Predicting Brain Age from Structural MRI Using ResNet-Based 3D Convolutional Neural Networks

Lincoln Dibler

Independent Research Project

July 2025

Abstract:

Brain age modeling uses neuroimaging data to estimate an individual's biological brain age, offering a non-invasive biomarker of brain health. In this project, I developed a 3D convolutional neural network (ResNet3D) to predict brain age from structural MRI scans using the OASIS dataset. The model was trained on a cognitively healthy cohort with stratified age sampling and age-weighted loss to address data imbalance. Evaluation on a held-out test set yielded a mean absolute error (MAE) of 9.28 years and R^2 of 0.73. However, systematic underestimation was observed in middle-aged and older adults. To mitigate this bias, I applied Target Age Transformation (tAT) as a post-hoc correction method, reducing the test MAE to 7.54 years and improving R^2 to 0.83. These results demonstrate the potential of deep learning for brain age estimation and highlight the importance of bias correction in lifespan modeling. Future work should explore model interpretability and generalization to clinical populations.

Introduction:

Background:

Aging influences the brain at structural, functional and cognitive levels. While chronological age is a straightforward measure of time since birth, it does not always reflect the true biological state of an individual's brain. The concept of **brain age** refers to the estimated biological age of the brain, inferred from neuroimaging data using computational models. This estimate can deviate from chronological age, revealing insights into brain health, resilience, or risk for disease.

Machine learning—particularly deep learning—has made it possible to estimate brain age from structural MRI data with increasing accuracy. Convolutional neural networks (CNNs) can detect subtle, age-related features in brain anatomy, such as cortical thinning, ventricular expansion, and changes in white matter integrity, that might be challenging for human observers to quantify consistently. The difference between predicted brain age and chronological age, known as the brain

age gap, has been proposed as a useful biomarker in studies of neurological and psychiatric conditions.

A positive brain age gap (i.e., a brain that appears older than expected) has been associated with cognitive decline, Alzheimer's disease, traumatic brain injury, and other forms of neurodegeneration. Conversely, a negative gap may reflect cognitive resilience or protective factors such as education, physical activity, or genetic predisposition.

Understanding what drives these deviations—and how they relate to health outcomes—is a key goal in neuroscience and aging research. Brain age modeling offers a non-invasive, quantitative measure of brain health and has become a promising biomarker for early diagnosis, longitudinal monitoring, and treatment evaluation. Deep learning approaches, especially convolutional neural networks, have shown strong performance in capturing subtle anatomical changes—such as cortical thinning, ventricular enlargement, and white matter alterations—that may elude human observers.

Motivation:

In this project, I build on existing research to gain hands-on experience training CNNs on MRI data, with the goal of replicating and understanding these techniques as part of my portfolio development.

Research Question:

This project investigates the practical challenges and performance outcomes of applying 3D convolutional neural network (CNN) architectures to brain age regression using publicly available structural MRI datasets, specifically the OASIS dataset. The aim is to explore how well a 3D ResNet model can predict brain age across a cognitively diverse population, and to identify key technical hurdles encountered during model development and evaluation. By addressing these

questions, this work contributes a hands-on perspective on the feasibility and limitations of deep learning methods for brain age modeling in real-world neuroimaging data.

Related Work:

Brain age estimation has become an important area of neuroimaging research, with many studies applying machine learning techniques—especially deep convolutional neural networks (CNNs)—to structural MRI data. The OASIS dataset is a widely used resource in this field, providing diverse samples of cognitively healthy and impaired individuals. Prior work has demonstrated that 3D CNN architectures, such as ResNet variants, can effectively capture volumetric brain features and predict chronological age with reasonable accuracy. However, challenges remain in addressing age-related bias in predictions, often requiring post-hoc calibration or bias correction to improve the reliability and interpretability of brain age gap estimates.

Methodology:

Dataset and Preprocessing:

I utilized the OASIS-1 cross-sectional dataset, comprising structural MRI scans and associated metadata for aging and cognitive decline analysis. From the original data repository, Iextracted T1-weighted MRI scans stored in NIfTI format. To consolidate the dataset, I combined data across all 12 available ‘disc’ folders and retained only subjects with successfully converted NIfTI images.

Subject metadata, including age and Clinical Dementia Rating (CDR) scores, were loaded from the provided Excel sheet and matched to the MRI scans based on unique subject IDs. Many young subjects had missing CDR values, while older individuals often had recorded CDR scores, reflecting their cognitive status.

