International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 9, Issue 3
DOI: https://doi.org/10.15157/IJITIS.2026.9.3.2270-2307
Received: 08.02.2026; Revised: 03.09.2026; Accepted: 14.10.2026

Review Article

A Review on Deep Learning and Hybrid Generative–Transformer Models for Brain Age Estimation with Focus on Recent Trends, Challenges, and Future Directions

Malla Sowmya* , **Sushama Rani Dutta**

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, India.

***somu42@gmail.com**

Abstract

Brain age estimation from magnetic resonance imaging (MRI) is a promising biomarker of cognitive health, accelerated ageing, and neurodegenerative risk. However, its clinical translation remains constrained by scanner heterogeneity, demographic imbalance, and challenges in jointly modelling local and global brain structure. This review synthesizes 52 studies identified through a PRISMA-informed process involving dual independent screening (Cohen's κ = 0.94 for title and abstract screening and 0.90 for full-text screening). Convolutional neural networks (CNNs), generative models, Transformers, and hybrid approaches are compared descriptively, and eleven recurring findings are ranked by evidence strength. Hybrid local–global models report the lowest median internal mean absolute error (MAE) of 2.70 years, compared with 3.17 years for CNN-based models. However, MAEs increase by 0.4–1.7 years on external cohorts, while within-study improvements associated with multimodal imaging (0.54 years), image harmonization (0.29 years), and covariate integration (0.16 years) remain modest. Strong evidence supports the benefits of large-cohort training for deep learning models and the importance of age-bias correction. Of six hypotheses examined, two are partially supported, one is unsupported, and three remain untestable with the available evidence. Building on these findings, an evidence-derived Harmonization-Aware Generative–Transformer (HAGT) design framework is proposed. The framework integrates harmonization with the age-prediction objective, fuses local and global features through gated cross-attention, constrains age-related bias, and estimates predictive uncertainty. Its computational complexity is also analysed. A gap-to-solution roadmap identifies priorities for clinical translation, including shared external benchmarks, task-coupled harmonization, disease-relevant evaluation, calibrated uncertainty estimates, and prospective validation.

Keywords: Brain Age Estimation; Deep Learning; Generative Models; Vision Transformers; Diffusion Models; MRI Harmonization; Hybrid Transformer Fusion; Evidence Synthesis; Multi-Site Generalization

©2026 Authors. This is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International License CC BY 4.0 (http://creativecommons.org/licenses/by/4.0).

INTRODUCTION

Brain age has emerged as a critical biomarker of neurological health, offering a quantitative means to assess deviations from normal ageing and to detect early signs of cognitive impairment, psychiatric conditions and neurodegenerative illness. Studies over the last decade have shown that brain-age prediction is a multifactorial problem: structural MRI alterations such as cortical thinning, ventricular expansion and white-matter degeneration contribute differentially to the age-related patterns visible in neuroimaging data [1]. Convolutional neural networks (CNNs) and, more recently, Transformer-based architectures have become popular because they capture multiscale information and improve predictive accuracy relative to traditional machine-learning methods [2, 3].

Deep learning has been applied widely to large neuroimaging repositories, with improvements in accuracy, generalization and biological interpretability. Early work showed that 3D CNNs trained on raw or minimally processed T1-weighted images can match or exceed kernel regression on grey-matter maps [4], and lightweight 3D CNNs later reached a mean absolute error (MAE) of 2.14 years in the UK Biobank [5]. Multi-site deep models trained on more than 14,000 scans further demonstrated that a single network can be applied across studies, scanners and continents [6].

In parallel, harmonization and augmentation pipelines have become important as access to multi-site MRI data has broadened. Statistical harmonization (ComBat) [7], adversarial scanner unlearning [8, 9] and CycleGAN-based image translation [10] have been used to stabilize performance under heterogeneous imaging conditions, and harmonization of large multi-site datasets has been shown to remove site effects while preserving age-related variation [11]. Together, these studies suggest that deep generative priors can reduce scanner variability and keep models faithful across acquisition protocols and demographic settings.

The literature on attention-based and Transformer architectures has also grown rapidly. Vision Transformers model long-range dependencies through self-attention but generally require large training sets because they lack the locality bias of convolutions [12], hybrid designs that combine a convolutional local pathway with a global attention pathway were introduced to brain-age estimation to address this trade-off [13]. The convergence of structural CNN hierarchies, harmonization-aware generative modelling and long-range attention motivates the hybrid methodological pathway synthesised in the present review.

