Received: 12 April 2022 | Revised: 25 May 2022 | Accepted: 27 June 2022
DOI: 10.1002/hbm.26010

TECHNICAL REPORT

WILEY

Systematic evaluation of machine learning algorithms for neuroanatomically-based age prediction in youth

Amirhossein Modabbernia[1] | Heather C. Whalley[2] | David C. Glahn[3] | Paul M. Thompson[4] | Rene S. Kahn[1] | Sophia Frangou[1,5]

[1]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA

[2]Division of Psychiatry, University of Edinburgh, Kennedy Tower, Royal Edinburgh Hospital, Edinburgh, UK

[3]Boston Children's Hospital and Harvard Medical School, Boston, Massachusetts, USA

[4]Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Los Angeles, California, USA

[5]Department of Psychiatry, Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, British Columbia, Canada

Correspondence
Sophia Frangou, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, 1 Gustave L Levy Place, New York, NY 10029, USA.
Email: sophia.frangou@mssm.edu

Funding information
Icahn School of Medicine at Mount Sinai; National Institute of Mental Health, Grant/Award Numbers: R01-MH113619, T32-MH122394; National Institutes of Health, Grant/Award Numbers: U01DA041025, U01DA041093, U24DA041147, U24DA041123, U01DA041174, U01DA041156, U01DA041148, U01DA041134, U01DA041120, U01DA041117, U01DA041106, U01DA041089, U01DA041048, U01DA041028, U01DA041022

Abstract

Application of machine learning (ML) algorithms to structural magnetic resonance imaging (sMRI) data has yielded behaviorally meaningful estimates of the biological age of the brain (brain-age). The choice of the ML approach in estimating brain-age in youth is important because age-related brain changes in this age-group are dynamic. However, the comparative performance of the available ML algorithms has not been systematically appraised. To address this gap, the present study evaluated the accuracy (mean absolute error [MAE]) and computational efficiency of 21 machine learning algorithms using sMRI data from 2105 typically developing individuals aged 5–22 years from five cohorts. The trained models were then tested in two independent holdout datasets, one comprising 4078 individuals aged 9–10 years and another comprising 594 individuals aged 5–21 years. The algorithms encompassed parametric and nonparametric, Bayesian, linear and nonlinear, tree-based, and kernel-based models. Sensitivity analyses were performed for parcellation scheme, number of neuroimaging input features, number of cross-validation folds, number of extreme outliers, and sample size. Tree-based models and algorithms with a nonlinear kernel performed comparably well, with the latter being especially computationally efficient. Extreme Gradient Boosting (MAE of 1.49 years), Random Forest Regression (MAE of 1.58 years), and Support Vector Regression (SVR) with Radial Basis Function (RBF) Kernel (MAE of 1.64 years) emerged as the three most accurate models. Linear algorithms, with the exception of Elastic Net Regression, performed poorly. Findings of the present study could be used as a guide for optimizing methodology when quantifying brain-age in youth.

KEYWORDS
brain age, development, machine learning, neuroimaging, youth

1 | INTRODUCTION

Brain development involves highly organized multistep processes (Tau & Peterson, 2010) that lead to the emergence of adult levels of cognitive and behavioral competency (Paus, 2005; Spear, 2000). Brain development involves numerous cellular and noncellular events (Tau & Peterson, 2010), which are below the resolution of magnetic resonance imaging (MRI) but underpin morphological changes in brain organization that can be captured using structural MRI (sMRI) techniques. Multiple studies have shown that the volume of subcortical structures typically peaks in late childhood and adolescence and decreases thereafter (Dima et al., 2021; Raznahan et al., 2014). Cortical thickness shows a steep reduction in late childhood and adolescence that continues at a slower rate throughout adult life (Frangou et al., 2021; Wierenga et al., 2020). Cortical surface area expands during childhood and most of adolescence showing gradual decrements thereafter (Fjell et al., 2015; Tamnes et al., 2017). These age-related changes demonstrate marked inter-regional and inter-individual variation (Mills et al., 2021; Wierenga et al., 2020).

Machine learning (ML) algorithms applied to sMRI data can harness the multidimensional nature of age-related brain changes at the individual-level to predict age, as a proxy for the biological age of the brain (i.e., brain-age). The difference between brain-age and chronological age is referred to here as brain-age-gap-estimation (BrainAGE; Franke & Gaser, 2019), which is equivalent to terms such as brain-predicted-age-difference (brainPAD; Luna et al., 2021), brain-age-gap (BAG; Anatürk et al., 2021), and brain-age delta (Beheshti et al., 2019) used in other studies. In adults, higher brain-age relative to chronological age (i.e., higher BrainAGE) has been associated with adverse physical (Cole et al., 2018), cognitive (Anatürk et al., 2021; Boyle et al., 2021; Elliott et al., 2019) and mental health phenotypes (Kaufmann et al., 2019; Lee et al., 2021). By contrast, in children and adolescents higher BrainAGE has been associated with better cognitive test performance (Boyle et al., 2021; Erus et al., 2015; Luna et al., 2021) while associations with clinical phenotypes show a more complex pattern which may depend on the nature of the phenotype and/or the developmental stage of the sample (Chung et al., 2018; Luna et al., 2021). These findings underscore the importance of accuracy in brain-based age-prediction in youth, as childhood and adolescence are periods of dynamic brain re-organization.

Therefore, the current study focuses exclusively on the evaluation of the methods used to compute brain-age in youth from sMRI data as a foundation for guiding study design into its functional significance. We have previously shown that age prediction from sMRI data in adults is influenced by the choice of algorithm (Lee et al., 2021). Here addressed this knowledge gap in youth because with few exceptions (Ball et al., 2021; Brouwer et al., 2021; Lee et al., 2021; Luna et al., 2021), studies on brain-age prediction in this population have typically employed a single ML algorithm, most commonly relevance vector regression (RVR), Gaussian process regression (GPR), or support vector regression (SVR; Cole et al., 2018; Franke et al., 2010; Franke et al., 2012; Gaser et al., 2013; Liem et al., 2017; Valizadeh et al., 2017). We systematically evaluated the performance of 21 ML algorithms applied to sMRI data from youth from five different cohorts and then tested their performance in two independent samples. The algorithms encompassed parametric and nonparametric, Bayesian, linear and nonlinear, tree-based, and kernel-based models. These algorithms were selected to include those that are commonly used in brain-age prediction studies as well representative examples of a range of algorithms that provide reasonable and potentially better alternatives. We evaluated the ML methods for accuracy and for their sensitivity to key parameters known to affect model performance pertaining to parcellation scheme, number of neuroimaging input features (Valizadeh et al., 2017), number of cross-validation folds, sample size (by resampling the available data), and number of extreme outliers. Our prediction was that nonlinear kernel-based and ensemble algorithms would outperform other algorithms because they are theoretically better at handling collinear data and non-linear relationships with age and, in the case of ensemble algorithms, they improve predictive performance by aggregating results from multiple nodes. Collectively, these analyses may assist in optimizing the design of future investigations on brain predicted age in youth.

2 | METHODS

2.1 | Samples

We used T1-weighted scans from six separate cohorts: Autism Brain Imaging Data Exchange (ABIDE; Di Martino et al., 2014; Di Martino et al., 2017); ABIDE II (Di Martino et al., 2017); ADHD-200 (ADHD-200 Consortium, 2012); Human Connectome Project Development (HCP-D; Harms et al., 2018); Child Mind Institute (CMI; Alexander et al., 2017), Adolescent Brain Cognitive Development (ABCD; Alexander et al., 2017; Garavan et al., 2018), Pediatric Imaging, Neurocognition, and Genetics (PING) Data Repository (Jernigan et al., 2016; details of the cohorts in the Data S1 and Table S1). Data collection for these cohorts was conducted at multiple independent sites located in eight countries: The United States, Germany, Ireland, Belgium, the Netherlands, Switzerland, China, and France. Only psychiatrically healthy participants with high-quality anatomical brain scans from each cohort were included (details of quality assurance in Data S1). Data from five cohorts (total $n = 2105$; 41% female, age-range: 9–10 years; Figure S1) were used to train the ML algorithms (training set) while data from the ABCD sample ($n = 4078$; 52% female; age range: 9–10 years) and the PING sample ($n = 594$; female $= 49.6\%$; age-range: 5–21 years) comprised the independent hold-out test-sets.

2.2 | Image processing

Across all cohorts, more than 98% of the participants were scanned using 3-T MRI machines; Siemens Prisma and Trio Tim scanners were each used for 31% of the participants of the total training sample (Table S1). The T1-weighted images were downloaded from the respective cohort repositories and processed at the Icahn School of Medicine at Mount Sinai (ISMMS) using identical pipelines. Image processing was implemented using standard pipelines in the FreeSurfer 7.1.0 software to generate cortical parcels based on the Schaefer scheme (Schaefer et al., 2018) by projecting the parcellation onto individual surface space (https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal/Parcellations/project_to_individual) and using the *mri_anatomical_stats* function to extract cortical values. We used the 400-parcel resolution (i.e., 400 cortical thickness and 400 cortical surface area values; Figure S2) in the main analyses. The 400-parcellation scheme has been shown to have good stability, signal to noise ratio, and performance in different contexts, and correspondence to histology (Bryce et al., 2021; Valk et al., 2020). Participants with missing values on any parcellations were excluded, because it was assumed that the image quality was compromised; participants in the training dataset only were also excluded if more than 5% of their parcellation features had extreme values (details in Data S1). We did not exclude participants based on outlier values in the hold-out test sets but instead studied the effect of outliers on model performance.

