Biology

Human Transcriptomic Clock: Improvement and Validation of GTEx Clock

Submitted by,

Santhosh Gojjam Kantharaju

Under the Guidance of,

Prof. Dr. Bjorn Schumacher

Department of Biology

University of Cologne, Germany

Contents

1 Abstract

The difference between a person's biological age and chronological age can be predicted using an aging clock. These aging clocks are created using supervised machine learning methods, such as penalized regression techniques like elastic-net. Variations in RNA molecule abundance reflect various regulatory influences, resulting in significant changes in gene expression throughout an individual's life. Identifying these changes is crucial for developing transcriptomic aging biomarkers that are specific to RNA expression.

We developed a human transcriptomic clock using a quantization method and supervised machine learning techniques, including elastic-net regression and a simple neural network regression model, to predict biological age. Additionally, we are working on dimensionality reduction through principal component analysis (PCA) to build a PCAge clock. We are also validating our in-house developed quantized GTEx Clock using various publicly available datasets.

Keywords: Aging Clocks, Transcriptomic Clocks, Quantilization

2 Introduction

Biological age is a significant risk factor influencing an individual's health and mortality. Typically, a person's biological age is completely different from their chronological age [4]. Measuring biological age could be useful in identifying targeted personal health-promoting interventions which can also be specific to different diseases. It can also help in testing interventions aimed at modifying the aging process [2].

In the last decade, we have uncovered the remarkable ability of epigenetic changes to estimate a person's age. Epigenetics refers to the chemical modifications and organization of the genome that affect or indicate its activity, with strict definitions requiring inheritance through cell division. For over 50 years, researchers have reported observations of how age influences this mechanism, suggesting its role in age-related diseases [2].

The connection between epigenetic modifications and age became particularly clear with the introduction of high-throughput arrays [2], combined with advancements in machine learning and bioinformatics, has significantly accelerated the discovery of biomarkers. This progress has also enhanced the use of computational modeling to understand complex biological phenomena [6].

Aging clocks are usually created using supervised machine-learning methods. Linear regression serves as the foundation for many modern modeling tools. It offers the simplest form of a model, representing the regression function as a linear combination of predictors. Due to its linear structure, the model parameters are easy to interpret. The performance of a model can be assessed using its mean squared error (MSE), which is the sum of its squared bias and variance [11]. Additionally, R2 scores quantify the proportion of variance in the test data that can be explained by each model [8].

Regularization techniques are employed to improve model performance by introducing penalties that help stabilize it. Ridge linear regression applies an L2-norm penalty, which adds the squared values of the coefficients to the loss function. Lasso linear regression uses an L1-norm penalty, incorporating the absolute values of the coefficients

into the loss function. Elastic Net is a hybrid method that combines both regularization and variable selection. Ridge regression is particularly effective in addressing high multicollinearity issues, while Lasso regression is useful for feature selection. The L1 ratio determines the proportion of each regularization method to be utilized [1].

Hyperparameter tuning is crucial for improving the performance of the model, but it can result in significant overfitting. Therefore, it is common to split the available data into training and validation sets. The model that performs best on the validation data is then selected as the final model. One effective technique for this process is K-fold cross-validation, which partitions the data in such a way that each example is eventually used for both training and validation. In each of the splits, some examples are set aside for validation while the remainder are used for training [5].

Two core strategies feature selection and feature extraction improve the performance of regression models:

- Feature selection is a crucial step in improving the performance of a model. Often, the features in a dataset are not suitable for processing by subsequent modules due to high dimensionality and redundancy. The challenges associated with high-dimensional data, commonly referred to as the "curse of dimensionality" in statistical pattern recognition, indicate that the number of training examples must increase exponentially with the number of features to learn an accurate model. Since only a limited number of examples are usually available, there is an optimal number of feature dimensions. Beyond this point, the performance of the pattern analysis model tends to degrade [5].

  Random Forest is an ensemble technique that can also used for feature selection. It involves the random selection of features to construct a collection of decision trees with controlled variation. In each interior node of these trees, a subset of features is evaluated using the Gini index heuristic. The feature with the highest Gini index is selected as the splitting feature for that node. The Gini index measures the impurity of data, indicating how uncertain we are about the occurrence of an event [3].