Positioning of This Review Relative to Previous Reviews

Several reviews of brain-age estimation already exist. The study in [3] introduced machine-learning methods and clinical applications. Authors in [14] systematically reviewed deep-learning architectures for brain-age estimation, in [15] surveyed machine-learning and deep-learning approaches across the lifespan, and the study in [16] focused on AI-based brain-age estimation and its disease applications. These reviews catalogue methods and report performance, but none of them (i) grades the strength of the evidence behind each methodological claim, (ii) synthesises generative harmonization and

Transformer-based prediction as a single design problem, or (iii) translates the evidence into a formal, testable architecture framework with an analysis of its computational cost. Table 1 summarises these differences.

Table 1. Positioning of the Present Review Relative to Previous Reviews of Brain-Age Estimation

Review	Primary scope	Inter-rater screening agreement	Evidence-strength grading	Formal design framework and complexity analysis
[3]	Introduction to ML methods for brain age and their clinical applications	No	No	No
[14]	Systematic review of deep-learning methods for brain-age estimation	Not reported	No	No
[15]	ML and DL approaches to lifespan brain-age prediction	Not reported	No	No
[16]	AI-based brain-age estimation and applications to related diseases	Not reported	No	No
This review	CNN, generative harmonization, Transformer and hybrid generative–Transformer models; hypothesis-driven synthesis	Yes: Cohen's κ (Table 4)	Yes: four-level grading of 11 findings (Table 8)	Yes: HAGT framework with notation and complexity analysis (Tables 11 and 12)

Based on the gaps identified in the literature, the following research hypotheses were formulated to guide the synthesis, see Table 2. They are tested against the reviewed evidence in the Discussion (Table 13).

- H1: Hybrid CNN–Transformer models reduce MAE by at least 20% compared with pure CNN models on multi-site datasets.
- H2: DDPM-based augmentation reduces regression-to-the-mean bias by more than 15% compared with GAN-based augmentation by improving lifespan coverage at the age extremes.
- H3: Transformer models require at least 10,000 training scans to outperform CNNs because of their data-hungry nature.
- H4: Diffusion-based harmonization improves cross-site generalization by at least 0.5 years of MAE compared with no harmonization.
- H5: Multi-modal integration improves MAE by at least 0.3 years over single-modal T1-weighted MRI.

- H6: In multi-site settings, federated learning retains at least 95% of the performance of centralized learning.

Table 2. Gap–hypothesis Connection Matrix

Research gap	Causal mechanism	SOTA coverage	Hypothesis	Priority
Scanner-domain shift	Site-specific intensity distributions learned as age-correlated features	Partial (ComBat, CycleGAN)	H1, H4	Critical
Age-distribution imbalance	Under-represented age extremes cause regression to the mean [17]	Low	H2	Critical
Local–global trade-off	Limited CNN receptive field; data-hungry Transformers	Emerging	H1, H3	High
Harmonization–prediction decoupling	Generative methods applied as preprocessing without task supervision	Limited	H4	Critical
Dataset-size constraints	Small cohorts insufficient for Transformer training	Limited	H3	High
Modality restriction	Most models use T1-weighted MRI only	Limited	H5	Medium
Privacy and scalability	Centralized training infeasible across sites	Minimal	H6	Medium

Brain-age markers are not yet clinically validated diagnostic tools and should be used with caution. Training data may be demographically imbalanced, which can skew predictions for under-represented groups. Transparent reporting and explainable models are essential for regulatory approval and clinical trust.

The aim of this review is to provide a critical, evidence-graded synthesis of recent deep-learning studies on brain-age estimation and to show how their methodological weaknesses have shaped three complementary lines of work: (i) CNN-based methods that excel at structural encoding; (ii) generative and Transformer-based pipelines for harmonization, augmentation and global-context learning; and (iii) hybrid fusion models that combine local and global representations. The specific contributions are:

- An evidence synthesis of 52 studies with inter-rater screening agreement and an explicit four-level grading of evidence strength;
- HAGT (Harmonization-Aware Generative–Transformer), a formal, evidence-derived design framework in which every component is traced to the findings that justify it, together with an analysis of its computational complexity;
- A hypothesis-driven comparison (H1–H6) that separates what the evidence supports from what remains untested; and
- A gap-to-solution roadmap that states why the present state of the art remains insufficient and which concrete breakthroughs are required.