2.3 | Algorithms for brain-based age prediction

We used the *caret* package (version 6.0.84) in R (version 3.5.3) to conduct the ML analyses because it interfaces with multiple ML packages and standardizes data preprocessing, model training and testing. Several of the regression algorithms evaluated can be extended to accommodate non-linear associations using kernel functions. A kernel function transforms the original non-linear data into a higher-dimensional space in which they can become linearly separable. The kernelized models evaluated here incorporated polynomial and radial basis function (RBF) kernels. The former adds features using the polynomial combinations of the original data up to a specified degree and the latter adds features using the distance of the original data from specified reference values. Below we describe the 15 base models, six of which have non-linear kernelized variations; together, they amount to 21 different algorithms:

1. Generalized linear model: This is a standard algorithm for regression that minimizes the sum of squared errors between the observed variables and predicted outcomes. Models have no tuning parameters and were implemented using the “glm” function.
2. Bayesian general linear model (Gelman et al., 2008): This is a linear regression model in which the outcome variable and the model parameters are assumed to be drawn from a probability distribution; it therefore provides estimates of model uncertainty. Models have no tuning parameters and were implemented using the “bayesglm” function.
3. Gaussian Processes Regression (Williams & Barber, 1998): This is a regression model that follows Bayesian principles. The covariance function here was defined by using either a linear, or a polynomial function or a RBF kernel as a prior. The polynomial kernels were tuned using degree and scale and the RBF kernels were tuned using the *sigma* parameter (the inverse kernel width parameter). Models were implemented using “gaussprRadial,” “gaussprLinear,” and “gaussprPoly.”
4. Independent Component Regression (Shao et al., 2006): This is a linear regression model in which components from a prior independent component analysis are used as the explanatory variables. The number of components was tuned, and the models were implemented using the “icr” function.
5. Principal Component Regression: This is a linear regression model in which components from a prior principal component analysis are used as the explanatory variables. The number of components was tuned, and the models were implemented using the “pcr” function.
6. Kernel Partial Least Squares Regression (Dayal & MacGregor, 1997): This is an extension of the partial least squares (PLS) regression which creates components by using the correlations between explanatory variables and outcome variables. The kernelized version used here (K-PLS) maps the data vector from the sample space to a higher-dimensional, Euclidean space; models were tuned for the number of components and implemented using the “kernelpls” function.
7. Sparse Partial Least Squares Regression (SPLS; Chun & Keleş, 2010): This is a different extension of PLS that reduces the number of explanatory variables (sparsity) through a least absolute shrinkage and selection operator (LASSO) approach. The models were tuned for the number of components, and *eta* (the sparsity parameter), and were implemented using the “spls” function.
8. Quantile Regression with least absolute shrinkage and selection operator (LASSO) Penalty (Wu & Liu, 2009): This algorithm models the relationship between explanatory variables and specific percentiles (or “quantiles”) of the outcome variable; in this variation, sparsity was introduced through the LASSO approach. The number of selected variables was tuned, and models were implemented with the “rqlasso” function.
9. Elastic Net Regression (Zou & Hastie, 2005): This is a linear regression that adds two penalties, LASSO regression (L1-norm) and ridge regression (L2-norm), in the loss function to encourage simpler models and avoid overfitting. Models were tuned for *lambda* (weight decay) and fraction of the full solution (equivalent to ordinary least squares) and were implemented using the “enet” function.
10. Boosted Generalized Additive Model (Bühlmann & Yu, 2003): This generalized additive model is fitted using a gradient-based boosting algorithm based on penalized B-splines. Overfitting was reduced by pruning the number of iterations using the optimal value of the Akaike Information Criterion. Models were implemented using the “gamboost” function.
11. Random Forest Regression (Breiman, 2001): This an ensemble machine learning method, which involves construction of multiple decision trees (i.e., forests) via bootstrap (bagging) and aggregates the predictions from these multiple trees to reduce the variance and improve the robustness and precision of the results. Models were implemented using the “rf” function and were tuned with regard to the number of trees.
12. Support Vector Regression (Cortes & Vapnik, 1995): Support Vector Regression (SVR) is characterized by the use of kernels, sparsity, and control of the margin of tolerance (epsilon; ε) and the number of support vectors (Awad & Khanna, 2015). It

identifies a symmetrical ε-insensitive region, called the ε-tube, which approaches the loss function as an optimization problem; the ε-value determines the width of the tube and maximization of the "flatness" aims to ensure that it contains most of the values in the training sample. Here flatness maximization was subject to the L2-norm penalty. In addition to the linear kernel, we also tested a version with polynomial and RBF kernels. The corresponding functions were "svmLinear3," "svmPoly," and "svmRadial." The regularization parameter (C) was used to optimize all models, while scale and degree were also considered in polynomial models and *sigma* for RBF models.

13. Relevance Vector Regression (Tipping, 2001): Relevance Vector Regression (RVR) is an extension of SVR embedded in a Bayesian framework. Its characteristic feature is that it imposes an explicit zero-mean Gaussian prior on the model parameters leading to a vector of independent hyperparameters that reduces the dataset. The behavior of the RVR is controlled by the type of kernel, which has to be chosen, while all other parameters are automatically estimated by the learning procedure itself. Here we used a linear, polynomial, or RBF kernel implemented with functions "rvmLinear," "rvmPoly," and "rvmRadial," respectively. The latter two kernels require tuning for scale and degree (polynomial) and for sigma (RBF).
14. Bayesian Regularized Neural Networks (Perez-Rodriguez et al., 2013): This is a version of the feedforward artificial neural network (ANN) architecture, in which robustness is improved through Bayesian regularization of the ANN parameters. The model includes two layers: the input layer—consisting of independent variables—and the hidden layer of *S* number of neurons. Models were implemented using the "brnn" function and tuned for the number of neurons.
15. Extreme Gradient Boosting (Chen & Guestrin, 2016): Extreme Gradient Boosting (XGBoost) is an ensemble decision-tree based gradient boosting algorithm that allows for modeling complex nonlinear relationships and interactions. The algorithm optimizes model performance through parallel (simultaneous) processing, regularization, tree pruning, optimal split (through a weighted quantile sketch algorithm), automatic missing data handling, and built-in cross-validation. Tuning parameters involved the number of boosting iterations; maximum tree depth; eta (shrinkage parameter); gamma (minimum loss reduction); subsample ratio of columns; minimum sum of instance weights, and column subsample percentage. Models were implemented using "xgbTree" function.

For clarity we refer to each algorithm by the name of the specific function used for its implementation.

Computational efficiency for each algorithm was assessed by recording the total Central Processing Unit (CPU) time, and the average and maximum memory usage. All models were run on the ISMMS high-performance computing cluster.

Several analytic steps were common to all algorithms. As there are known sex differences in the rate of age-related changes (Brouwer et al., 2021; Wierenga et al., 2019; Wierenga et al., 2020), models were separately trained for males and females. Hyperparameter tuning (when required) was performed in the combined training set ($n = 2105$), using a grid search in a fivefold cross-validation scheme across five repeats. In each cross-validation 80% of the training sample was used to train the model and 20% was used to test the model parameters. Subsequently, the model was re-trained on the whole training dataset using the optimal hyperparameters identified through cross-validation. Finally, the generalizability of the model was tested in two hold-out datasets (ABCD $n = 4078$ and PING $n = 594$).

The primary accuracy measure for each algorithm was the Mean Absolute Error (MAE) which represents the absolute difference between the neuroimaging-predicted age and the chronological age. For each algorithm, the abbreviation MAE_T refers to values obtained in the hold-out test dataset and MAE_cv refers to the mean cross-validation value in the training dataset. We also report two other commonly used accuracy measures: the Root Mean Square Error (RMSE), which is the standard deviation of the prediction errors, and the correlation between predicted and actual age. This correlation coefficient was not calculated in the ABCD data because of the narrow age range (<2 years). Based on these criteria we identified the three best performing algorithms which we evaluate further in the subsequent sections.

2.4 | Calculating BrainAGE and corrected BrainAGE for the three best performing algorithms

BrainAGE in each individual was calculated by subtracting the chronological age from the age predicted by each of the three best performing algorithms. Positive BrainAGE values indicate an older than expected brain-age for the given chronological age, and the opposite is the case for negative BrainAGE values. BrainAGE is typically overestimated in younger individuals and underestimated in older individuals. To counter this bias, multiple methods have been proposed (Beheshti et al., 2019; Cole et al., 2018). Here we used a robust approach introduced by Beheshti and colleagues (Beheshti et al., 2019), which relies on the slope (α) and intercept (β) of a linear regression model of BrainAGE against chronological age in the training set. This way an offset is calculated (as $\alpha\Omega + \beta$) and then subtracted from the estimated brain-age to yield a bias-free BrainAGE (Beheshti et al., 2019), hereafter referred to as "corrected BrainAGE" ($\text{BrainAGE}_\text{corr}$).

2.5 | Quantifying feature importance for age prediction in the three best performing algorithms

Estimates of the contribution of individual neuroimaging features to age prediction in each of the three best performing algorithms were obtained using Shapley Values (SV) implemented via the *fastshap* package Version 0.0.5 (Greenwell & Greenwell, 2020) in R. SVs derive from the cooperative game theory (Lundberg & Lee, 2017) and measure the contribution of each feature value in the model by abstracting away from the model specification. They accommodate non-linearity and have properties that make their interpretation intuitive. For example, the sum of all the SVs of a model is equal to the accuracy of the model and features with the same SV contribute equally to the model.

2.6 | Sensitivity and supplemental analyses in the three best performing algorithms

To test the effect of sex on model generalizability, we applied the parameters trained on one sex to the other and compared differences in BrainAGE using a two-sample Student's *t*-test. Sensitivity analyses focused on the parcellation scheme, number of input features, sample size and number of repeats, and cross-validation folds and outliers. Accordingly, we repeated the analyses using features from (a) the Desikan–Killiany (DK) atlas (*n* of features = 136); (b) the DK and subcortical Aseg atlas in FreeSurfer combined (*n* of features = 157); (c) the 400-parcel Schaefer atlas with Aseg atlas (*n* of features = 821); and (d) the 1000-parcel Schaefer atlas (*n* of features 2000). To test effect of sample size, the training dataset was randomly resampled with replacement in increments of 100, from 100 to 1500 (20 times each). Additionally, we conducted the same analyses using 10 repeats and 10 cross-validation folds. Finally, we tested the effect of number of extreme outliers (potential indicators of low-quality segmentation) on the model performance, by calculating the Spearman's correlation coefficient between the number of outliers and absolute error value among the subjects in the hold-out test sets. An outlier was defined as three median absolute deviation above or below the median for each brain region.