- An effective way to enhance performance is through feature extraction. The objective of feature extraction is to create a low-dimensional representation that retains most of the information from the original feature vector. This process associates relevant information with the structure of the data such as variance and is particularly useful when conducting exploratory data analysis. Principal Component Analysis (PCA) is a technique used for signal representation that produces projections along the directions of maximum variance, which are determined by the first eigenvectors of the covariance matrix.

  A viable solution to the problem of collinearity in regression is to apply PCA and retain only a few of the principal components as regressors, also known as "latent variables." This method is referred to as Principal Components Regression (PCR). By using PCR, the regressors become effectively decorrelated, and the smaller eigenvalues that can lead to infinite values when calculating the pseudo-inverse are removed. The optimal number of principal components to retain for regression can be determined through cross-validation[5].

Additionally, artificial neural networks can be used for building the age prediction model using Multi-Layer Perceptron (MLP) regressors. These networks are structured as feed-forward systems, consisting of simple processing elements known as neurons, which connect similarly to biological neuronal circuits. During the training process, the MLP forwards all of its inputs through the network, compares the resulting outputs to the desired outputs, and then back propagates the errors. This adjustment process modifies each weight in the network based on its contribution to the overall error. The model can be enhanced through hyperparameter tuning, which involves optimizing activation functions and selecting appropriate optimization algorithms. Activation functions are crucial because they introduce nonlinearity to the network and significantly influence its ability to capture complex patterns in the data. Optimization algorithms work to minimize the difference between the predicted values and the actual values in the training dataset [5].

In this paper, we developed a model to predict biological age using transcriptomic data. We applied various regression techniques and incorporated feature selection through Random Forests. Additionally, we implemented 5-fold cross-validation to minimize the mean squared error (MSE). Our results indicated that the elastic net model achieved the best performance.

Additionally, we collected the coefficients and intercept values from an in-house generated Genotype-Tissue Expression (GTEx) Clock which has not been published yet, and validated them using various RNA sequencing data obtained from publicly available databases.

3 Methods

3.1 Data Collection and Processing

The raw counts for bulk RNA sequencing of whole human blood samples, as well as metadata were collected from `https://github.com/korean-genomics-center/transcriptomic_clock`. The dataset comprises 969 patients as columns and 69222 genes as rows. Patients without gene expression data were screened out and removed from the original dataset.

A custom function was used to apply the Transcripts per Million (TPM) normalization technique to standardize the dataset.

3.2 Feature Selection

To manage the large dataset, we performed feature selection to identify the important features. Initially, we removed genes with zero expression values across all patients. Additionally, we excluded genes with a median expression of zero, retaining only those with a median expression greater than 0.1.

We then applied the random forest (sklearn.ensemble.RandomForestRegressor) ensemble method to the remaining filtered genes to identify the most significant ones. These selected genes were subsequently used to train the regression models [9].

3.3 Quantilization

To optimize the values, a quantilization step was performed. In this process, Q10 quantilization was applied, where the samples were labeled between 1 and 10. The data was divided into 10 quantiles based on their row-wise values. As a result, the final output contains quantile labels instead of the original TPM values.

3.4 Sample Selection for Training Model

The dataset includes three types of patients: "Healthy," "Mental Disorder," and "Viral Infection." The training process was conducted only on the healthy subset, which

was extracted from the original dataset.

3.5 Trancriptomic Age Prediction Clock

Different models of the transcriptomic clock were created using the 5-fold cross-validation technique (sklearn.model-selection.KFold) with shuffling and random-state=42. Various models were trained to identify the one with the best performance.

3.6 Linear Regression Models

Linear regression models were trained using training and testing sets obtained from KFold output. Four parts of the dataset were used for training, while one part was reserved for testing. This process was repeated in a loop five times to explore all possible combinations of features. The mean squared error (MSE) and R2 score for each loop were calculated and recorded.