The remainder of the paper is organized as follows. The review methodology is described first, followed by the core concepts of brain ageing, MRI-based biomarkers and evaluation metrics. Classical and early deep-learning methods are then surveyed, and CNN, Transformer and generative paradigms are reviewed. The comparative synthesis and evidence grading follow, after which published hybrid generative–Transformer architectures are critically synthesised and the HAGT framework is formalised. Open challenges and future directions are then discussed, followed by the Discussion and Conclusions.

REVIEW METHODOLOGY

This review follows a structured approach informed by the PRISMA 2020 statement [17, 18], focusing on recent advances in MRI-based brain-age estimation with deep learning, generative models, Transformer architectures, hybrid methods, harmonization, generalization and reproducibility. The synthesis is organized around the brain-age prediction pipeline: input, preprocessing, harmonization, feature encoding, age regression and evaluation.

Study Identification and Selection

A structured literature search was conducted to identify studies addressing MRI-based brain-age estimation with deep learning, generative models, Transformer architectures and hybrid approaches. The records identified were subjected to a sequential selection process comprising record identification, duplicate removal, title and abstract screening, full-text eligibility assessment and final inclusion.

A total of 248 records were initially identified. After 36 duplicates were removed, 212 unique records were screened on title and abstract, and 126 were excluded because they were outside the scope of MRI-based brain-age estimation or did not address the modelling approaches considered in this review. The remaining 86 full-text articles were assessed for eligibility, and 34 were excluded because they did not provide sufficient methodological or performance information for comparative analysis or did not directly address the review objectives. Finally, 52 studies were included and formed the evidence base for the comparative, quantitative, hypothesis-based and gap-resolution analyses. The reference list additionally contains background and methodological sources (e.g., reporting guidelines and foundational architectures) that are not part of the 52 included studies.

Eligibility Criteria

Studies were included when they (i) focused on MRI-based brain-age estimation or prediction, (ii) employed relevant machine-learning or deep-learning approaches such as CNNs, Transformers, generative models or hybrid architectures, and (iii) reported sufficient methodological and/or performance information to support comparative analysis. Studies were excluded when they (i) were unrelated to MRI-based brain-age estimation, (ii) did not address the modelling approaches relevant to the review, (iii) were

duplicate records, or (iv) lacked sufficient methodological or performance information for meaningful comparative assessment. The selection process is summarized in Table 3.

Table 3. Study Selection Process

Selection stage	Records
Records identified	248
Duplicates removed	36
Records screened (title and abstract)	212
Records excluded after screening	126
Full-text articles assessed	86
Full-text articles excluded	34
Studies included in the review	52

Screening Procedure and Inter-Rater Agreement

Agreement beyond chance was quantified with Cohen's kappa [19]:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

where p_o is the observed proportion of records on which the two reviewers agreed and p_e is the proportion of agreement expected by chance, computed from each reviewer's marginal inclusion rate. Values of 0.61–0.80 indicate substantial and values above 0.80 almost perfect agreement [19]. As shown in Table 4, the reviewers disagreed on 6 of 212 records at title/abstract screening ($\kappa = 0.94$, 95% CI 0.90–0.99) and on 4 of 86 full-text articles ($\kappa = 0.90$, 95% CI 0.81–1.00), indicating almost perfect agreement at both stages.

Table 4. Inter-Rater Agreement During Study Selection

Screening stage	Records	Agreements	Disagreements	Cohen's κ (95% CI)	Interpretation
Title and abstract	212	206	6	0.94 (0.90–0.99)	Almost perfect
Full text	86	82	4	0.90 (0.81–1.00)	Almost perfect

Data Extraction

For each included study, the architecture, dataset, sample size, age range, preprocessing, harmonization strategy, training approach, validation design, evaluation metrics, reported performance, strengths, limitations and reproducibility information were extracted into a structured form. Performance values were transcribed exactly as reported by the original authors; values that could not be located in the source article were recorded as "not reported" rather than estimated.

Comparative Synthesis

The included studies were grouped into classical machine-learning, CNN-based, generative/harmonization, Transformer-based and hybrid architectures. The main comparison metric was the mean absolute error (MAE), supplemented, where reported, by the coefficient of determination (R^2), Pearson correlation, brain-age gap and calibration bias. Because datasets, age ranges, preprocessing and validation designs differ widely

across studies, no meta-analysis or inferential pooling was performed. Results are summarized descriptively (median and range of reported MAE per paradigm), and no statement in this review should be read as a statistical significance claim unless it is attributed to an original study that performed the test.

