

AGEOptimizer: An Optimization Framework for Improving Transcriptomic Age Prediction Across Human Tissues

Andrew Shin¹

Received May 7, 2026

Accepted August 21, 2026

Electronic access September 30, 2026

Aging is a biological decline in the body, gradually impairing cellular function and raising susceptibility to disease. While age-prediction models can aid in the diagnosis and treatment of different medical conditions, many are hindered by poor data quality, lack of generalizability for a wide variety of samples, and inherent biases, reducing the confidence of applications in real-world situations. To address this issue, AGEOptimizer seeks to enhance the robustness of genomic age prediction through improved data preprocessing, progressive hyperparameter tuning, and rigorous model selection. Compared with the baseline XGBoost model evaluated in this study, AGEOptimizer reduced overall mean absolute error (MAE) by 12.37% and reduced MAE by 46.17% in the best-performing tissues. In the GTEx dataset, AGEOptimizer achieved its lowest prediction error in the brain amygdala, substantia nigra, and cortex. While AGEOptimizer predicted healthy tissue age with low error, predicted ages for low-grade glioma (LGG) samples were approximately 26% greater than their chronological ages. This difference suggests that LGG samples may share transcriptomic features with healthy tissues from older individuals. Overall, AGEOptimizer improved transcriptomic age prediction across the datasets examined in this study. Although additional validation using independent datasets is needed, these results suggest that systematic optimization can improve transcriptomic age-prediction performance.

Keywords: Biological Age, Machine Learning, GTEx, RNA-Seq, Hyperparameter Optimization, Low-Grade Glioma (LGG)

1 Introduction

Aging in humans is the temporal change in the body's functions to meet physiological demands¹. Thus, it is also the main risk factor in many prevalent diseases such as cancer, cardiovascular disease², and neurodegeneration^{3,4}. These conditions are classified as age-related diseases due to their strong association with transcriptomic changes and increasing incidence with age³⁻⁵. Therefore, understanding how aging contributes to disease risk has important implications for both biological research and clinical medicine.

The widespread use of high-throughput RNA sequencing has produced large transcriptomic datasets from many human tissues⁶. These datasets have enabled researchers to study age-related changes in gene expression and develop models that predict biological age from transcriptomic profiles rather than chronological age alone⁷⁻¹¹. Early transcriptomic aging studies demonstrated that gene expression profiles could be used to estimate chronological age across multiple tissues⁹⁻¹¹, while more recent studies have improved prediction accuracy using machine-learning methods trained on large transcriptomic datasets^{12,13}.

Predicting biological age from transcriptomic data remains challenging. Gene expression datasets contain tens of thousands of genes, and expression patterns vary across tissues. Differences in preprocessing methods, feature-selection approaches¹⁴⁻¹⁶, and optimization procedures¹⁷⁻¹⁹ also make it difficult to compare results across studies. In addition, the high dimensionality of transcriptomic data and variation between datasets can reduce model performance and limit how well models generalize to new data²⁰⁻²³.

Although previous studies have developed machine-learning models for transcriptomic age prediction^{12,13}, most have focused on improving the performance of a single prediction model. Other studies have examined feature selection separately¹⁴⁻¹⁶. Model interpretation²⁴ and optimization¹⁷⁻¹⁹ have likewise been investigated independently. Few studies have evaluated how combining standardized preprocessing, systematic model selection, and staged hyperparameter tuning influences transcriptomic age-prediction performance across multiple machine-learning algorithms¹⁷⁻¹⁹.

To address this, we developed AGEOptimizer, an optimization framework for transcriptomic age prediction. Figure 1 summarizes the overall workflow of the study. GTEx RNA-seq data were first preprocessed, and several machine-learning

¹ The harker school, California, USA

models were compared to identify the best-performing approach. Following staged hyperparameter optimization, the final model was evaluated across multiple tissue types and then applied to low-grade glioma (LGG) samples to compare predicted biological age with chronological age.