Additionally, linear models with L1 regularization, specifically Ridge and Lasso (sklearn.linear-model.Ridge/Lasso), were trained using alpha parameters that ranged from 0 to 1 in increments of 0.1. Elastic Net (sklearn.linear-model.ElasticNet), which incorporates both L1 and L2 regularization, was also trained using the same alpha values, along with an L1 ratio that varied from 0.0 to 1.0 in intervals of 0.2. Hyperparameter tuning was conducted to find optimal values for the alpha and L1 ratio to enhance the regression model's performance. The MSE and R2 scores for all models were calculated and documented.

3.6.1 Neural Network Model

A simple neural network was trained using the MLPRegressor from the sklearn.neuralnetwork module. Hyperparameter tuning was performed for various parameters, including activation functions (ReLU and Tanh), solvers (Adam, SGD, and lbfgs), maximum iterations (500, 1000, 2000, and 5000), and the architecture of the layers. For each model, the Mean Squared Error (MSE) and R2 scores were calculated and recorded.

3.6.2 PCAge Clock

The PCAge Clock was trained following the steps outlined in [4]. Data normalized by TPM was utilized, which was log-transformed using the np.log2 function and subsequently scaled through a quantization step. Principal Component Analysis (PCA) was then performed as a feature selection step. Only the most important features were retained, and the model was trained using linear regression, as previously described.

3.7 Validation of GTEx Clock

The coefficients and intercept values of in-house generated GTEx Clock were obtained. To validate the results, we selected common features (genes) that were present in both datasets. The validation was carried out using the formula $y = mx + c$, where y represents the predicted age, m is the coefficient derived from the GTEx Clock, x contains the values from the testing dataset, and c is the intercept.

Different datasets obtained from [12], [10], and `https://github.com/korean-genomics-center/transcriptomic_clock` were validated using this method, and a graph comparing actual age to predicted age was plotted.

4 Results and Findings

4.1 Chronological Age Prediction for Healthy Individuals

The dataset was initially visualized to examine the distributions among different groups of patients. A bar plot (Figure 1, 2) was created to illustrate this distribution. It was found that the dataset included 558 healthy samples, 287 cases of viral infections, and 124 cases of mental disorders. The dataset was also examined for sex distributions, resulting in the creation of an additional bar plot. From this analysis, we can conclude that age and sex are not confounded. The data included individuals from a variety of age groups, with ages ranging from a minimum of 20 years to a maximum of 79 years.

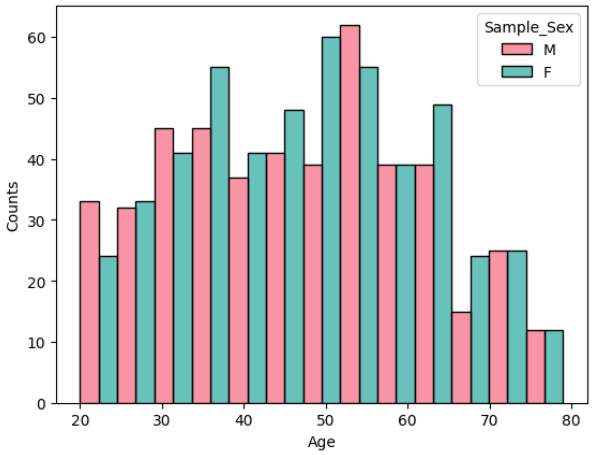

Figure 1: *Age Distribution in the dataset for both male and female*

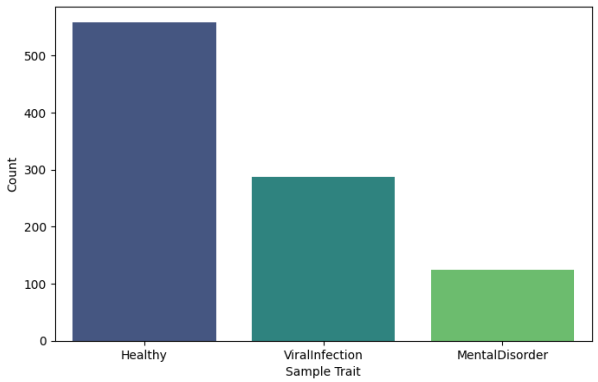

Figure 2: *Dataset Composition of Diseases*

4.1.1 Feature Selection

During feature selection, genes with a median expression value of zero were removed, as this indicated that they were not expressed in any patients. Alternatively, genes with

a median expression value greater than 0.1 were retained because they exhibited higher expression levels across the dataset, making them more reliable. This filtering process reduced the dataset from 69222 genes to 17294 genes. Consequently, only these 17294 genes with high expression levels are considered for analysis, leading to more accurate predictions.