To focus the model on healthy brain aging and avoid confounding effects from neurodegeneration, I defined a healthy cohort as subjects with either a CDR score of

zero or missing CDR with age under 65 years. Subjects with CDR scores greater than zero were labeled as unhealthy and excluded from model training. This approach preserved a large portion of the dataset while maintaining a health-based inclusion criterion.

After filtering, the healthy cohort comprised 288 subjects. I performed a stratified split of this healthy subset into training, validation, and test sets based on age quantiles (8 bins) to ensure balanced age distributions across splits.

MRI scans were loaded using a custom PyTorch Dataset class that paired subject metadata with corresponding NIfTI images. This setup supports streamlined batch loading and preprocessing during model training and evaluation. Data augmentation and other training-specific transformations were applied later during the training phase.

Model and Training:

Architecture:

I implemented a custom 3D ResNet architecture in PyTorch, designed to process full volumetric brain scans as described in the preprocessing section. The model input consisted of normalized T1-weighted MRI volumes with a resolution of 128×128×128 voxels.

The Architecture Included:

A **stem block**:

- 3D convolution (kernel size 7, stride 2, padding 3)
- Batch normalization
- ReLU activation
- Max pooling (kernel size 3, stride 2)

Three residual stages with configurations [1, 1, 1], each using a `BasicBlock3D`:

- Two 3D convolutions (kernel size 3, padding 1)
- Batch normalization after each convolution
- ReLU activations
- Optional downsampling via 1×1×1 convolution when input/output dimensions differed

A **global average pooling** layer followed by a fully connected regression head:

- Linear(256 → 128) → ReLU → Dropout(0.5) → Linear(128 → 1)

A final **learnable affine transformation** (`age_scale` and `age_bias`) applied to the output to allow post-training calibration.

Training Strategy:

The model was trained using the mean squared error (MSE) loss function. To address age imbalance in the training data, I computed custom age weights based on stratified bins and applied a weighted loss, emphasizing underrepresented age groups.

Training was conducted over 40 epochs with early stopping (patience = 12) based on validation loss. Optimization used the Adam optimizer with:

- Learning rate: 1e-4
- L2 weight decay: 5e-4

A cosine annealing warm restart scheduler adjusted the learning rate cyclically to improve convergence and escape local minima.

Data Augmentation:

To improve generalization and robustness, I applied 3D data augmentation during training:

- Random 3D rotations (±10°)
- Affine transformations (scaling and translation)
- Gaussian noise injection

These transformations were applied on-the-fly during batch loading to increase anatomical variability and reduce overfitting.

Evaluation and Results:

To evaluate the model's performance, I report Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) across the training, validation, and test datasets. Final evaluation was performed on a held-out test set consisting of healthy individuals. Metrics are summarized in *Table 1*.

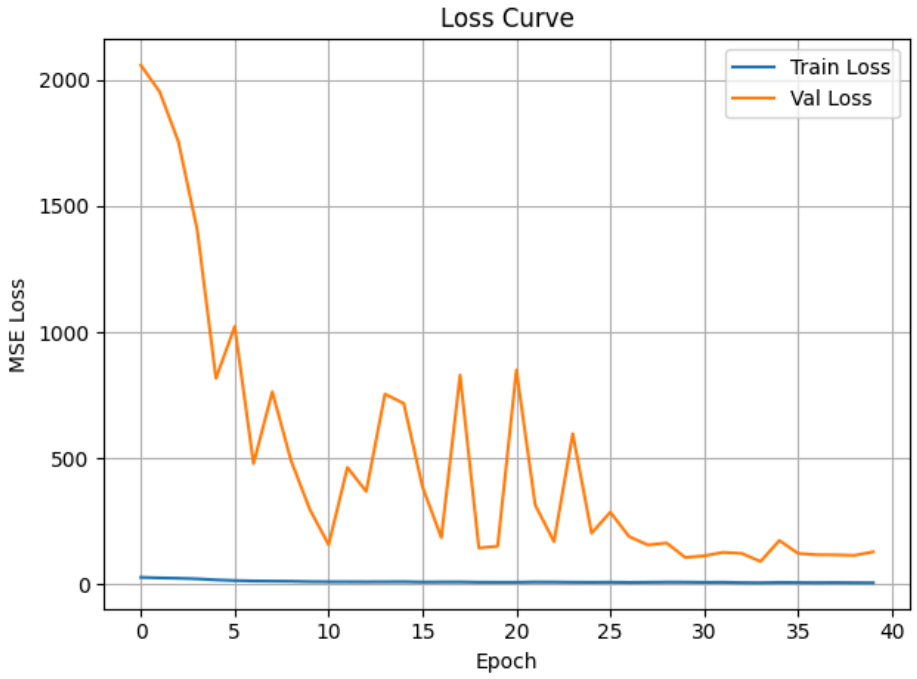
Table 1. Model Performance Across Datasets