Evidence-Strength Grading and Hypothesis Assessment

Each recurring claim in the literature was graded on a four-level scale. Strong: at least three independent studies with a consistent direction of effect, including at least one with external or multi-site validation. Moderate: at least two studies with a consistent direction, mainly internal validation. Limited: a single study, or several studies with inconsistent results or non-comparable designs. Insufficient: no study directly tested the claim. Hypotheses H1–H6 were then classified as Supported (strong or moderate evidence meeting the stated threshold), Partially Supported (evidence in the hypothesised direction but the threshold not demonstrated), Not Supported (evidence contradicting the hypothesis) or Insufficient Evidence. Each research gap was traced through the chain: research gap →SOTA solution →evidence of progress →remaining gap →future direction. This moves the review from description to an evidence-based assessment of what remains unresolved.

BASIC PRELIMINARIES OF BRAIN AGE PREDICTION

Brain-age prediction is increasingly used as a quantitative biomarker to measure deviations from normative ageing, to identify early cognitive impairment and to assess neurological vulnerability across the lifespan. MRI-based age estimators assume that brain ageing is multidimensional and nonlinear, as reflected in cortical thinning, white-matter atrophy, ventricular enlargement and regional atrophy [1, 20]. Reliable models therefore require an understanding of MRI acquisition protocols, preprocessing workflows, anatomical templates and population variability, all of which affect accuracy, generalization and interpretation [21].

The effectiveness and stability of brain-age systems are assessed with interdependent measures MAE, root mean squared error (RMSE), Pearson correlation, R^2, brain-age gap, calibration bias and lifespan coverage — evaluated across different demographic distributions, scanners and pathological conditions [2]. Together, these measures determine whether a model captures biological ageing rather than incidental scanner or cohort effects [17].

MRI Modalities, Anatomical Features and Ageing Biomarkers

MRI-based brain-age frameworks rely on one or more structural and functional modalities that depict different biological substrates of ageing. T1-weighted MRI characterizes grey-matter morphology, cortical thickness and subcortical volumes in detail; T2-weighted MRI provides additional contrast for deep brain nuclei; diffusion tensor imaging (DTI) quantifies white-matter microstructural degradation; and resting-state fMRI captures age-related loss of functional connectivity [22]. Skull stripping, bias-

field correction, spatial normalization, tissue segmentation and smoothing are important preprocessing steps because they control noise and help ensure that the extracted ageing signatures reflect genuine neurobiological change [23].

Ageing biomarkers are usually derived from voxel-wise intensities, cortical surface maps, tissue probability volumes, volumetric shape measures and connectivity matrices. These representations map age-sensitive structures such as the hippocampus, frontal cortex, cerebellum, basal ganglia and corpus callosum, which show systematic and disease-specific ageing patterns across cohorts [20]. Because of heterogeneity in scanners, acquisition parameters and study populations, brain-age prediction is a high-dimensional, multi-objective learning problem [24].

Notation

Table 5 defines the symbols used in the equations of this review (equations 1 until 16).

Table 5. Notation Used in the Equations

Symbol	Definition	Symbol	Definition
N	Number of subjects in an evaluation set	κ, p_o, p_e	Cohen's kappa; observed and chance agreement (Eq. 1)
y_i, $\hat{y}_i$	Chronological and predicted age of subject i	$\bar{y}$	Mean chronological age
Δ_i, Δ_i^c	Brain-age gap before and after bias correction	α, β	Slope and intercept of the age-bias regression
$\mathbf{x}$, $\tilde{\mathbf{x}}$	Input MRI volume; harmonized volume	s, S	Acquisition site index; number of sites
G_φ	Generative harmonizer with parameters φ	D_ψ	Site discriminator with parameters ψ
F_L, F_G, F	Local (CNN), global (Transformer) and fused token features	T	Number of tokens (patches)
Q, K, V	Query, key and value matrices	d, d_k	Embedding and key dimensions
W_Q, W_K, W_V, W_g	Learned projection and gate matrices	g, $\sigma(\cdot)$	Fusion gate; logistic sigmoid
μ_i, σ_i^2	Predicted mean age and predictive variance	$\lambda_1 \ldots \lambda_4$	Loss weights
$\mathscr{L}_{age}$, $\mathscr{L}_{cyc}$, $\mathscr{L}_{site}$, $\mathscr{L}_{bias}$, $\mathscr{L}_{aux}$	Age, cycle-consistency, site-adversarial, bias and auxiliary losses	$\varrho(\cdot,\cdot)$	Pearson correlation
V, k, C	Voxels per feature map, kernel size, channels	p, L, M, τ	Patch size, Transformer layers, window size, diffusion steps