3 | RESULTS

3.1 | Algorithm performance for age prediction

Linear algorithms, with the exception of Elastic Net Regression, performed poorly while the XGBoost, RF regression and SVR with RBF kernel emerged as the three top performing models in males and females in cross-validation (Table S2) and in the combined hold-out datasets (i.e., ABCD and PING; Figure 1, Table 1). Despite nominal ranking of the algorithms, the top 10 algorithms performed comparably (maximum difference <0.5 years). The results of the statistical comparison in absolute error values between each pair of algorithms are presented in Figures S3 and S4. The median correlation coefficient between age predicted by the 21 different algorithms was 0.92 for males and 0.94 for females (Figure 2). The wider age-range of the PING dataset enabled examination of the association between observed and predicted age from the different algorithms which had a median correlation coefficient of 0.84 for males and 0.86 for females (Table 1, Figures 3 and S5).

3.2 | Computational speed and memory usage of each algorithm

The highest maximum memory usage was observed while training the Bayesian regularized neural networks, the boosted generalized additive model, and XGBoost algorithms (in that order). Highest average memory usage was seen with Bayesian regularized neural networks, the boosted generalized additive model, and the quantile regression with LASSO penalty. Bayesian regularized neural networks, XGBoost, and SVR with a polynomial kernel engaged CPU for the longest time. The generalized linear model, Kernel-partial least squares and principal component regression were the fastest algorithms with the lowest memory usage. Among the algorithms that performed best nonlinear kernelized versions had a favorable memory-computational speed profile (Table S3).

3.3 | BrainAGE and BrainAGE$_{corr}$ for the three best performing algorithms

In Table 2 we report the BrainAGE and BrainAGE$_{corr}$ derived from the three best performing algorithms. The corresponding values for all algorithms are shown in Table S4. The values presented refer to the models' performance in the ABCD and the PING samples using the optimized model parameters in the training phase. There was a significantly negative correlation between BrainAGE and chronological age but not for BrainAGE$_{corr}$, as this association was mitigated by applying age-bias correction (Table S4, Figures S6–S10).

3.4 | Feature importance in brain-age prediction in the three best performing algorithms

Feature importance, as inferred by their SVs, varied considerably across models specified with XGBoost, RF regression, and SVR with the RBF kernel (Figure 4, Table S5). The values presented refer to the models' performance in the training sample using the optimized model parameters. In RF regression, a few regions made very large contributions, with minimal contributions from other regions. In SVR with the RBF kernel, most features contributed to the model although the contribution of each feature was small. The profile of feature contributions in XGBoost was intermediate between the other two algorithms.

3.5 | Sensitivity and supplemental analyses for the three best performing algorithms

3.5.1 | Sex

Application of parameters from models trained on males to the entire sample, yielded marginally higher BrainAGE values for females than

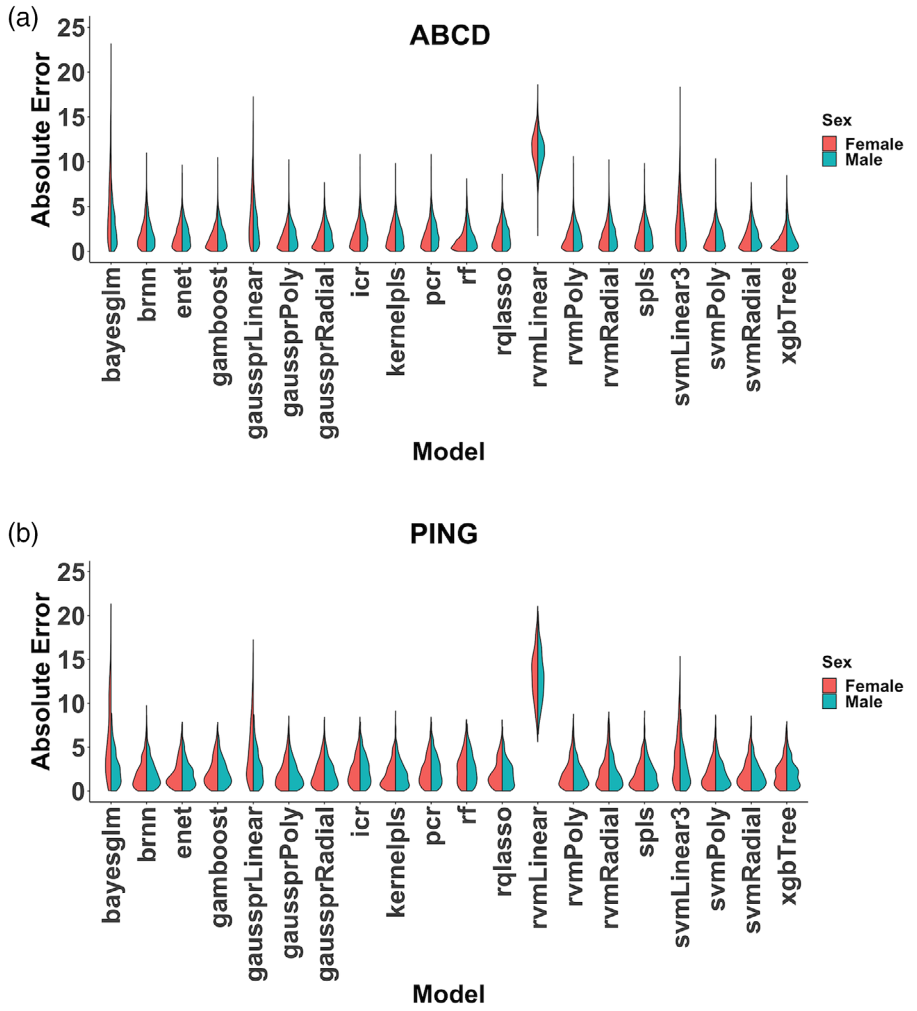

FIGURE 1 Absolute Mean Error of the 21 algorithms evaluated. The figure presents the model performance in males and females in the hold-out test sets: the Adolescent Brain Cognitive Development (ABCD) study (Panel a) and the Pediatric Imaging, Neurocognition, and Genetics Data Repository (PING) (Panel b). The different algorithms are referenced by the function used for their implementation. bayesglm, Bayesian Generalized Linear Model; brnn, Bayesian Regularized Neural Network; enet, Elastic Net Regression; gamboost, Generalized Additive Model with Boosting; gaussprLinear, Gaussian Processes Regression Linear; gaussprPoly, Gaussian Processes Regression Polynomial; gaussprRadial, Gaussian Processes Regression Radial; glm, Generalized Linear Model; icr, Independent Component Regression; kernelpls, Kernel Partial Least Squares; pcr, Principal Component Regression; rf, Random Forest; rqlasso, Quantile Regression with LASSO penalty; rvmLinear3, Relevance Vector Machine-Linear; rvmPoly, Relevance Vector Machine-Polynomial; rvmRadial, Relevance Vector Machine-Radial; spls, Sparse Partial Least Squares; svmeLinear3, Support Vector Regression-Linear; svmPoly, Support Vector Regression-Polynomial; svmRadial, Support Vector Regression-Radial; xgbTree, Extreme Gradient Boosting.

males (maximum difference across models = 0.2 years; Tables S6 and S7). Similarly, application of parameters from models trained on females to the entire sample yielded higher BrainAGE for females than males (maximum difference across models = 0.6 years; Tables S6 and S7).

3.5.2 | Parcellation scheme and number of input features

The different parcellation schemes had minimal influence on the MAE in any of the three best performing algorithms (Figure 5a). The number of features in the range examined (136–2000) had minimal impact on MAE and notably the Schaefer-1000 parcellation did not outperform the Schaefer-400 parcellation used in the main analyses. The same pattern was seen in females and males in the ABCD and PING datasets (Figures S11 and S12).

3.5.3 | Sample size

MAE_T improved in line with sample increase up to a size of 500 participants and it plateaued thereafter. The corrected $MAE_{T,}$ on the other hand sowed limited change across different sample sizes (Figures 5b and S13).