Dataset	MAE (years)	RMSE (years)	R^2
Training	8.48	11.29	0.784
Validation	7.98	11.33	0.774
Test	9.28	13.05	0.73

Post-Hoc (Test) (bias-corrected)	7.54	10.40	0.83

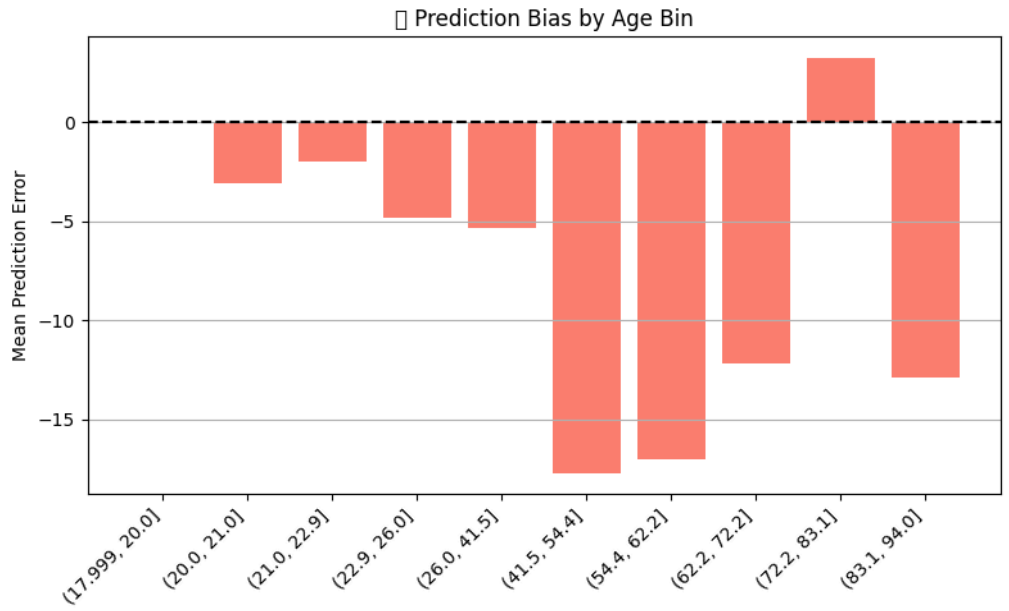
The model demonstrated strong generalization, with training and validation performance closely aligned. On the raw test set, the model achieved an MAE of 9.28 years and an R^2 of 0.73. After applying post-hoc bias correction, performance improved substantially, with the MAE reduced to 7.54 years and R^2 increasing to 0.83 — indicating a tighter fit and improved reliability across the age spectrum.

Figure 1: Training and Validation Loss



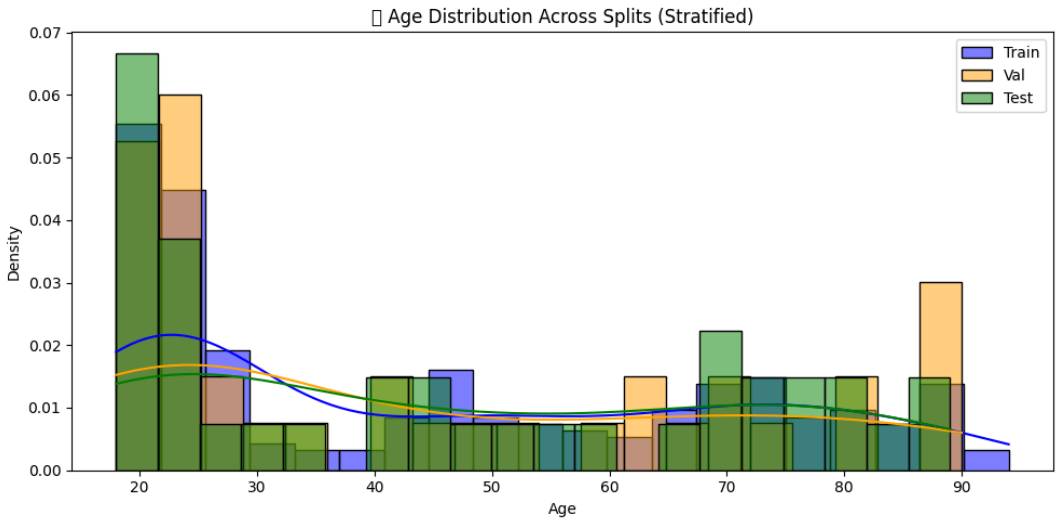
The plot shows the Mean Squared Error loss over training epochs. The training loss steadily decreases, indicating effective learning. The validation loss drops sharply early on, then fluctuates before stabilizing, suggesting good generalization with some variability. Overall, both curves converge, showing the model has trained sufficiently.