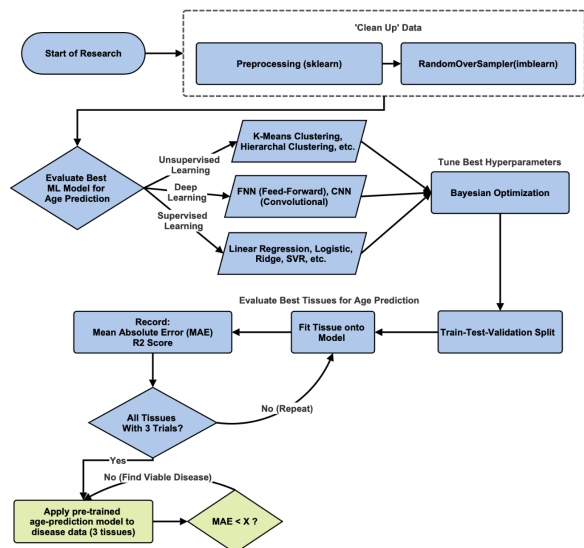


Figure. 1 Overview of the AGEOptimizer study design and analysis workflow.

Methods

Data Collection and Preprocessing

Gene expression data were obtained from the GTEx v10 repository and included 56,304 genes across multiple healthy human tissue types. For disease analysis, RNA-sequencing data from 506 Low-Grade Glioma (LGG) patient samples were obtained from The Cancer Genome Atlas (TCGA)²⁵. Only genes shared between the GTEx and TCGA datasets (approximately 36,000 genes) were retained for subsequent analyses.

The age-prediction model was developed exclusively using GTEx samples. The GTEx dataset was randomly divided at the sample level into training (70%), validation (15%), and test (15%) sets, while the LGG samples were kept separate and used only after model development had been completed.

To normalize the data, gene expression values were log-transformed using $\log_2(x + 0.001)$. The data were then feature-scaled before model training. Keeping the LGG dataset separate throughout model development helped prevent information leakage and ensured that disease samples did not influence model training or optimization.

Model Development

Several machine-learning algorithms with different levels of complexity, including Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, a feedforward Neural Network, and XGBoost, were compared to determine which framework performed best for transcriptomic age prediction. All models were trained using the same preprocessed GTEx dataset and identical training, validation, and test splits. Model performance was evaluated using mean absolute error (MAE). Among the models tested, XGBoost produced the lowest validation MAE and was selected for further optimization.

To further refine the selected model, a three-stage hyperparameter tuning process was employed. In the first stage, model complexity parameters (`max_depth` and `min_child_weight`) were adjusted. In the second stage, regularization and subsampling parameters (`subsample` and `colsample_bytree`) were tuned. Finally, in the third stage, the learning parameters (`n_estimators` and `eta`) were optimized. Dividing the hyperparameter search into stages reduced the search space while allowing improvements from earlier stages to be carried forward into later ones.

Hyperparameter tuning was performed using the Optuna framework¹⁹, with validation MAE used as the optimization objective. Early stopping was employed during hyperparameter optimization to reduce overfitting. The parameter combination that achieved the lowest validation MAE was selected for the final model. The final optimized model was then evaluated on the held-out test set before being applied to tissue-specific and LGG analyses.

Benchmarking

To evaluate performance, AGEOptimizer was compared with a baseline XGBoost model trained using the same dataset and data split but with default hyperparameters rather than the staged optimization procedure. Model performance was evaluated using mean absolute error (MAE), which measures the average difference between predicted biological age and chronological age. Lower MAE values indicate more accurate age prediction.

Tissue-Specific Performance Analysis

To identify the tissues in which the model performed most accurately, MAE values were compared across all GTEx tissue types. Tissues with the lowest MAE were considered to have the lowest prediction error.

LGG Disease Analysis

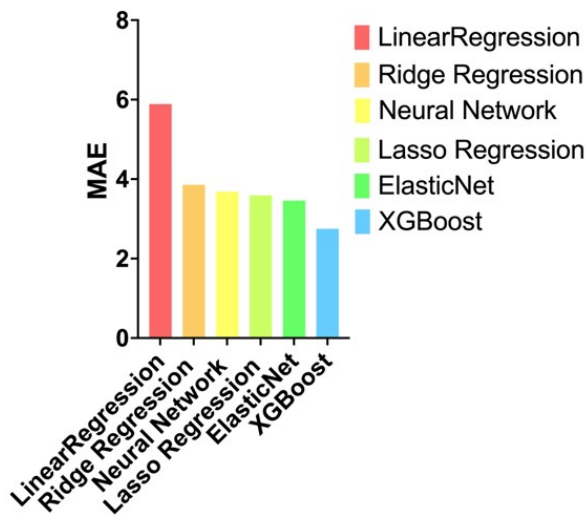
For the disease analysis, AGEOptimizer was applied to gene expression data from LGG samples after the model had been