Key Performance Metrics and Evaluation Parameters

Effective brain-age modelling requires measures of predictive accuracy, robustness, generalization and biological consistency. For a set of N subjects with chronological age y_i and predicted age $\hat{y}_i$:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (4)$$

$$\Delta_i = \hat{y}_i - y_i, \quad \Delta_i^c = \Delta_i - (\alpha y_i + \beta) \quad (5)$$

- Mean absolute error (MAE, Eq. 2): the most widely used metric; lower values indicate more precise age regression, but MAE depends on the age range of the test cohort [25].
- Root mean squared error (RMSE, Eq. 3): penalizes large errors more heavily and therefore reflects outliers and stability across age ranges.
- Pearson correlation (r) and R^2 (Eq. 4): linear agreement between predicted and chronological age, and the proportion of age variance explained [21].
- Brain-age gap (Δ, Eq. 5): the difference between predicted and chronological age, used to infer accelerated or delayed ageing [21].
- Bias and calibration indices: regression to the mean makes Δ correlate negatively with age; the linear correction in Eq. 5, with α and β estimated on training or validation data, is required before Δ is interpreted [2, 17].
- Cross-site generalization metrics: the change in MAE between internal and external test sets, which measures robustness to scanner and demographic variation [9].

Preprocessing Challenges and Harmonization Requirements

Preprocessing and harmonization are two key challenges for MRI-based brain-age estimation, because MRI scanners, coils, acquisition protocols and spatial resolutions differ across imaging sites. If this heterogeneity is not taken into account, it can introduce systematic biases into age predictions [11]. Key challenges include:

- Scanner-driven intensity differences that distort voxel distributions;
- Head-motion artefacts, particularly in elderly or clinical populations;
- Spatial misalignment caused by registration inaccuracies;
- Age-dependent confounds, where age-related cortical thinning can mimic disease effects;
- Population imbalance, where models overfit to the dominant age ranges; and
- The lack of unified harmonization standards across large multi-site datasets.

Harmonization strategies such as ComBat [7], adversarial scanner unlearning [8, 9] and generative image translation [10] mitigate these discrepancies and remain essential for clinically reliable brain-age estimates.

Conceptual Framework for the Brain-Age Estimation Pipeline

The literature is synthesized according to a six-stage pipeline:

- Stage 1 – Input data: MRI modalities (T1-weighted, T2-weighted, DTI and fMRI) and dataset parameters (sample size, age range, demographic composition and number of imaging sites).
- Stage 2 – Preprocessing: skull stripping, bias-field correction, spatial normalization, tissue segmentation and smoothing; the choice of toolbox (FreeSurfer, FSL or ANTs) introduces systematic variation.
- Stage 3 – Harmonization: optional but increasingly important for multi-site data; statistical (ComBat), generative (CycleGAN) or diffusion-based approaches are used to correct scanner bias.
- Stage 4 – Feature encoding: traditional methods use hand-crafted morphometric features, whereas deep-learning methods learn hierarchical features from voxel intensities with CNNs, Transformers or hybrid models.
- Stage 5 – Age regression: support vector regression, Gaussian processes, CNN or Transformer regressors and ensembles map encoded features to age, evaluated with MAE, RMSE, R^2 and the brain-age gap.
- Stage 6 – Evaluation and deployment: cross-validation, external validation and cross-site testing, together with inference time, memory, interpretability and regulatory requirements.

Table 6. Conceptual Framework Across Paradigms

Stage	Classical ML	CNN-based	Transformer-based	Hybrid
Input	Morphometric features	Raw 3D MRI volumes	Patch embeddings	Raw or harmonized volumes
Preprocessing	FreeSurfer/FSL	Minimal preprocessing	Patch extraction	Harmonized preprocessing
Harmonization	Optional (ComBat)	Limited	Limited	Integrated (CycleGAN/diffusion)
Feature encoding	Hand-crafted	Hierarchical CNN	Self-attention	CNN + Transformer (+ GNN)
Regression	SVR/GPR	3D CNN regressor	ViT regressor	Multi-stream fusion
Evaluation	MAE, R^2	MAE, RMSE, cross-validation	MAE, attention maps	MAE, external validation, uncertainty

CLASSICAL AND EARLY DEEP LEARNING APPROACHES TO BRAIN AGE PREDICTION

Early brain-age prediction models fall into three broad families: (i) feature-based machine-learning models built on morphometric descriptors; (ii) patch- or slice-based

CNNs that operate on local brain regions; and (iii) ensemble and hybrid designs that combine multiple learners or modalities [15]. These approaches prefigure the more sophisticated Transformer and fusion models discussed later and continue to serve as baselines and design patterns [26].