After applying the random forest technique, the number of features was further reduced, selecting only those genes that could enhance the model's performance, ultimately decreasing the count from 17294 genes to 8647 genes.

4.1.2 Regression Model for Age Prediction

A linear regression model was trained only using the healthy cohort, utilizing optimal features. A scatterplot was created in (Figure 3), with the x-axis representing the actual age and the y-axis showing the predicted age according to the model. Each scatter point corresponds to a healthy individual, while the red line illustrates the regression trend. Various models were trained with different regularization parameters, and it was found that the Elastic-Net model, with an alpha value of 1.0 and, an l1 ratio of 0.4 achieved the best prediction performance, with an R2 score of 0.75 and MSE of 73 years.

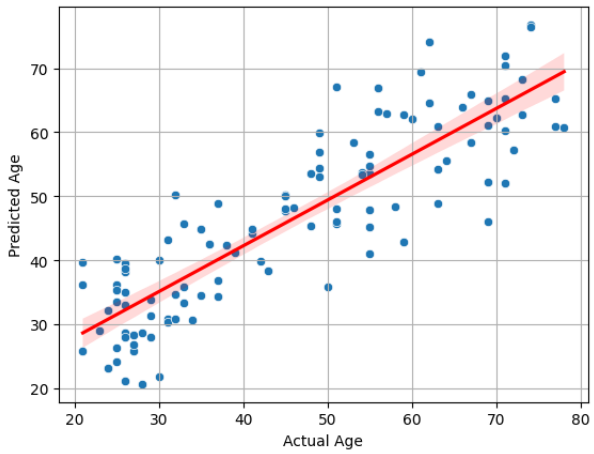

Figure 3: *Scatterplot of Predicted Age vs Actual Age, with the red line representing the regression line, colored dots representing the data points and red shaded region representing 95% confidence level. The model was build using Elastic-Net regression, alpha=1.0, and l1 ratio=0.4. The prediction performance of the model was R2 score=0.75*

4.1.3 PCAge Clock

The PCAge clock was constructed in a similar manner, but before using it in the linear regression model, optimal parameters were first determined. It was trained with various regularization parameters and different regression models. Ultimately, it was found that the Elastic-Net model provided the best prediction, and a scatterplot (Figure4) was plotted to show the result in a similar way mentioned previously. The model achieved an R2 score of 0.35.

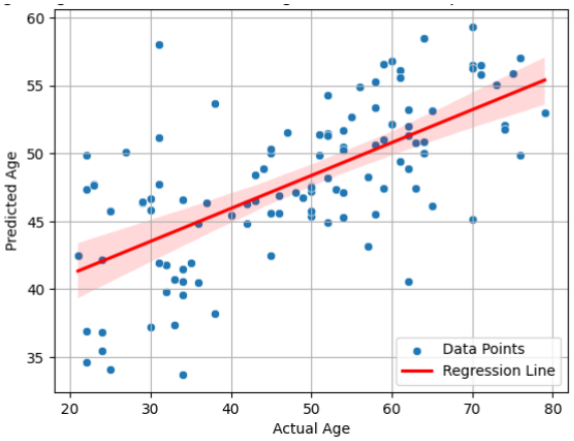

Figure 4: *Scatterplot of Predicted Age vs Actual Age, with the red line representing the regression line, colored dots representing the data points and red shaded region representing 95% confidence level. The model was build using Ridge regression, alpha=0.0. The prediction performance of the model was R2 score=0.35*

4.1.4 Neural Network Model for Age Prediction

A neural network was constructed and trained with 100 layers, using the ReLU activation function and the Adam optimizer. A scatter plot (Figure 5) for this model was created as mentioned previously. It was found that the neural network achieved prediction performance of an R2 score of 0.75 and MSE over 73 years.