117 developed using only GTEx data. Predicted biological ages
118 for the LGG samples were compared with those of healthy
119 GTEx brain tissues to determine whether the LGG samples
120 consistently showed higher predicted biological ages than the
121 healthy reference tissues.

122 **Results**

123 **Model Development and Hyperparameter Tuning**

124 Multiple machine-learning models, including Linear Regression,
125 Ridge Regression, Lasso Regression, Elastic Net, a feed-
126 forward Neural Network, and XGBoost, were evaluated using
127 the same preprocessed GTEx dataset and identical training,
128 validation, and test splits. Each model was first trained using
129 its default parameters, and performance was assessed using
130 mean absolute error (MAE). Among the models evaluated,
131 XGBoost had the lowest validation MAE and was selected for
132 further optimization (Figure 2).

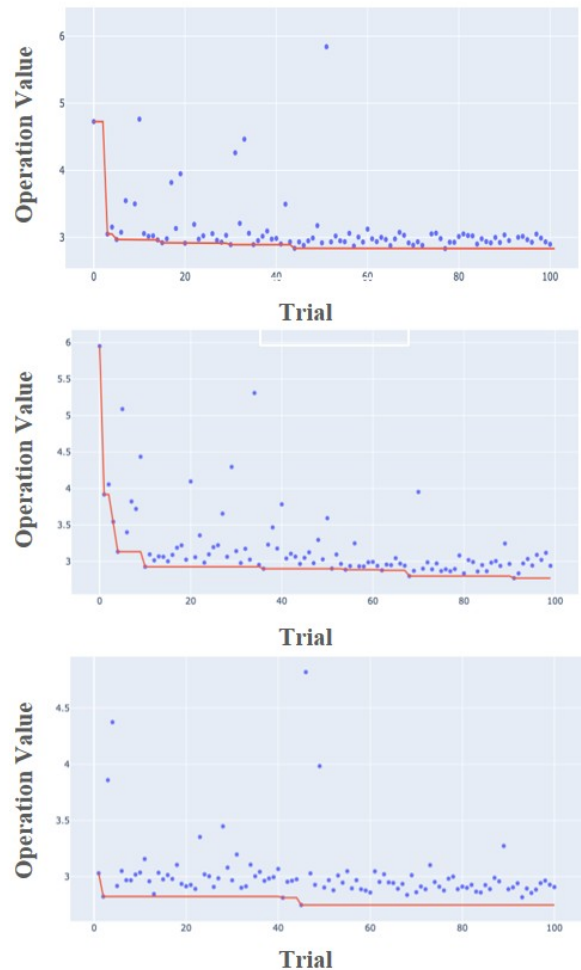


133 **Figure. 2** Comparison of Model Performance / Bar graph comparing
134 the mean absolute error (MAE) of multiple machine-learning models
135 trained using their default hyperparameters. All models were trained
136 using the same training, validation, and test splits, and model performance
137 was evaluated on the held-out test set using MAE.

138 Hyperparameter tuning was performed in three stages. The
139 largest reduction in validation MAE occurred during the first
stage, while the second and third stages produced smaller
improvements as performance began to level off (Figure 3). Validation
MAE decreased from 3.136 for the baseline XGBoost model to 2.927
after the first stage, 2.879 after the second stage, and 2.748 after the
final stage (Table 1). Overall, staged

hyperparameter tuning improved model performance and resulted in the final AGEOptimizer model.

The optimized XGBoost model had the lowest validation MAE and was used for all subsequent analyses, including benchmarking, tissue-specific performance analysis, and LGG disease analysis.