TABLE 1 Algorithm performance in the hold-out sets

	ABCD			PING			
Algorithm (function name in caret package)	MAE$_{\text{T}}$	RMSE	Bias-adjusted MAE$_{\text{T}}$	MAE$_{\text{T}}$	RMSE	Bias-adjusted MAE$_{\text{T}}$	Correlation
Males							
Extreme Gradient Boosting (xgbTree)	1.57	2.03	1.16	2.02	2.5	1.32	0.86
Random Forest Regression (rf)	1.65	2.13	1.09	2.57	3.11	1.14	0.81
Support Vector Regression-Radial Basis Function (svmRadial)	1.72	2.14	1.29	1.9	2.38	1.41	0.87
Support Vector Regression-Polynomial (svmPoly)	1.74	2.15	1.3	1.92	2.41	1.43	0.86
Relevance Vector Regression-Polynomial (rvmPoly)	1.78	2.2	1.38	1.9	2.41	1.46	0.86
Gaussian Processes Polynomial (gaussprPoly)	1.8	2.22	1.2	1.92	2.43	1.43	0.86
Gaussian Processes Radial (gaussprRadial)	1.81	2.22	1.34	1.99	2.48	1.26	0.87
Generalized Additive Model with Boosting (gamboost)	1.82	2.25	1.3	2.11	2.59	1.38	0.85
Sparse Partial Least Squares (spls)	1.86	2.29	1.44	1.93	2.44	1.53	0.86
Kernel Partial Least Squares (kernelpls)	1.86	2.29	1.44	1.93	2.44	1.53	0.86
Elastic Net Regression (enet)	1.9	2.32	1.44	1.99	2.5	1.49	0.85
Quantile Regression with LASSO penalty (rqlasso)	1.91	2.33	1.42	2.06	2.56	1.48	0.84
Relevance Vector Regression-Radial (rvmRadial)	1.94	2.39	1.53	2	2.6	1.6	0.83
Bayesian Regularized Neural Network (brnn)	1.99	2.51	1.69	2.09	2.62	1.74	0.83
Independent Component Regression (icr)	2.1	2.55	1.37	2.47	2.99	1.4	0.79
Principal Component Regression (pcr)	2.1	2.55	1.37	2.47	2.99	1.4	0.79
Support Vector Regression-Linear (svmLinear3)	2.7	3.37	2.5	2.7	3.44	2.52	0.73
Gaussian Processes-Linear (gaussprLinear)	2.77	3.45	2.52	2.73	3.38	2.56	0.74
Generalized Linear Model (glm)	2.81	3.49	2.56	2.76	3.41	2.6	0.74
Bayesian Generalized Linear Model (bayesglm)	2.81	3.5	2.56	2.76	3.41	2.60	0.74
Relevance Vector Machine-Linear (rvmRaidal)	11.09	11.22	1.32	12.95	13.25	1.39	0.81
Females							
Random Forest Regression (rf)	1.23	1.66	1.08	2.62	3.15	1.2	0.84
Extreme Gradient Boosting (xgbTree)	1.25	1.69	1.19	2.17	2.69	1.41	0.86
Support Vector Regression-Radial (svmRadial)	1.47	1.89	1.21	2.08	2.63	1.38	0.87
Support Vector Regression-Polynomial (svmPoly)	1.47	1.89	1.24	2	2.56	1.48	0.87
Gaussian Processes Polynomial (gaussprPoly)	1.48	1.9	1.2	2.05	2.62	1.5	0.86
Generalized Additive Model with Boosting (gamboost)	1.49	1.91	1.33	2.19	2.73	1.57	0.86
Relevance Vector Regression-Polynomial (rvmPoly)	1.51	1.94	1.28	2.04	2.62	1.53	0.86
Gaussian Processes Radial (gaussprRadial)	1.52	1.95	1.24	2.1	2.62	1.28	0.88
Relevance Vector Regression-Radial (rvmRadial)	1.64	2.08	1.44	2.23	2.89	1.66	0.83
Quantile Regression with LASSO penalty (rqlasso)	1.66	2.08	1.43	2.13	2.65	1.61	0.85
Independent Component Regression (icr)	1.67	2.11	1.44	2.32	2.85	1.52	0.83
Principal Component Regression (pcr)	1.67	2.11	1.44	2.32	2.85	1.52	0.83
Elastic Net Regression (enet)	1.7	2.15	1.51	2	2.54	1.73	0.86
Kernel Partial Least Squares (kernelpls)	1.78	2.21	1.5	1.91	2.45	1.62	0.87
Sparse Partial Least Squares (spls)	1.78	2.21	1.5	1.91	2.45	1.62	0.87
Bayesian Regularized Neural Network (brnn)	1.97	2.52	1.79	2	2.59	1.7	0.86
Support Vector Regression-Linear (svmLinear3)	3.66	4.58	3.51	3.87	4.88	3.74	0.61
Gaussian Processes-Linear (gaussprLinear)	4.17	5.23	3.97	4.48	5.62	4.35	0.54
Bayesian Generalized Linear Model (bayesglm)	5.24	6.61	5.06	5.68	7.15	5.53	0.42
Generalized Linear Model (glm)	5.3	6.69	10.47	5.74	7.23	10.13	0.41
Relevance Vector Machine-Linear (rvmLinear)	11.67	11.83	1.47	13.17	13.49	1.58	0.82

Note: Correlations were only conducted in the PING dataset which has a wider age-range and not in the ABCD dataset were the age-range was very restricted (9–10 years).

Abbreviations: ABCD, adolescent brain cognitive development; LASSO, least absolute shrinkage and selection operator; MAE$_{\text{T}}$, mean absolute error; PING, Pediatric Imaging, Neurocognition, and Genetics Data Repository; RMSE, root mean squared error.

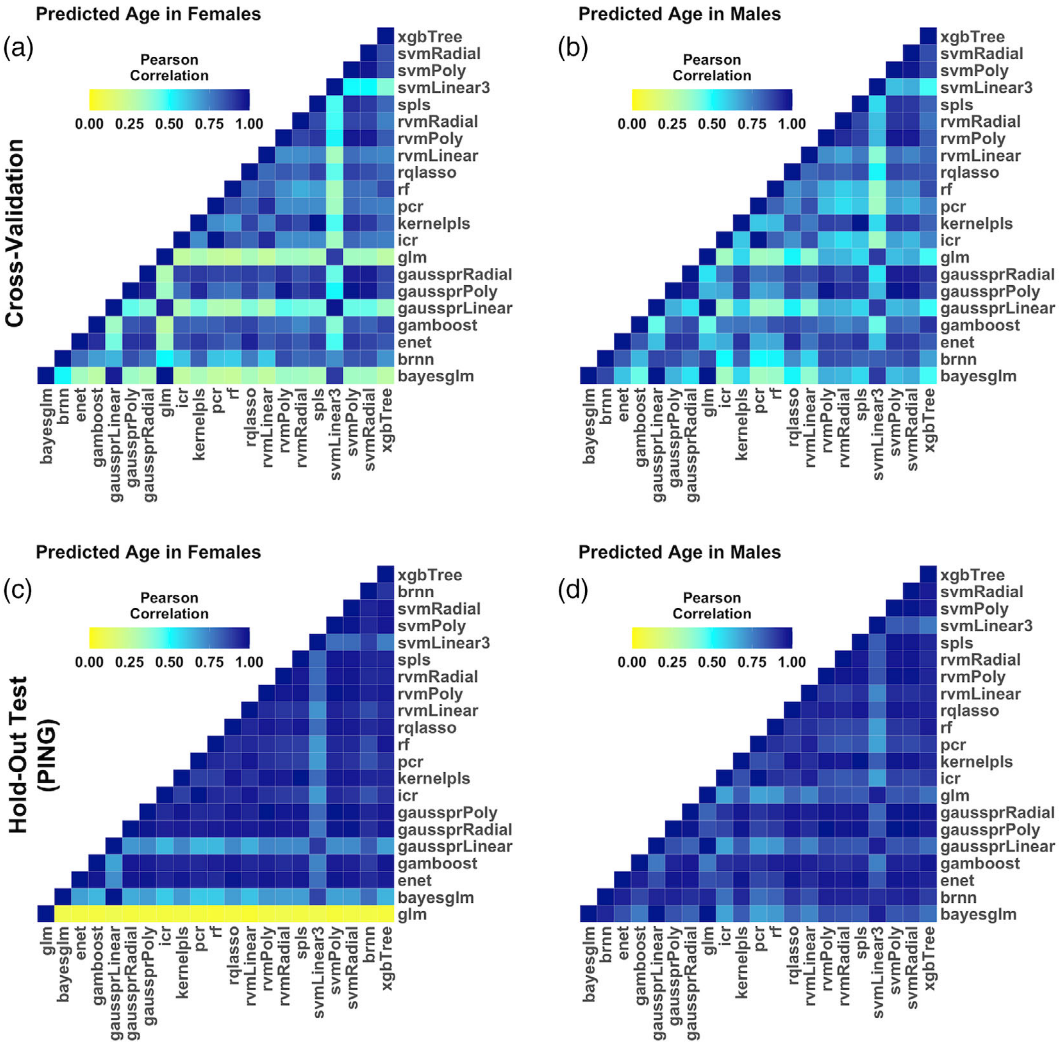

FIGURE 2 Pairwise correlations of the predicted age of the 21 algorithms. Figure demonstrates correlations between predicted age as estimated by different models in the females and males in the cross-validation set (Panels a and b), and in females and males in the hold-out Pediatric Imaging, Neurocognition, and Genetics Data Repository (PING) dataset (Panels c and d). The different algorithms are referenced by the function used for their implementation. bayesglm, Bayesian Generalized Linear Model; brnn, Bayesian Regularized Neural Network; enet, Elastic Net Regression; gamboost, Generalized Additive Model with Boosting; gaussprLinear, Gaussian Processes Regression Linear; gaussprPoly, Gaussian Processes Regression Polynomial; gaussprRadial, Gaussian Processes Regression Radial; glm, Generalized Linear Model; icr, Independent Component Regression; kernelpls, Kernel Partial Least Squares; pcr, Principal Component Regression; rf, Random Forest; rqlasso, Quantile Regression with LASSO penalty; rvmLinear3, Relevance Vector Machine-Linear; rvmPoly, Relevance Vector Machine-Polynomial; rvmRadial, Relevance Vector Machine-Radial; spls, Sparse Partial Least Squares; svmeLinear3, Support Vector Regression-Linear; svmPoly, Support Vector Regression-Polynomial; svmRadial, Support Vector Regression-Radial; xgbTree, Extreme Gradient Boosting.

3.5.4 | Number of cross-validation folds and repeats

In the main analyses, we used fivefold repeats and fivefold cross-validations. Using 10 instead of fivefolds and repeats did not improve performance and in the case of XGBoost, we noted markedly worse performance in the MAE_T (Table S8).

3.5.5 | Effect of extreme outliers in the test set

In the PING dataset, spearman's correlation coefficients between the number of outliers and the MAE derived from the three best performing algorithms was small (all rho <0.15). In the ABCD dataset, the corresponding values were of similar magnitude with a maximum rho of 0.25 for RF regression. The magnitude of these associations was reduced when using age-bias-corrected MAE (max rho <0.2). Among the three best performing models, the performance of SVR with the RBF kernel was the least impacted by extreme outliers (Table S9).