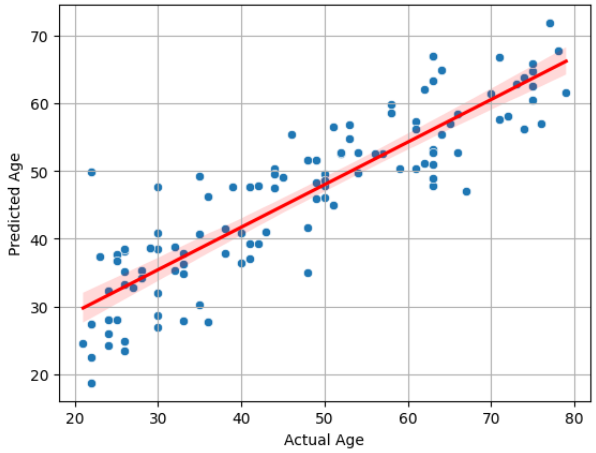

Figure 5: *Scatterplot of Predicted Age vs Actual Age, with the red line representing the regression line, colored dots representing the data points and red shaded region representing 95% confidence level. The model was build using MLP regressor (Neural Network), activation=ReLU, optimizer=Adam. The prediction performance of the model was R2 score=0.75*

4.2 Validation of GTEx Clock

Coefficients for the genes and the intercepts were obtained. These values were used to validate the clock. The scatter plot illustrates the actual ages compared to the GTEx Clock's predicted ages. Validation was carried out using both raw TPM values and Q10 TPM values.

The plots below indicate that the clock's performance with the unseen data is insufficient, and its overall performance is poor. In some cases, there is a negative correlation, while in others, it predicts an unusually high and negative age. The primary reason for this issue is the lack of data used in training the model, also the presence of outliers in the validation dataset. In particular, the dataset obtained from [10] contains a significant amount of outliers, resulting in negative values for predicted ages. This problem is also evident in the dataset obtained from `https://github.com/korean-genomics-center/transcriptomic_clock`. Therefore, careful screening of the data is essential to remove outliers before validating the clock. Additionally, the dataset from [12] lacks sufficient data, which can also lead to inaccurate results.

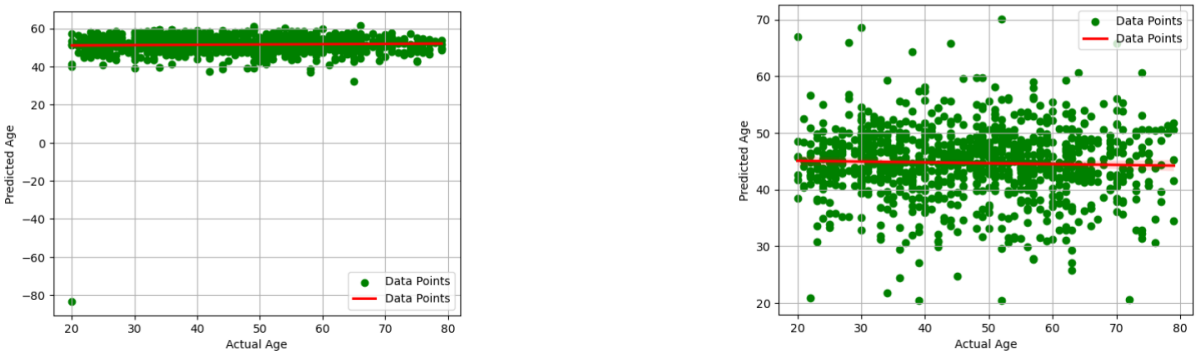

Figure 6: *Scatterplot of Predicted Age vs Actual Age, with the red line representing the regression line and green dots representing the data points. The GTEx Clock validation for the Korean dataset performs poorly using both quantilized (left) and TPM values (right).*