140 **Figure. 3** Hyperparameter Tuning / Mean absolute error (MAE)
141 across sequential hyperparameter trials for three tuning groups: (top)
142 max_depth and min_child_weight, (middle) subsample and colsam-
143 ple_bytree, and (bottom) n_estimators and eta (learning rate). Blue
144 points show the validation MAE for each XGBoost trial, and the red
145 curve indicates the best (lowest) MAE achieved up to that trial (best-
so-far). Rapid early gains are followed by convergence as additional
trials yield diminishing improvements.

Prediction Accuracy and Tissue-Specific Performance

On the held-out test set, AGEOptimizer reduced overall mean absolute error (MAE) by 12.37% compared with the baseline

Table 1 Hyperparameter tuning / Progressive hyperparameter optimization of the XGBoost model. Hyperparameters were optimized in three sequential stages using Optuna. At each stage, the parameter combination producing the lowest validation mean absolute error (MAE) was carried forward to the next optimization stage.

Optimization Stage	Hyperparameter	Search Range	Selected Value	Validation MAE
Baseline	Default XGBoost	Default	Default	3.136
Stage 1	max_depth	3–10	5	2.927
	min_child_weight	1–300	154.82	
Stage 2	subsample	0.5–1.0	0.85	2.879
	colsample_bytree	0.5–1.0	0.68	
Stage 3	n_estimators	100–1500	1164	2.748
	eta	0.01–0.30	0.101	

XGBoost model and reduced MAE by 46.17% in the best-performing tissues (Figure 4A). In the tissue-specific analysis, several brain tissues, including the amygdala, substantia nigra, and cortex, showed the lowest MAE values. These tissues therefore showed the lowest prediction error under the AGEOptimizer framework (Figure 4B).

Biological Age Prediction in LGG

AGEOptimizer predicted healthy GTEx tissue age with an average error of approximately 7%. By comparison, predicted biological ages for the low-grade glioma (LGG) samples were, on average, approximately 26% higher than their chronological ages (Figure 5). This observation suggests that LGG samples differ from healthy tissues in ways that result in higher predicted biological ages under the AGEOptimizer model.

Discussion

In this study, we developed AGEOptimizer to improve transcriptomic age prediction by combining standardized preprocessing, progressive hyperparameter tuning, and systematic model selection. Compared with the baseline model, the optimized framework reduced prediction error throughout the benchmarking analyses. Most of the improvement occurred during the early stages of hyperparameter tuning, while later stages produced smaller gains as performance began to level off. Overall, these findings suggest that systematic optimization can improve transcriptomic age prediction beyond the use of default model settings^{17–19}.

Prediction accuracy also varied across tissue types. AGEOptimizer achieved the lowest mean absolute error (MAE) in several brain tissues, particularly the amygdala, substantia nigra, and cortex. These tissues had the lowest prediction errors among those included in the GTEx dataset. However, MAE can also be influenced by factors such as sample

size, age distribution, tissue heterogeneity, RNA quality, and other technical differences^{20–23}. Therefore, these findings reflect model performance within the GTEx dataset and should not be interpreted as evidence that these tissues are inherently more informative for biological aging.

When AGEOptimizer was applied to low-grade glioma (LGG) samples, the predicted biological ages were, on average, higher than the corresponding chronological ages. This suggests that LGG samples share some transcriptomic features with older healthy tissues^{5,11,12}. However, this finding should not be interpreted as evidence that aging causes LGG or that LGG undergoes accelerated biological aging. Because the model was trained only on GTEx data and applied independently to the TCGA-LGG dataset, the observed differences should be interpreted as an association rather than a causal relationship.

This study also has several limitations. The model was developed and evaluated primarily using GTEx data, so additional validation with independent transcriptomic datasets will be needed to determine how well the approach generalizes. Although GTEx and TCGA-LGG were kept separate during model development, differences in sequencing protocols, sample processing, and other technical factors may still contribute to differences between healthy and disease samples^{20–23}. In addition, while hyperparameter tuning improved model performance, future studies should include donor-level validation and external test datasets to provide a stronger assessment of generalizability.