4 | DISCUSSION

In the present study, we undertook a comprehensive comparison of machine learning algorithms for sMRI-based age prediction as a proxy

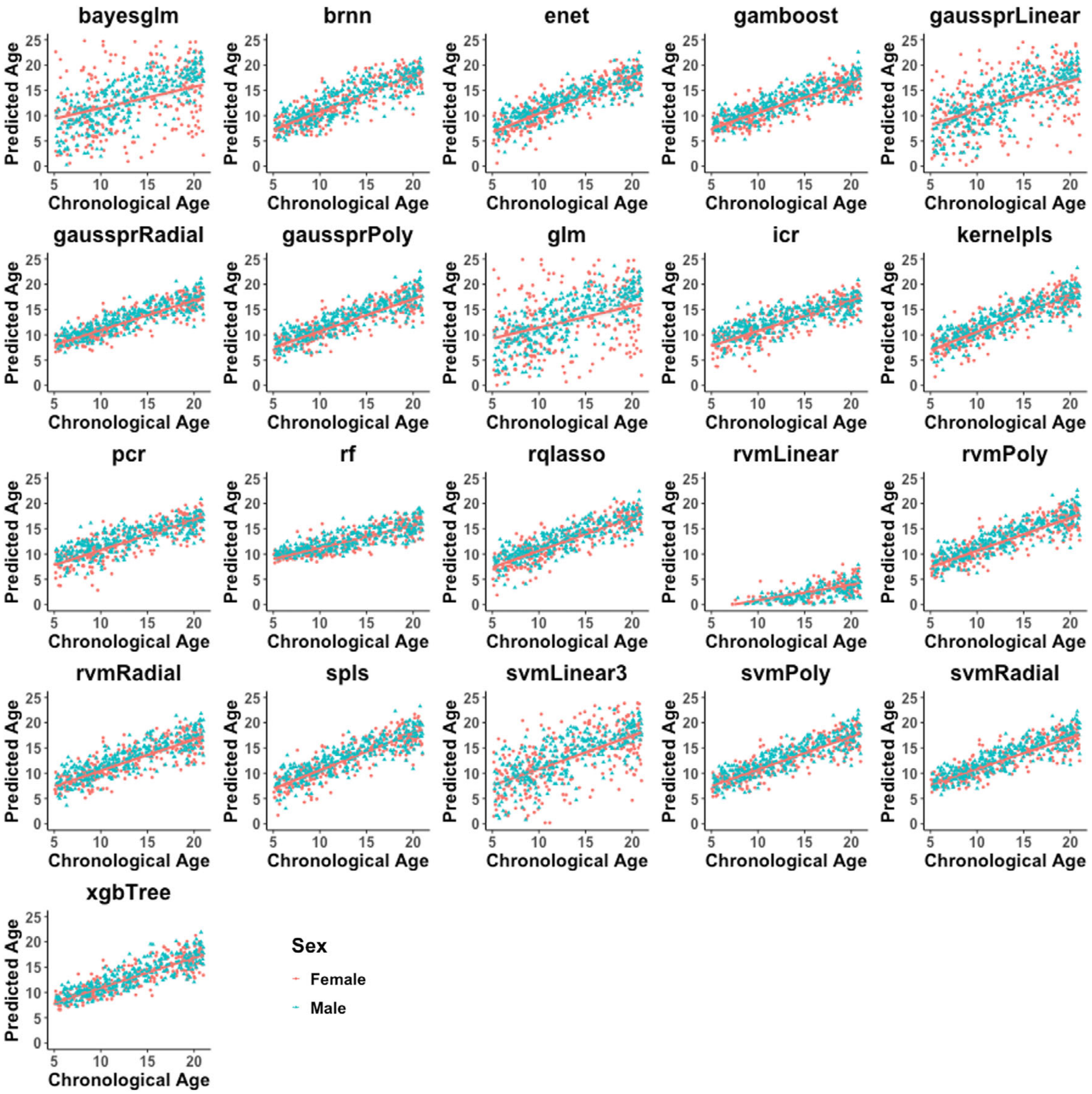

FIGURE 3 Correlations between chronological age (years) and predicted age across 21 algorithms in the PING dataset. The figure shows the correlation of chronological age with sMRI-age in each of the 21 algorithms tested in males and females in the Pediatric Imaging, Neurocognition, and Genetics Data Repository (PING). The different algorithms are referenced by the function used for their implementation. bayesglm, Bayesian Generalized Linear Model; brnn, Bayesian Regularized Neural Network; enet, Elastic Net Regression; gamboost, Generalized Additive Model with Boosting; gaussprLinear, Gaussian Processes Regression Linear; gaussprPoly, Gaussian Processes Regression Polynomial; gaussprRadial, Gaussian Processes Regression Radial; glm, Generalized Linear Model; icr, Independent Component Regression; kernelpls, Kernel Partial Least Squares; pcr, Principal Component Regression; rf, Random Forest; rqlasso, Quantile Regression with LASSO penalty; rvmLinear3, Relevance Vector Machine-Linear; rvmPoly, Relevance Vector Machine-Polynomial; rvmRadial, Relevance Vector Machine-Radial; spls, Sparse Partial Least Squares; svmeLinear3, Support Vector Regression-Linear; svmPoly, Support Vector Regression-Polynomial; svmRadial, Support Vector Regression-Radial; xgbTree, Extreme Gradient Boosting.

for the biological age of the brain in youth. We identified three algorithms, namely XGBoost, RF regression, and SVR with the RBF kernel, that outperformed all others in terms of accuracy while being computationally efficient. Notably we also show that sMRI-based age prediction was suboptimal in models using linear algorithms.

Linear algorithms consistently underperformed compared with other algorithms probably because of the multicollinearity of the neuroimaging data, as suggested by the relative better performance of those linear algorithms that are based on covariance (such as SPLS regression or PCA regression). Further, the general underperformance of linear models may also reflect the fact that they do not account for nonlinear and interactive associations between brain imaging features and age.

As predicted XGBoost and RF regression, which are both ensemble tree-based algorithms, performed well in terms of their accuracy and generalizability to unseen samples. A decision tree is a machine learning algorithm that partitions the data into subsets based on conditional statements. Although each tree has generally low predictive performance, their combination (ensemble) improves generalizability

TABLE 2 BrainAGE and BrainAGE$_{corr}$ in the ABCD and PING sample

Algorithm (function name in caret package)	ABCD				PING			
	Males BrainAGE	Males BrainAGE$_{corr}$	Females BrainAGE	Females BrainAGE$_{corr}$	Males BrainAGE	Males BrainAGE$_{corr}$	Females BrainAGE	Females BrainAGE$_{corr}$
Extreme Gradient Boosting (xgbTree)	1.33 (1.53)	0.22 (1.50)	0.73 (1.52)	−0.15 (1.49)	0.06 (2.51)	0.05 (1.66)	−0.28 (2.68)	−0.19 (1.80)
Random Forest Regression (rf)	1.55 (1.46)	−0.03 (1.38)	0.97 (1.34)	−0.37 (1.27)	−0.13 (3.10)	−0.19 (1.42)	−0.43 (3.10)	−0.34 (1.50)
Support Vector Regression-Radial (svmRadial)	1.44 (1.57)	0.56 (1.55)	1.04 (1.57)	0.21 (1.55)	0.24 (2.40)	0.26 (1.73)	−0.22 (2.62)	−0.16 (1.88)

Note: Values shown as mean (standard deviation).

Abbreviations: ABCD, adolescent brain cognitive development; BrainAGE$_{corr}$, Age-bias corrected BrainAGE; MAE$_T$, mean absolute error; PING, Pediatric Imaging, Neurocognition, and Genetics Data Repository.

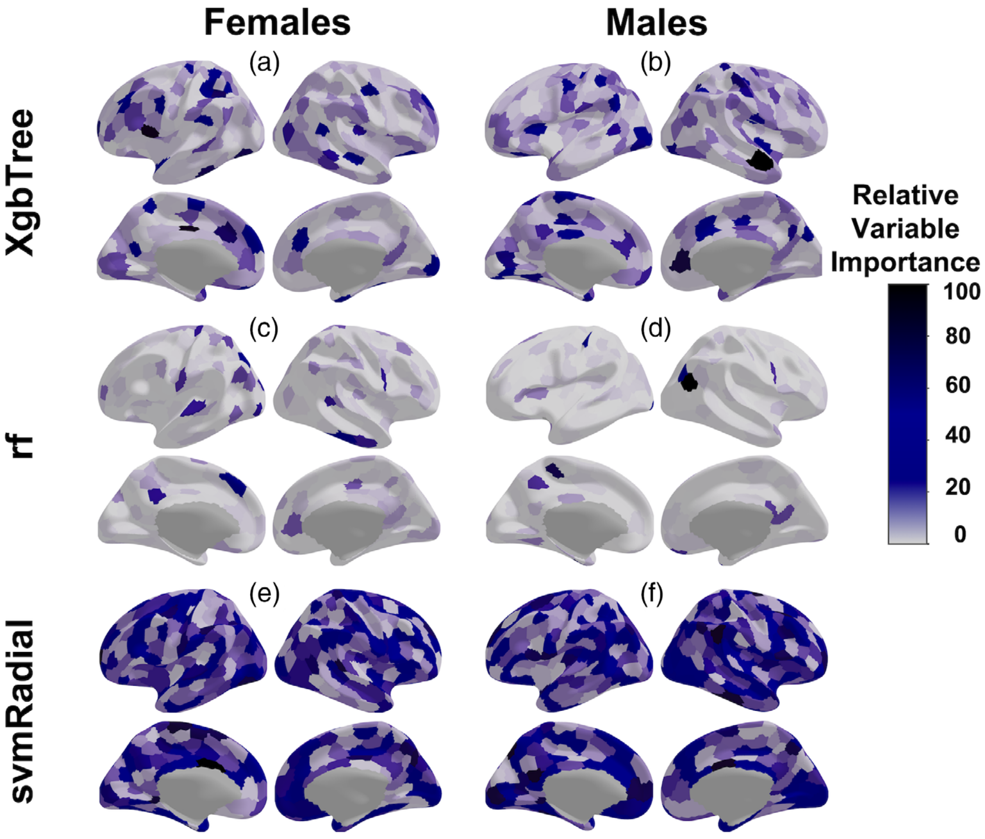

FIGURE 4 Relative importance of neuroimaging features for age prediction in the three best performing algorithms. The figure shows the relative variable importance of the features of the 400-parcel Schaefer Atlas for age-prediction based on their Shapley Values in females (Panel a) and males (Panel b). The relative importance values shown were rescaled such that the feature with the maximum average absolute Shapley Value in each model was assigned a value of 100. The algorithms are referenced by the function used for their implementation: rf, Random Forest Regression; svmRadial, Support Vector Regression-Radial; xgbTree, Extreme Gradient Boosting.

without sacrificing accuracy (Qi, 2012). Further advantages of these methods, particularly in the context of neuroimaging datasets, is that they are nonparametric, they do not involve assumptions about the distribution for the data and can account for nonlinear effects and interactions, which may be particularly relevant in modeling developmental brain-age. They also require minimal preparation of the input sMRI features as they can handle multicollinear data without losing accuracy. Notably, these algorithms were relatively insensitive to the number of the neuroimaging features, when the size of the feature sets ranged between 136 and 2000, with the mid-point feature set ($n = 841$) being relatively better.