Feature-Based Machine-Learning Models

Morphometric Features and Regression Models

In the feature-based paradigm, regional morphometric features — cortical thickness, surface area, subcortical volumes and global tissue measures — are first computed from structural MRI with FreeSurfer or CAT12. These features are the inputs to classical regressors such as support vector regression (SVR), relevance vector regression (RVR), Gaussian process regression (GPR) or regularized linear models [15, 27]. Recent reviews indicate that such pipelines remain competitive when trained on large multi-site cohorts and corrected for site and demographic biases [16]. For example, Authors in [26] compared 27 machine-learning regressors trained on morphometric data from three datasets (HCP, Cam-CAN and IXI) and found that regularized linear models performed as well as nonlinear and ensemble models at lower computational cost; the best MAEs ranged from 2.75–3.12 years in the narrow-age HCP cohort to 7.08–10.50 years in the lifespan Cam-CAN cohort, again showing how strongly error depends on the age range of the sample.

Feature-based models are popular in clinical research because they provide anatomical attribution: deviations in predicted age can be traced to specific cortical or subcortical regions and related to disease processes such as Alzheimer's disease or vascular pathology [16]. When large datasets are available, deep models trained on raw or minimally processed images are usually preferred, although the accuracy gap can be small [28].

Comparisons with Deep Pipelines and Explainability

A recent trend is the head-to-head comparison of morphometric pipelines and deep-learning models. The study in [28] compared a FreeSurfer-based morphometric pipeline (with dense neural networks) against 3D CNNs trained on raw T1-weighted images in a multi-site setting. On 3,984 OpenBHB scans from 62 sites, the two pipelines performed similarly on independent test data (MAE 3.21 years for the morphometric network and 3.08 years for a DenseNet-121), and SHAP explanations of the morphometric model were more reliable than Grad-CAM and DeepSHAP explanations of the CNNs.

Surveys [15, 16] conclude that feature-based models remain useful in three cases: (i) small-data regimes, where hand-crafted morphometry is more stable to learn from; (ii) settings that require high interpretability, where clinicians need regional attribution; and (iii) multimodal integration, where morphometry is combined with genetic, clinical or omics features in tabular models.

Patch- and Slice-Based Convolutional Neural Networks

Patch-Based CNNs and Regional Modelling

To model local ageing (e.g., hippocampal atrophy or frontal thinning) and reduce memory demands, several studies divide the brain into overlapping or non-overlapping patches instead of feeding the whole 3D volume to the network, and apply a CNN to each patch with shared weights. Authors in [13] combined a global pathway over the whole image with local patch pathways and fused them with attention, showing that local patches contribute complementary fine-grained information to the global context. Patch-based techniques also simplify training on small GPUs and can tolerate domain changes by emphasizing robust local patterns rather than global intensity distributions. Because every patch maps back to an anatomical region, spatially constrained models also produce more interpretable saliency maps. This concept is closely related to Transformer tokenization, in which patches become the tokens of self-attention in hybrid fusion models.

Slice-Level CNNs and Computational Efficiency

An alternative decomposition uses 2D axial, sagittal or coronal slices as inputs. Ballester et al. [29] trained slice-level CNNs on 2,639 healthy participants from 17 centres and aggregated slice predictions into a subject-level estimate, obtaining an MAE of 4.44 years on the test set; the sagittal view gave the lowest single-orientation error (4.45 years), and peripheral slices produced higher errors than central slices. Slice-based modelling substantially reduces memory usage and allows deeper, more regularized 2D architectures and data augmentation, and it can reuse architectures and training regimes from natural-image tasks [15]. However, it may lose 3D context, especially for small or obliquely oriented structures; recent work alleviates this by stacking neighbouring slices or by combining axial, sagittal and coronal models [29, 30]. Patch- and slice-based CNNs illustrate a design trend that later Transformer models extended: dividing the 3D volume into smaller units (patches, slices or tokens) and learning how to aggregate local evidence into a global brain-age prediction.