Future studies should evaluate AGEOptimizer using additional cohorts and tissue types to determine whether these findings can be reproduced in independent datasets. Comparing AGEOptimizer with additional machine-learning models and evaluating its performance across different age groups would also provide a more comprehensive assessment of its robustness. For the LGG analysis, future work should account for factors such as tumor purity, cell-type composition, se-

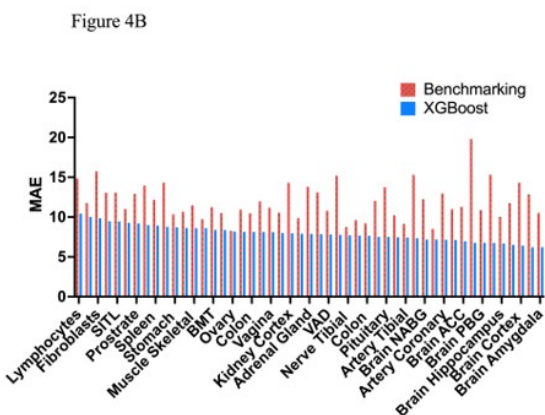
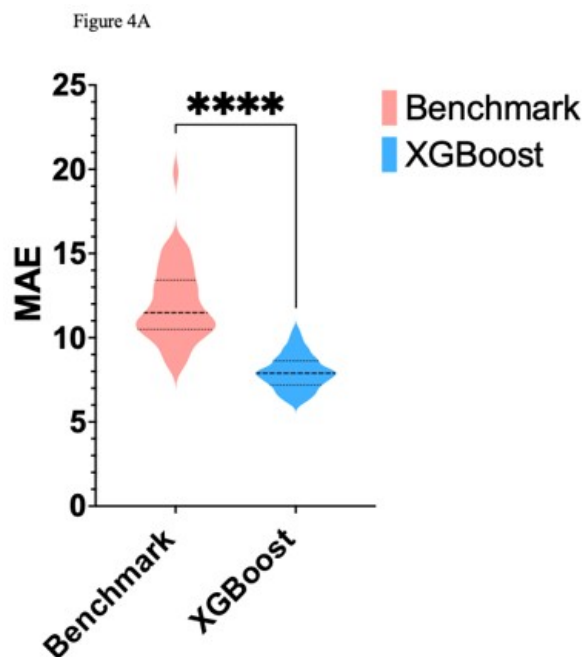


Figure. 4 Prediction Accuracy and Tissue-Specific Performance / (A) Overall model performance comparison between the baseline XGBoost model and AGEOptimizer, reported as mean absolute error (MAE). AGEOptimizer reduced overall MAE by 12.37% and reduced MAE by 46.17% in the best-performing tissues relative to the baseline XGBoost model. Groups were compared using an unpaired t-test with Welch's correction (**** $p < 0.0001$). (B) Tissue-specific MAE across GTEx tissue types using AGEOptimizer. Lower MAE indicates higher age-prediction accuracy. Brain amygdala, substantia nigra, and cortex showed among the lowest MAE values, indicating that AGEOptimizer performed most accurately in these tissues. Bars represent mean $\pm$ SEM where applicable, and MAE values were calculated using the held-out test samples.

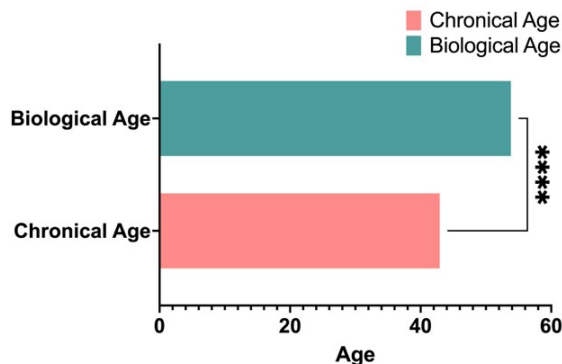


Figure. 5 AGEOptimizer Prediction of Biological Age in LGG / Predicted biological ages generated by AGEOptimizer were compared with chronological ages for healthy GTEx samples and LGG samples. Healthy tissues were predicted within approximately 7% error, whereas predicted biological ages for LGG samples were approximately 26% higher than their chronological ages. Group differences in predicted age were evaluated using an unpaired t-test with Welch's correction (**** $p < 0.0001$).

quencing batch, and immune infiltration to determine whether the higher predicted ages remain after these variables are considered.