Similar considerations applied to SVR with RBF algorithm which had the additional advantage of being particularly robust to outliers. This may represent a particular strength of this algorithm for studies with relatively small datasets where strict rules for outlier exclusion may result in significant data loss. Despite similar performance in terms of accuracy, the sMRI features contributing to age prediction differed across the three best-performing algorithms. Relative to the other two algorithms, more features contributed to age prediction in SVR with RBF which may contribute to its robustness to outliers.

Two recent studies also undertook benchmarking of methods used for brainAGE computation in adult (Baecker et al., 2021;

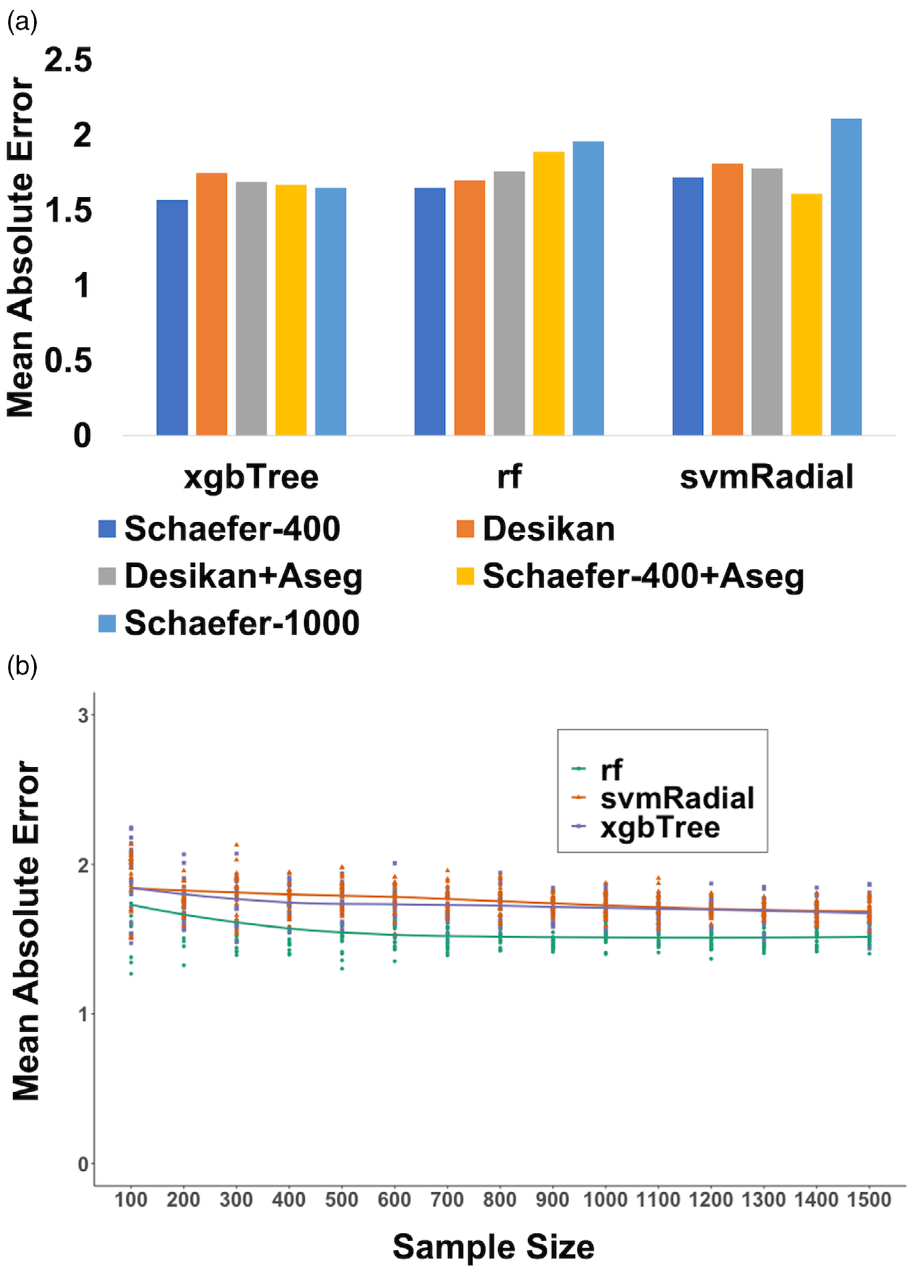

FIGURE 5 Mean absolute error (MAE) as a function of Parcellation scheme and sample size. Panel a: The mean absolute error for the three best performing algorithms as a function of parcellation scheme in male participants from the Adolescent Brain Cognitive Development (ABCD) study. The corresponding information from female ABCD participants and the PING sample are shown in Figures S8 and S9); Panel b: Mean absolute error of each of the three best performing algorithms as a function of sample size in the ABCD sample; model parameters for each algorithm were obtained by randomly resampling the training dataset without replacement generating subsamples of 100–1500. The algorithms are referenced by the function used for their implementation: rf, Random Forest Regression; svmRadial, Support Vector Regression-Radial; xgbTree, Extreme Gradient Boosting.

Beheshti et al., 2022). Baecker et al. (2021) examined the performance of three of the algorithms tested here, namely support vector regression, relevance vector regression and Gaussian process regression, in 10,824 participants in the UK Biobank, aged range 47–73 years. They reported minimal differences in accuracy in the three algorithms tested. Beheshti et al. (2022) tested the performance of 22 different algorithms in a sample of 876 healthy adults, aged 18–94 years. The algorithms overlapped with those used here and included linear, non-linear, and tree-based models. They also found that linear models underperformed compared with kernel-based and tree-based models. The range of MAE values in both studies was 3.7–7.1 years which is numerically higher than that observed here probably because of the wider age-range.

A major concern in neuroimaging research is the effect of site on the generalizability of ML models (Dockes et al., 2021; Solanes et al., 2021). Sites may differ in terms of scanner infrastructure, acquisition protocols, and neuroimaging feature extraction pipelines as well as sample composition. Here the post-acquisition extraction of neuroimaging features was undertaken for all cohorts using the same pipeline which may have reduced variability in the neuroimaging feature set. However, all other parameters differed between cohorts (and between recruitment sites within cohorts). Yet the three best performing algorithms showed excellent generalizability to the hold-out datasets, which is likely to reflect the robustness of these algorithms. Additionally, the inclusion of observations from multiple sites in the training dataset may have forced the ML algorithms to select and weight features that are robust to site differences, therefore reducing the dependence of the model on the effects of site.

We observed a lower MAE for females compared with males across most models. This has been reported in prior studies (Brouwer et al., 2021; Wierenga et al., 2019; Wierenga et al., 2020), and can be attributed to either biological differences, that is, female brain showing less variability or confounding, that is, males may move more, on average, than females which could make their brain measurements less accurate. Brouwer et al. (2021) demonstrated that in individuals aged 9–23 years, females have higher sMRI-derived BrainAGE than their male counterparts. The same pattern was reported by Tu et al. (2019) using sMRI data from 118 males and 147 females, aged 5–18, from the NIH MRI Study of Normal Brain Development. In these studies, as well as in ours, sex differences in BrainAGE are small and within the range of the MAE for brain-predicted age.

We acknowledge that the list of algorithms evaluated is not exhaustive but provides a good coverage of the many models that are currently available. We were unable to account for potential influences of race and ethnicity as such information was either absent or not uniformly coded in the cohorts used for model training. Based on the racial constitution of the general population, in the countries of the recruitment sites, we anticipate an over-representation of white individuals. As more data becomes available on other racial/ethnic groups, it should be possible to address this issue in future studies.

In summary, using a wide range of ML algorithms on geographically diverse datasets of young people, we showed that tree-based followed by nonlinear kernel-based algorithms offer robust, accurate, and generalizable solutions for predicting age based on brain morphological features. Findings of the present study can be used as a guide for quantifying brain maturation during development and its contribution to functional and behavioral outcomes.

AUTHOR CONTRIBUTIONS

Amirhossein Modabbernia: Conceptualization, Methodology, Investigation, Data Curation, Formal analysis, Writing - Original Draft, Writing - Review & Editing. Sophia Frangou: Supervision, Conceptualization, Writing - Original Draft, Writing - Review & Editing. Rene S. Kahn: Supervision, Conceptualization, Writing - Review & Editing. Heather C. Whalley: Supervision, Conceptualization, Review & Editing. David C. Glahn: Supervision - Review & Editing. Paul M. Thompson: Supervision - Review & Editing.