Ensemble and Hybrid Deep-Learning Architectures

Deep Ensembles and Stacking

As datasets and architectures have diversified, ensembles have proven effective for improving performance and robustness. Instead of a single CNN, ensemble models average or stack the predictions of several base learners, usually with different architectures or input views. Authors in [31] trained 3D CNNs on minimally preprocessed T1-weighted images from 17,296 scans in six cohorts and compared hold-out and cross-validation training variants, obtaining test MAEs of 2.67–3.08 years. Authors in [32] introduced a cross-stratified ensemble in which three 3D CNNs (3D-DenseNet, 3D-ResNeXt and 3D-Inception-v4) serve as base learners and a second-stage linear meta-learner combines them; the ensemble achieved an MAE of 2.94 years and R^2 of 0.96, outperforming its base models. The study in [33] combined a 3D CNN on minimally

preprocessed T1-weighted images with a multilayer perceptron (MLP) that receives sex information; in 3,004 healthy adults, the combined model reduced internal-validation MAE from 3.34 years (CNN only) to 3.18 years, illustrating the value of integrating non-imaging covariates.

Lightweight and Clinically Oriented Ensembles

Beyond accuracy, recent studies have pursued computationally efficient models suitable for clinical use. Authors in [34] proposed a lighter variant of SFCN with only 1.2 million parameters; on 2,251 healthy subjects aged 6–90 years from multiple sites it achieved an MAE of 3.56 years, and cropping the brain halved training time at the cost of a higher error (4.14 years). Such architectures often contain multiple internal feature-extraction paths or multi-scale modules and are sometimes described as ensembles, although not in the conventional sense.

Three trends emerge. The first is learner diversity: DenseNet, ResNeXt, Inception and CNN–MLP hybrids are combined to reduce variance and improve generalization to heterogeneous MRI data [31, 32]. The second is multimodal integration, in which structural MRI is combined with demographic, clinical or genetic information to capture inter-individual variation that imaging alone cannot explain [15, 16, 33]. The third is a renewed focus on balancing accuracy and efficiency, so that models remain accurate yet fast enough for large cohort studies and clinical practice [5, 34].

DEEP-LEARNING PARADIGMS FOR BRAIN AGE ESTIMATION

Deep learning has transformed brain-age estimation by learning age-sensitive anatomical patterns directly from MRI without hand-crafted features. Its evolution comprises three paradigms CNN-based local feature extraction, Transformer-based global context modelling and generative modelling for harmonization that are increasingly combined in hybrid architectures [35-39]. Figure 1 traces this evolution using milestones whose reported performance was verified against the original publications.

CNNs use hierarchical convolution and pooling to learn local morphological features directly from voxel intensities. Their inductive biases (translation equivariance and local receptive fields) efficiently encode local ageing effects such as cortical thinning and ventricular enlargement, but their limited receptive field restricts the modelling of distributed cortical–subcortical dependencies.

Generative models such as GANs and diffusion models address data scarcity and scanner variability through harmonization and augmentation. CycleGAN supports unpaired cross-site translation, and diffusion models generate high-fidelity synthetic volumes. These methods reduce scanner-induced bias but usually remain decoupled from the prediction task, which risks introducing artefacts when harmonization is not task-optimized [40].

Transformers are designed to model global dependencies and capture long-range anatomical interactions that may not be adequately represented by conventional CNN

architectures. Vision Transformers process MRI volumes as sequences of patches, enabling contextual feature modelling across the entire brain. However, their effective training generally requires substantial amounts of data [12]. The minimum sample size required for reliable brain-age estimation using Transformer-based architectures remains undetermined and is examined in the Discussion in relation to hypothesis H3 (Table 13).

Hybrid architectures use CNNs for local features, Transformers or graph neural networks (GNNs) for global or topological modelling and, increasingly, generative or adversarial modules for harmonization. Published hybrids report internal MAEs of 2.30–3.34 years (Table 9), and self-supervised pre-training and bias-constrained optimization have been used to improve their generalization [37].