Overall, AGEOptimizer improved transcriptomic age-prediction accuracy within the datasets examined in this study. Although additional validation is still needed, these findings suggest that systematic optimization can improve prediction performance and provide a reproducible framework for transcriptomic age prediction^{18,19}.

In conclusion, AGEOptimizer improved transcriptomic age-prediction accuracy through systematic optimization. The model performed best in several brain tissues within the GTEx dataset and predicted higher biological ages for LGG samples than their chronological ages. Although these findings require further validation, they suggest that systematic optimization can improve transcriptomic age prediction and provide a foundation for future studies of age-related transcriptomic changes.

References

- 1 T. Niccoli and L. Partridge, *Ageing as a risk factor for disease*, 2012.
- 2 B. J. North and D. A. Sinclair, *The intersection between aging and cardiovascular disease*, 2012.
- 3 T. Wyss-Coray, *Ageing, neurodegeneration and brain rejuvenation*, 2016.
- 4 C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano and G. Kroemer, *The hallmarks of aging*, 2013.
- 5 J. Wang *et al.*, *Accurate estimation of biological age and its application in disease prediction using a multimodal image transformer system*, 2024.
- 6 GTEx Consortium, *The GTEx Consortium atlas of genetic regulatory effects across human tissues*, 2020.
- 7 S. Horvath, *DNA methylation age of human tissues and cell types*, 2013.

-
- 247 8 J. Jylhävä, N. L. Pedersen and S. Hägg, *Biological age predictors*, 2017.
- 248 9 M. J. Peters, R. Joehanes, L. C. Pilling *et al.*, *The transcriptional land-*
- 249 *scape of age in human peripheral blood*, 2015.
- 250 10 J. G. Fleischer, R. Schulte, H. H. Tsai *et al.*, *Predicting age from the*
- 251 *transcriptome of human dermal fibroblasts*, 2018.
- 252 11 X. Ren and P. F. Kuan, *RNAAgeCalc: A multi-tissue transcriptional age*
- 253 *calculator*, 2020.
- 254 12 P. Mamoshina *et al.*, *Machine learning on human muscle transcriptomic*
- 255 *data for biomarker discovery and tissue-specific drug target identifica-*
- 256 *tion*, 2018.
- 257 13 A. Zhavoronkov, P. Mamoshina, Q. Vanhaelen *et al.*, *Deep aging clocks:*
- 258 *the emergence of AI-based biomarkers of aging and longevity*, 2019.
- 259 14 Y. Saeys, I. Inza and P. Larrañaga, *A review of feature selection techniques*
- 260 *in bioinformatics*, 2007.
- 261 15 I. Guyon and A. Elisseeff, *An introduction to variable and feature selec-*
- 262 *tion*, 2003.
- 263 16 T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical*
- 264 *Learning: Data Mining, Inference, and Prediction*, 2009.
- 265 17 T. Chen and C. Guestrin, *XGBoost: A scalable tree boosting system*, 2016.
- 266 18 P. Probst, M. N. Wright and A. L. Boulesteix, *Hyperparameters and tun-*
- 267 *ing strategies for random forest*, 2019.
- 268 19 T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, *Optuna: A next-*
- 269 *generation hyperparameter optimization framework*, 2019.
- 270 20 J. T. Leek, R. B. Scharpf, H. C. Bravo *et al.*, *Tackling the widespread and*
- 271 *critical impact of batch effects in high-throughput data*, 2010.
- 272 21 W. E. Johnson, C. Li and A. Rabinovic, *Adjusting batch effects in mi-*
- 273 *croarray expression data using empirical Bayes methods*, 2007.
- 274 22 W. W. B. Goh, W. Wang and L. Wong, *Why batch effects matter in omics*
- 275 *data, and how to avoid them*, 2017.
- 276 23 J. T. Leek and J. D. Storey, *Capturing heterogeneity in gene expression*
- 277 *studies by surrogate variable analysis*, 2007.
- 278 24 S. M. Lundberg and S.-I. Lee, *A unified approach to interpreting model*
- 279 *predictions*, 2017.
- 280 25 The Cancer Genome Atlas Research Network, *Comprehensive, Integra-*
- 281 *tive Genomic Analysis of Diffuse Lower-Grade Gliomas*, 2015.