ACKNOWLEDGMENTS

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children aged 9–10 and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041022, U01DA041028, U01DA041048, U01DA041089, U01DA041106, U01DA041117, U01DA041120, U01DA041134, U01DA041148, U01DA041156, U01DA041174, U24DA041123, U24DA041147, U01DA041093, and U01DA041025. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/scientists/workgroups. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The ABCD data repository grows and changes over time. Dr. A.M. received support from the National Institute of Mental Health T32-MH122394. Dr S.F. received support from the National Institute of Mental Health (R01-MH113619). This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The UKB Data described in this manuscript, is available to all researchers and can be accessed upon approval. Specifically, Data from the Autism Brain Imaging Data Exchange (ABIDE) and ABIDE II can be accessed through http://fcon_1000.projects.nitrc.org/indi/abide/ and http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html. Details of these initiatives have been published by (Di Martino et al., 2017; Di Martino et al., 2014). Data from the Attention Deficit Hyperactivity Disorder (ADHD)-200 Consortium can be accessed through http://fcon_1000.projects.nitrc.org/indi/adhd200/#. Details of the Consortium have been published by the ADHD consortium (ADHD-200 Consortium, 2012). Data from the Child Mind Institute-Healthy Brain Network (CMI-HBN) can be accessed through http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/. Details of the CMI-HBN have been by published by (Alexander et al., 2017). Data from the Human Connectome Project- Development (HCP-D) can be accessed through https://www.humanconnectome.org/study/hcp-lifespan-development. Details of the HCP-D have been published by Harms et al., 2018 (Harms et al., 2018). Data from the Adolescent Brain Cognitive Development (ABCD) study can be accessed through https://nda.nih.gov/abcd/. Details of the ABCD study design and of the neuroimaging acquisition and quality assurance procedures have been published by (Casey et al., 2018; Hagler et al., 2019). Data from the Pediatric Imaging, Neurocognition, and Genetics (PING) Study can be accessed through http://pingstudy.ucsd.edu/. Details of the PING study design and of the neuroimaging acquisition and quality assurance procedures have been published by (Jernigan et al., 2016). Analytic codes used in this study can be accessed through https://github.com/AmirhosseinModabbernia/DevelopmentalBrainAge.

ORCID

Sophia Frangou https://orcid.org/0000-0002-3210-6470

REFERENCES

Alexander, L. M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., Kovacs, M., Litke, S., O'Hagan, B., Andersen, J., Bronstein, B., Bui, A., Bushey, M., Butler, H., Castagna, V., Camacho, N., … Milham, M. P. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, *4*, 170181. https://doi.org/10.1038/sdata.2017.181

Anatürk, M., Kaufmann, T., Cole, J. H., Suri, S., Griffanti, L., Zsoldos, E., Filippini, N., Singh-Manoux, A., Kivimäki, M., Westlye, L. T., Ebmeier, K. P., & de Lange, A. G. (2021). Prediction of brain age and cognitive age: Quantifying brain and cognitive maintenance in aging. *Human Brain Mapping*, *42*(6), 1626–1640. https://doi.org/10.1002/hbm.25316

Awad, M., & Khanna, R. (2015). Support vector regression. In M. Awad & R. Khanna (Eds.), *Efficient learning machines* (pp. 67–80). Springer.

Baecker, L., Dafflon, J., da Costa, P. F., Garcia-Dias, R., Vieira, S., Scarpazza, C., Calhoun, V. D., Sato, J. R., Mechelli, A., & Pinaya, W. (2021). Brain age prediction: A comparison between machine learning models using region- and voxel-based morphometric data. *Human Brain Mapping*, *42*(8), 2332–2346. https://doi.org/10.1002/hbm.25368

Ball, G., Kelly, C. E., Beare, R., & Seal, M. L. (2021). Individual variation underlying brain age estimates in typical development. *NeuroImage*, *235*, 118036. https://doi.org/10.1016/j.neuroimage.2021.118036

Beheshti, I., Ganaie, M. A., Paliwal, V., Rastogi, A., Razzak, I., & Tanveer, M. (2022). Predicting brain age using machine learning algorithms: A comprehensive evaluation. *IEEE Journal of Biomedical and Health Informatics*, *26*(4), 1432–1440. https://doi.org/10.1109/JBHI.2021.3083187

Beheshti, I., Nugent, S., Potvin, O., & Duchesne, S. (2019). Bias-adjustment in neuroimaging-based brain age frameworks: A robust scheme. *Neuroimage: Clinical*, *24*, 102063. https://doi.org/10.1016/j.nicl.2019.102063

Boyle, R., Jollans, L., Rueda-Delgado, L. M., Rizzo, R., Yener, G. G., McMorrow, J. P., Knight, S. P., Carey, D., Robertson, I. H., Emek-Savaş, D. D., Stern, Y., Kenny, R. A., & Whelan, R. (2021). Brain-predicted age difference score is related to specific cognitive functions: A multi-site replication analysis. *Brain Imaging and Behavior*, *15*(1), 327–345. https://doi.org/10.1007/s11682-020-00260-3

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brouwer, R. M., Schutte, J., Janssen, R., Boomsma, D. I., Hulshoff Pol, H. E., & Schnack, H. G. (2021). The speed of development of adolescent brain age depends on sex and is genetically determined. *Cerebral Cortex*, *31*(2), 1296–1306. https://doi.org/10.1093/cercor/bhaa296

Bryce, N. V., Flournoy, J. C., Guassi Moreira, J. F., Rosen, M. L., Sambook, K. A., Mair, P., & McLaughlin, K. A. (2021). Brain parcellation selection: An overlooked decision point with meaningful effects on individual differences in resting-state functional connectivity. *NeuroImage*, *243*, 118487. https://doi.org/10.1016/j.neuroimage.2021.118487

Bühlmann, P., & Yu, B. (2003). Boosting with the L 2 loss: Regression and classification. *Journal of the American Statistical Association*, *98*(462), 324–339.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Paper presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Chun, H., & Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(1), 3–25.

Chung, Y., Addington, J., Bearden, C. E., Cadenhead, K., Cornblatt, B., Mathalon, D. H., McGlashan, T., Perkins, D., Seidman, L. J., Tsuang, M., Walker, E., Woods, S. W., McEwen, S., Van Erp, T., Cannon, T. D., & North American Prodrome Longitudinal Study (NAPLS) Consortium and the Pediatric Imaging, Neurocognition, and Genetics (PING) Study Consortium. (2018). Use of machine learning to determine deviance in neuroanatomical maturity associated with future psychosis in youths at clinically high risk. *JAMA Psychiatry*, *75*(9), 960–968. https://doi.org/10.1001/jamapsychiatry.2018.1543

Cole, J. H., Ritchie, S. J., Bastin, M. E., Valdés Hernández, M. C., Muñoz Maniega, S., Royle, N., Corley, J., Pattie, A., Harris, S. E., Zhang, Q., Wray, N. R., Redmond, P., Marioni, R. E., Starr, J. M., Cox, S. R., Wardlaw, J. M., Sharp, D. J., & Deary, I. J. (2018). Brain age predicts mortality. *Molecular Psychiatry*, *23*(5), 1385–1392. https://doi.org/10.1038/mp.2017.62

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Dayal, B. S., & MacGregor, J. F. (1997). Improved PLS algorithms. *Journal of Chemometrics*, *11*(1), 73–85.

Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., Balsters, J. H., Baxter, L., Beggiato, A., Bernaerts, S., Blanken, L. M., Bookheimer, S. Y., Braden, B. B., Byrge, L., Castellanos, F. X., Dapretto, M., Delorme, R., Fair, D. A., Fishman, I., … Milham, M. P. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific Data*, *4*, 170010. https://doi.org/10.1038/sdata.2017.10

Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., … Milham, M. P. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, *19*(6), 659–667. https://doi.org/10.1038/mp.2013.78

Dima, D., Modabbernia, A., Papachristou, E., Doucet, G. E., Agartz, I., Aghajani, M., Akudjedu, T. N., Albajes-Eizagirre, A., Alnaes, D., Alpert, K. I., Andersson, M., Andreasen, N. C., Andreassen, O. A., Asherson, P., Banaschewski, T., Bargallo, N., Baumeister, S., Baur-Streubel, R., Bertolino, A., … Karolinska Schizophrenia Project (KaSP). (2022). Subcortical volumes across the lifespan: Data from 18,605 healthy individuals aged 3-90 years. *Human Brain Mapping*, *43*, 452–469. https://doi.org/10.1002/hbm.25320

Dockes, J., Varoquaux, G., & Poline, J. B. (2021). Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience*, *10*(9), 1–11. https://doi.org/10.1093/gigascience/giab055

Elliott, M. L., Belsky, D. W., Knodt, A. R., Ireland, D., Melzer, T. R., Poulton, R., Ramrakha, S., Caspi, A., Moffitt, T. E., & Hariri, A. R. (2021). Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort. *Molecular Psychiatry*, *26*, 3829–3838. https://doi.org/10.1038/s41380-019-0626-7

Erus, G., Battapady, H., Satterthwaite, T. D., Hakonarson, H., Gur, R. E., Davatzikos, C., & Gur, R. C. (2015). Imaging patterns of brain development and their relationship to cognition. *Cerebral Cortex*, *25*(6), 1676–1684. https://doi.org/10.1093/cercor/bht425

Fjell, A. M., Westlye, L. T., Amlien, I., Tamnes, C. K., Grydeland, H., Engvig, A., Espeseth, T., Reinvang, I., Lundervold, A. J., Lundervold, A., & Walhovd, K. B. (2015). High-expanding cortical regions in human development and evolution are related to higher intellectual abilities. *Cerebral Cortex*, *25*(1), 26–34. https://doi.org/10.1093/cercor/bht201

Frangou, S., Modabbernia, A., Williams, S. C. R., Papachristou, E., Doucet, G. E., Agartz, I., Aghajani, M., Akudjedu, T. N., Albajes-Eizagirre, A., Alnaes, D., Alpert, K. I., Andersson, M., Andreasen, N. C., Andreassen, O. A., Asherson, P., Banaschewski, T., Bargallo, N., Baumeister, S., Baur-Streubel, R., … Dima, D. (2022). Cortical thickness across the lifespan: Data from 17,075 healthy individuals aged 3-90 years. *Human Brain Mapping*, *43*, 431–451. https://doi.org/10.1002/hbm.25364