Figure 2: Prediction Bias Across Age Bins



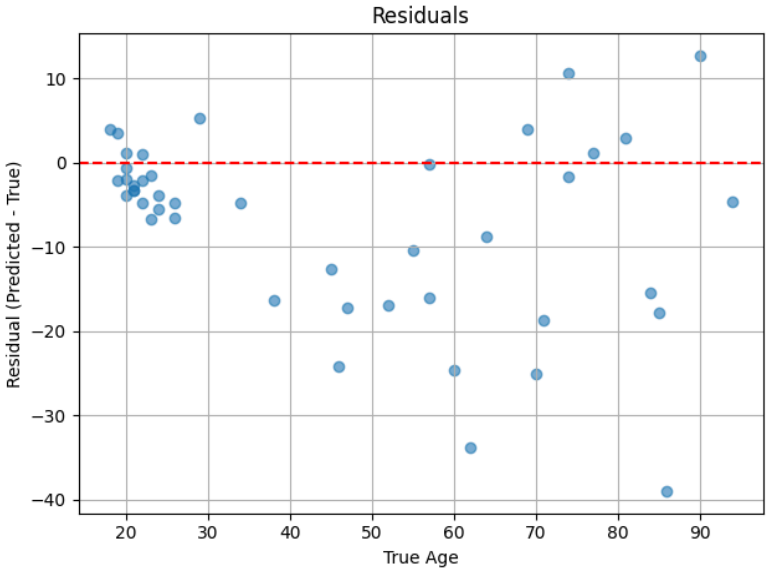
The model shows minimal bias in younger age bins (18–42), with errors under −5 years. A sharp underestimation emerges between 42 and 72, reaching up to −17 years, indicating consistent age regression in mid-to-late adulthood despite boosting strategies. Interestingly, the 72–83 bin shows mild overestimation (+4 years), before underestimation returns in the 83–94 range (−13 years). These results highlight the model's difficulty in accurately predicting middle-aged and older individuals, likely due to data imbalance and regression toward the mean.

Figure 3. Age Distribution Across Data Splits



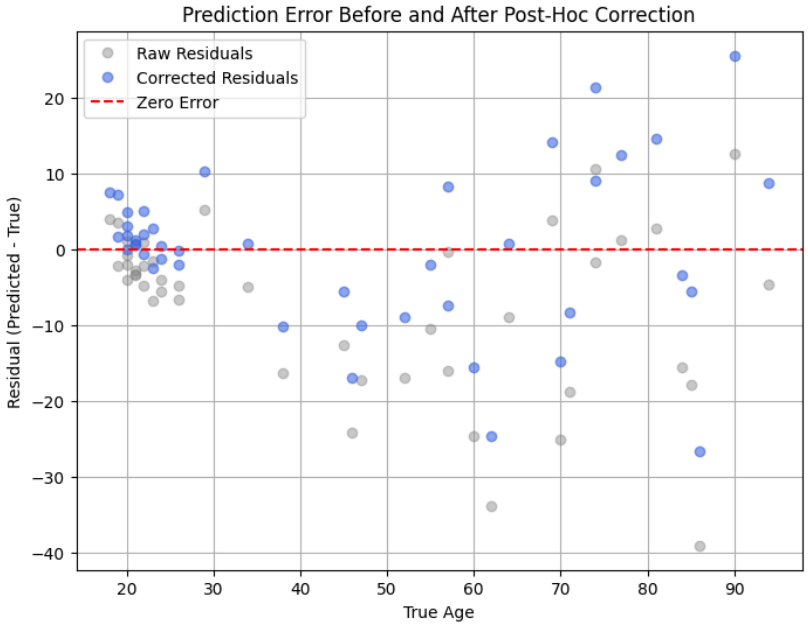
The age distribution across splits is relatively balanced, except for notably higher density in young adults (17–30 years) and sparse representation in the 30–40 range. This imbalance likely contributes to the observed bias in older predictions. Stratified sampling ensured exposure across the lifespan, but boosting alone couldn't fully correct for underrepresented age bins.