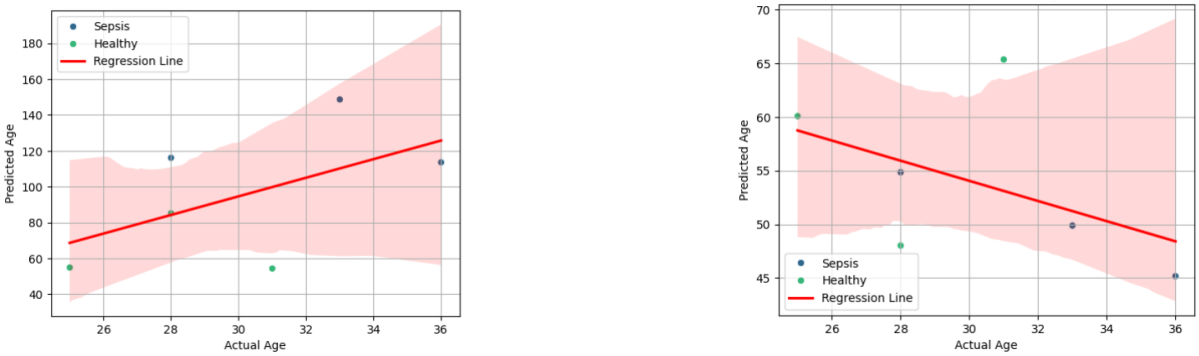

Figure 7: *Scatterplot of Predicted Age vs Actual Age, with the red line representing the regression line, colored dots representing the data points and red shaded region representing 95% confidence level. The GTEx Clock validation for the Sepsis dataset shows poor performance using both quantilized (left) and TPM values (right), indicating a negative correlation when TPM values are used.*

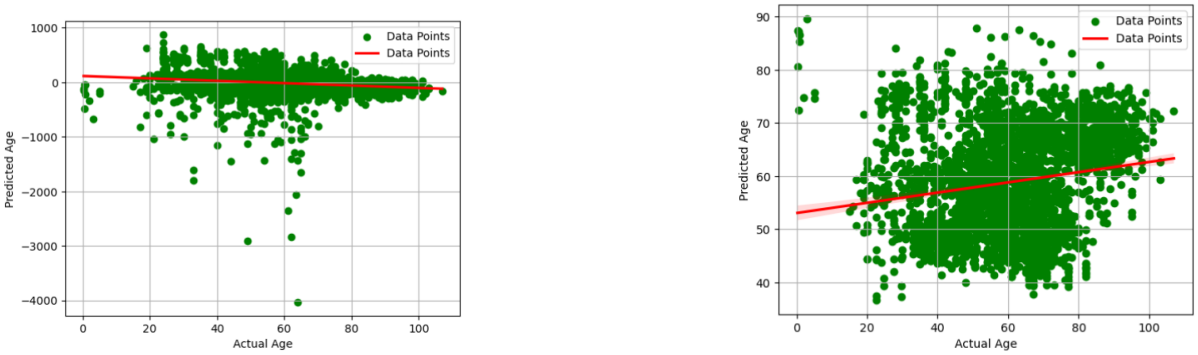

Figure 8: *Scatterplot of Predicted Age vs Actual Age, with the red line representing the regression line and green dots representing the data points. The GTEx Clock validation for the Shokhirev dataset shows poor performance using both quantilized (left) and TPM values (right), indicating a negative correlation when quantilized values are used.*

5 Discussion

Currently, there is no single model capable of accurately predicting the biological age of various organisms across different strains, treatments, and conditions [7]. In our study, we employed the quantilization method for gene expression data, which enhances the accuracy of biological age predictions. By training these quantilized values alongside feature selection, we can achieve better predictive accuracy. Additionally, the results obtained were consistent when using both linear regression and neural networks.

Using PCA for dimensionality reduction and pattern identification in large datasets offers significant advantages. However, caution is necessary when interpreting principal components, as they can be sensitive to outliers and threshold effects, making them less effective for isolating distinct pathways. This challenge of balancing interpretability and efficiency is common [4]. Although quantization and feature selection can improve results, the accuracy remains problematic. Therefore, careful fine-tuning is essential when selecting PCs and training the model.