Franke, K., & Gaser, C. (2019). Ten years of BrainAGE as a neuroimaging biomarker of brain aging: What insights have we gained? *Frontiers in Neurology*, *10*, 789. https://doi.org/10.3389/fneur.2019.00789

Franke, K., Luders, E., May, A., Wilke, M., & Gaser, C. (2012). Brain maturation: Predicting individual BrainAGE in children and adolescents using structural MRI. *NeuroImage*, *63*(3), 1305–1312. https://doi.org/10.1016/j.neuroimage.2012.08.001

Franke, K., Ziegler, G., Kloppel, S., Gaser, C., & Alzheimer's Disease Neuroimaging Initiative. (2010). Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, *50*(3), 883–892. https://doi.org/10.1016/j.neuroimage.2010.01.005

Garavan, H., Bartsch, H., Conway, K., Decastro, A., Goldstein, R. Z., Heeringa, S., Jernigan, T., Potter, A., Thompson, W., & Zahs, D. (2018). Recruiting the ABCD sample: Design considerations and procedures. *Developmental Cognitive Neuroscience*, *32*, 16–22. https://doi.org/10.1016/j.dcn.2018.04.004

Gaser, C., Franke, K., Kloppel, S., Koutsouleris, N., Sauer, H., & Alzheimer's Disease Neuroimaging Initiative. (2013). BrainAGE in mild cognitive impaired patients: Predicting the conversion to Alzheimer's disease. *PLoS One*, *8*(6), e67346. https://doi.org/10.1371/journal.pone.0067346

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

Greenwell, B., & Greenwell, M. B. (2020). Package 'fastshap'.

Harms, M. P., Somerville, L. H., Ances, B. M., Andersson, J., Barch, D. M., Bastiani, M., Bookheimer, S. Y., Brown, T. B., Buckner, R. L., Burgess, G. C., Coalson, T. S., Chappell, M. A., Dapretto, M., Douaud, G., Fischl, B., Glasser, M. F., Greve, D. N., Hodge, C., Jamison, K. W., … Yacoub, E. (2018). Extending the human connectome project across ages: Imaging protocols for the lifespan development and aging projects. *NeuroImage*, *183*, 972–984. https://doi.org/10.1016/j.neuroimage.2018.09.060

Jernigan, T. L., Brown, T. T., Hagler, D. J., Jr, Akshoomoff, N., Bartsch, H., Newman, E., Thompson, W. K., Bloss, C. S., Murray, S. S., Schork, N., Kennedy, D. N., Kuperman, J. M., McCabe, C., Chung, Y., Libiger, O., Maddox, M., Casey, B. J., Chang, L., Ernst, T. M., … Frazier, J. A. (2016). The Pediatric Imaging, Neurocognition, and Genetics (PING) data repository. *NeuroImage*, *124*(Pt B), 1149–1154.

Kaufmann, T., van der Meer, D., Doan, N. T., Schwarz, E., Lund, M. J., Agartz, I., Alnæs, D., Barch, D. M., Baur-Streubel, R., Bertolino, A., Bettella, F., Beyer, M. K., Bøen, E., Borgwardt, S., Brandt, C. L., Buitelaar, J., Celius, E. G., Cervenka, S., Conzelmann, A., … Westlye, L. T. (2019). Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature Neuroscience*, *22*(10), 1617–1623. https://doi.org/10.1038/s41593-019-0471-7

Lee, W. H., Antoniades, M., Schnack, H. G., Kahn, R. S., & Frangou, S. (2021). Brain age prediction in schizophrenia: Does the choice of machine learning algorithm matter? *Psychiatry Research: Neuroimaging*, *310*, 111270. https://doi.org/10.1016/j.pscychresns.2021.111270

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Kharabian Masouleh, S., Huntenburg, J. M., Lampe, L., Rahim, M., Abraham, A., Craddock, R. C., Riedel-Heller, S., Luck, T., Loeffler, M., Schroeter, M. L., Witte, A. V., Villringer, A., & Margulies, D. S. (2017). Predicting brain-age from multimodal imaging data captures cognitive impairment. *NeuroImage*, *148*, 179–188. https://doi.org/10.1016/j.neuroimage.2016.11.005

Luna, A., Bernanke, J., Kim, K., Aw, N., Dworkin, J. D., Cha, J., & Posner, J. (2021). Maturity of gray matter structures and white matter connectomes, and their relationship with psychiatric symptoms in youth. *Human Brain Mapping*, *42*, 4568–4579. https://doi.org/10.1002/hbm.25565

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Paper presented at the Proceedings of the 31st international conference on neural information processing systems.

Mills, K. L., Siegmund, K. D., Tamnes, C. K., Ferschmann, L., Wierenga, L. M., Bos, M. G. N., Luna, B., Li, C., & Herting, M. M. (2021). Inter-individual variability in structural brain development from late childhood to young adulthood. *NeuroImage*, *242*, 118450. https://doi.org/10.1016/j.neuroimage.2021.118450

Paus, T. (2005). Mapping brain maturation and cognitive development during adolescence. *Trends in Cognitive Sciences*, *9*(2), 60–68. https://doi.org/10.1016/j.tics.2004.12.008

Perez-Rodriguez, P., Gianola, D., Weigel, K. A., Rosa, G. J., & Crossa, J. (2013). Technical note: An R package for fitting Bayesian regularized neural networks with applications in animal breeding. *Journal of Animal Science*, *91*(8), 3522–3531. https://doi.org/10.2527/jas.2012-6162

Qi, Y. (2012). Random forest for bioinformatics. In C. Zhang & Y. Q. Ma (Eds.), *Ensemble machine learning* (pp. 307–323). Springer.

Raznahan, A., Shaw, P. W., Lerch, J. P., Clasen, L. S., Greenstein, D., Berman, R., Pipitone, J., Chakravarty, M. M., & Giedd, J. N. (2014). Longitudinal four-dimensional mapping of subcortical anatomy in human development. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(4), 1592–1597. https://doi.org/10.1073/pnas.1316911111

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X. N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, *28*(9), 3095–3114. https://doi.org/10.1093/cercor/bhx179

Shao, X., Wang, W., Hou, Z., & Cai, W. (2006). A new regression method based on independent component analysis. *Talanta*, *69*(3), 676–680. https://doi.org/10.1016/j.talanta.2005.10.039

Solanes, A., Palau, P., Fortea, L., Salvador, R., González-Navarro, L., Llach, C. D., Valentí, M., Vieta, E., & Radua, J. (2021). Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site. *Psychiatry Research: Neuroimaging*, *314*, 111313. https://doi.org/10.1016/j.pscychresns.2021.111313

Spear, L. P. (2000). The adolescent brain and age-related behavioral manifestations. *Neuroscience and Biobehavioral Reviews*, *24*(4), 417–463. https://doi.org/10.1016/s0149-7634(00)00014-2

Tamnes, C. K., Herting, M. M., Goddings, A. L., Meuwese, R., Blakemore, S. J., Dahl, R. E., Güroğlu, B., Raznahan, A., Sowell, E. R., Crone, E. A., & Mills, K. L. (2017). Development of the cerebral cortex across adolescence: A multisample study of inter-related longitudinal changes in cortical volume, surface area, and thickness. *The Journal of Neuroscience*, *37*(12), 3402–3412. https://doi.org/10.1523/JNEUROSCI.3302-16.2017

Tau, G. Z., & Peterson, B. S. (2010). Normal development of brain circuits. *Neuropsychopharmacology*, *35*(1), 147–168. https://doi.org/10.1038/npp.2009.115

The ADHD-200 Consortium. (2012). The ADHD-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, *6*, 62. https://doi.org/10.3389/fnsys.2012.00062

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, *1*, 211–244.

Tu, Y., Fu, Z., & Maleki, N. (2019). When does the youthfulness of the female brain emerge? *Proceedings of the National Academy of Sciences of the United States of America*, *116*(22), 10632–10633. https://doi.org/10.1073/pnas.1905356116

Valizadeh, S. A., Hanggi, J., Merillat, S., & Jancke, L. (2017). Age prediction on the basis of brain anatomical measures. *Human Brain Mapping*, *38*(2), 997–1008. https://doi.org/10.1002/hbm.23434

Valk, S. L., Xu, T., Margulies, D. S., Masouleh, S. K., Paquola, C., Goulas, A., Kochunov, P., Smallwood, J., Yeo, B. T. T., Bernhardt, B. C., & Eickhoff, S. B. (2020). Shaping brain structure: Genetic and phylogenetic axes of macroscale organization of cortical thickness. *Science Advances*, *6*(39), eabb3417. https://doi.org/10.1126/sciadv.abb3417

Wierenga, L. M., Bos, M. G. N., van Rossenberg, F., & Crone, E. A. (2019). Sex effects on development of brain structure and executive functions: Greater variance than mean effects. *Journal of Cognitive Neuroscience*, *31*(5), 730–753. https://doi.org/10.1162/jocn_a_01375

Wierenga, L. M., Doucet, G. E., Dima, D., Agartz, I., Aghajani, M., Akudjedu, T. N., Albajes-Eizagirre, A., Alnaes, D., Alpert, K. I., Andreassen, O. A., Anticevic, A., Asherson, P., Banaschewski, T., Bargallo, N., Baumeister, S., Baur-Streubel, R., Bertolino, A., Bonvino, A., Boomsma, D. I., … Tamnes, C. K. (2022). Greater male than female variability in regional brain structure across the lifespan. *Human Brain Mapping*, *43*, 470–499. https://doi.org/10.1002/hbm.25204

Williams, C. K., & Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(12), 1342–1351.

Wu, Y., & Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, *19*(2), 801–817.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Modabbernia, A., Whalley, H. C., Glahn, D. C., Thompson, P. M., Kahn, R. S., & Frangou, S. (2022). Systematic evaluation of machine learning algorithms for neuroanatomically-based age prediction in youth. *Human Brain Mapping*, *43*(17), 5126–5140. https://doi.org/10.1002/hbm.26010