Figure 4. Residuals vs True Age



Residuals are minimal and centered around zero for younger individuals (under ~40), indicating accurate predictions in early adulthood. A clear underestimation trend appears between ages 40 and 72, with residuals dropping as low as −15 years, especially around age 50–60. A brief overestimation emerges in the 72–83 range, followed by renewed underestimation in the oldest group (83–94). This pattern reflects the model's difficulty with mid-to-late adulthood, likely due to data imbalance and regression toward the mean.

Figure 5. Post-Hoc Prediction Error Comparison Scatter Plot



To address age-related bias, I applied Target Age Transformation (tAT) as a post-hoc calibration method. This technique models residual error by age and adjusts predictions accordingly. As shown in Figure 5, tAT significantly reduced underestimation in the 42–72 age range, where raw residuals had reached up to −17 years. After correction, residuals clustered more tightly around zero, indicating improved calibration.

tAT reduced the test MAE from 9.28 to 7.54 years and improved R^2 from 0.73 to 0.83, reflecting substantial gains in accuracy and explained variance. These results demonstrate

the value of post-hoc correction in mitigating bias and improving fairness, especially in age-imbalanced datasets.

Discussion:

Interpretation of Results:

The 3D convolutional neural network demonstrated strong predictive performance, with a post-hoc bias-corrected MAE of 7.54 years and R^2 of 0.83 on the held-out test set *(Table 1)*. These results indicate high alignment between predicted brain age and chronological age in cognitively healthy individuals. However, as shown in Figures 2 and 4, the model consistently underestimated age in middle-aged and older adults, with residuals reaching as low as −17 years between ages 42 and 72. This suggests sensitivity to age distribution imbalances within the training data, despite the use of stratified sampling and age-weighted loss.

Strengths:

A key strength of this study is the use of a 3D ResNet architecture, which processes full brain volumes rather than 2D slices, allowing the model to capture richer spatial features. Focusing on cognitively healthy participants reduced confounding from disease-related structural changes, improving interpretability. As shown in Figure 1, the model trained stably, with convergence between training and validation loss. Additionally, custom boosting and stratified sampling helped mitigate age-related bias, and the application of Target Age Transformation (tAT) further improved performance — reducing test MAE from 9.28 to 7.54 years and increasing R^2 from 0.73 to 0.83 (Table 1).

Limitations:

Despite these strengths, the dataset was skewed toward younger adults (Figure 3), with sparse representation in the 30–40 age range and fewer samples in older bins. This imbalance likely contributed to the underestimation trend observed in Figures 2 and 4. Excluding individuals with cognitive impairment also limits the model's applicability to clinical populations. Furthermore, the model's decision-making process remains a black box, as no interpretability methods (e.g., saliency maps or feature attribution) were applied. Finally, the model was only evaluated on the OASIS dataset, so its generalizability to other cohorts remains untested.

Comparison to Prior Work:

The model's performance aligns with prior studies in brain age prediction, which typically report test MAEs in the range of 4 to 7 years. The age-related bias pattern — particularly underestimation in older adults — is also consistent with findings in the literature. This suggests that the observed limitations are not unique to this model but reflect broader challenges in lifespan modeling with imbalanced neuroimaging datasets.

Conclusion:

This study developed a 3D convolutional neural network to estimate brain age from structural MRI scans, with a focus on minimizing age-related prediction bias. The model achieved strong performance on a held-out test set of cognitively healthy individuals, with a post-hoc MAE of 7.54 years and an R^2 of 0.83. While raw predictions showed systematic underestimation in middle-aged and older adults, this bias was substantially reduced through Target Age Transformation (tAT), improving both accuracy and fairness across the lifespan.

These findings contribute to the growing field of brain age modeling, which offers a powerful, non-invasive biomarker of brain health. Accurate brain age estimation can support early detection of neurodegenerative conditions, track cognitive aging, and inform personalized interventions. By addressing bias and improving calibration, this work enhances the reliability and interpretability of brain age predictions — key steps toward clinical translation.

Future directions include applying interpretability methods to uncover the neuroanatomical features driving predictions, validating the model on external datasets, and extending the approach to clinical populations to assess its utility in detecting pathological aging.