The validation of the GTEx clock indicates that accurately predicting unseen data is challenging. This difficulty may be due to several factors, including the limited amount of training data, the quality of the data (expression values), and the significant impact of batch effects on accuracy.

Aging is a complex process that involves multiple systems. There is a need for specific evidence to confirm that the transcriptomic changes observed in biological clocks are driving the aging process. Additionally, there is uncertainty regarding how deviations between predicted and actual age contribute to biological age or prediction error. Clocks that are trained on small samples are likely to be influenced by confounding factors related to cell composition. Furthermore, the robustness of these clocks across environments, populations, and tissues remains unknown [2].

To address these challenges, it is important to clarify for those outside the epigenomics field that epigenetic observations typically serve as biomarkers of aging. Large studies, in addition to increasingly focused research on tissue- and disease-specific clocks

and cell type-specific information, are necessary. Additionally, it is essential to assess variability by analyzing large, diverse, and well-powered datasets across a range of tissues that are likely involved [2].

6 Bibliography

References

[1] Abdullah S. Al-Jawarneh, Mohd Tahir Ismail, and Ahmad M. Awajan. Elastic net regression and empirical mode decomposition for enhancing the accuracy of the model selection. *International Journal of Mathematical, Engineering and Management Sciences*, 6:564–583, 4 2021.

[2] Christopher G. Bell, Robert Lowe, Peter D. Adams, Andrea A. Baccarelli, Stephan Beck, Jordana T. Bell, Brock C. Christensen, Vadim N. Gladyshev, Bastiaan T. Heijmans, Steve Horvath, Trey Ideker, Jean Pierre J. Issa, Karl T. Kelsey, Riccardo E. Marioni, Wolf Reik, Caroline L. Relton, Leonard C. Schalkwyk, Andrew E. Teschendorff, Wolfgang Wagner, Kang Zhang, and Vardhman K. Rakyan. Dna methylation aging clocks: Challenges and recommendations, 11 2019.

[3] Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan. Random forests: From early developments to recent advancements. *Systems Science and Control Engineering*, 2:602–609, 2014.

[4] Sheng Fong, Kamil Pabis, Djakim Latumalea, Nomuundari Dugersuren, Maximilian Unfried, Nicholas Tolwinski, Brian Kennedy, and Jan Gruber. Principal component-based clinical aging clocks identify signatures of healthy aging and targets for clinical intervention. *Nature Aging*, 4:1137–1152, 8 2024.

[5] Ricardo Gutierrez-Osuna. Pattern analysis for machine olfaction: A review, 2002.

[6] Nicholas Holzscheck, Cassandra Falckenhayn, Jörn Söhle, Boris Kristof, Ralf Siegner, André Werner, Janka Schössow, Clemens Jürgens, Henry Völzke, Horst Wenck, Marc Winnefeld, Elke Grönniger, and Lars Kaderali. Modeling transcriptomic age using knowledge-primed artificial neural networks. *npj Aging and Mechanisms of Disease*, 7, 12 2021.

[7] David H. Meyer and Björn Schumacher. Bit age: A transcriptome-based aging clock near the theoretical limit of accuracy. *Aging Cell*, 20, 3 2021.

[8] Srikari Rallabandi and Sourav Yadav. Comparing ridge and logistic regression with neural networks. pages 1065–1073. Institute of Electrical and Electronics Engineers Inc., 2023.

[9] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics, 10 2007.

[10] Maxim N. Shokhirev and Adiv A. Johnson. Modeling the human aging transcriptome across tissues, health status, and sex. *Aging Cell*, 20, 1 2021.

[11] Xiaogang Su, Xin Yan, and Chih Ling Tsai. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4:275–294, 5 2012.

[12] Dai Xiao-Kang, Ding Zhen-Xing, Tan Yuan-Yuan, Bao Hua-Rui, Wang Dong-Yao, and Hong Zhang. Neutrophils inhibit cd8+ t cells immune response by arginase-1 signaling in patients with sepsis. *World Journal of Emergency Medicine*, 13:136–143, 2022.