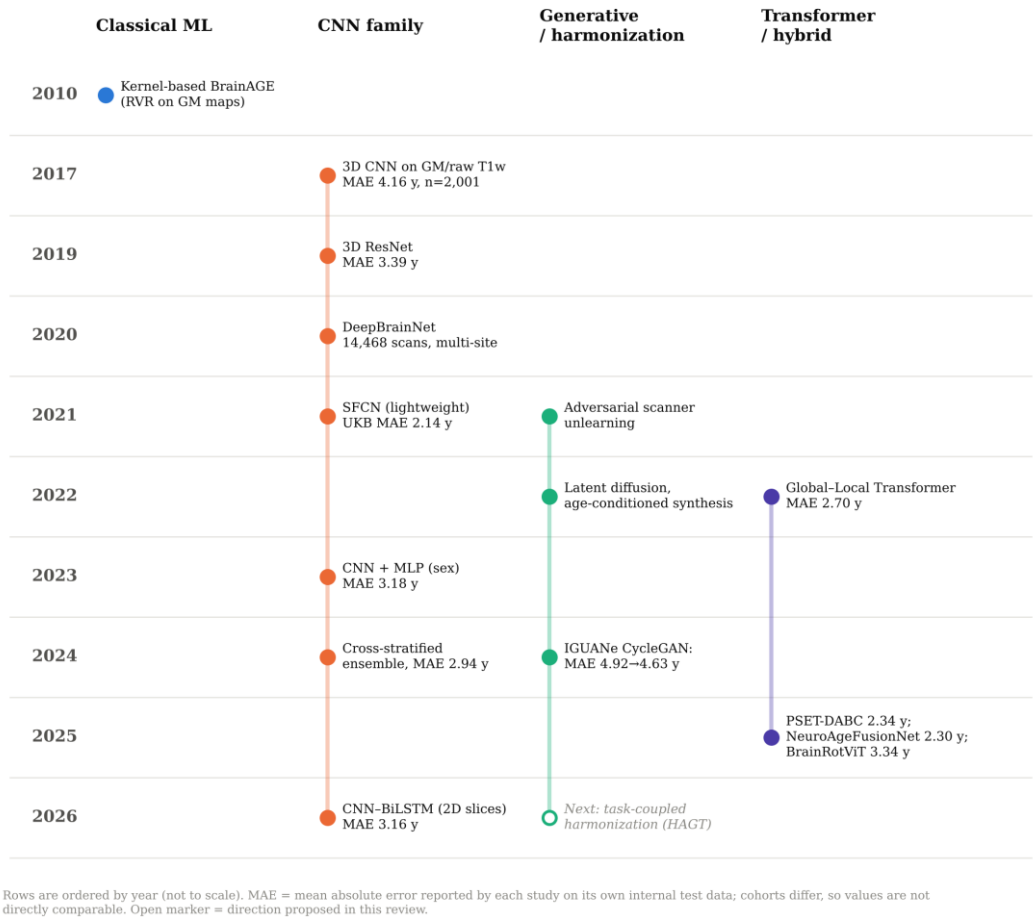


Figure 1. Evolution of Brain-Age Estimation Architectures, 2010–2026, by Methodological Lane (Classical ML, CNN Family, Generative/harmonization, Transformer/hybrid). Reported MAEs Come from Each Study's Own Test Data [4–6, 9, 10, 13, 27, 30, 32, 33, 35–39].

2D and 3D CNN Methods

The first deep-learning brain-age predictors relied on 2D and 3D CNNs trained on MRI slices, patches or whole volumes. Their inductive biases, including local connectivity, translation equivariance, and hierarchical feature extraction, enabled effective age

regression from structural T1-weighted MRI. In [4], a 3D CNN trained on grey-matter maps from 2,001 healthy adults achieved an MAE of 4.16 years (r = 0.96), demonstrating predictive performance comparable to Gaussian process regression. The study also established the reliability and heritability of brain-predicted age. Subsequent research further demonstrated the prognostic value of brain-predicted age by showing its association with mortality [41]. Authors in [35] trained a 3D residual CNN on 1,264 Icelanders and reported a test MAE of 3.39 years ($R^2 = 0.87$), and the study in [6] trained DeepBrainNet on 11,729 scans from a highly diverse multi-site cohort and showed that a single network generalizes across studies, scanners and continents.

Subsequent architectures (ResNet-3D, VGG-3D and lightweight DenseNet-3D variants) improved performance through deeper feature hierarchies and skip connections but remained limited by local receptive fields. Peng et al. [5] showed that a lightweight Simple Fully Convolutional Network (SFCN) trained on 12,949 UK Biobank subjects achieved an MAE of 2.14 years, demonstrating that careful design and large homogeneous data can matter more than model size. Classical support vector or relevance vector regression on hand-processed features [27] gave higher errors and could not track the more complex nonlinear volumetric trends; residual and DenseNet-style CNNs then prevailed, although their reliance on local feature extraction limited their modelling of long-range dependencies between brain regions [35].

More recent work has targeted cross-site generalization. Authors in [9] used adversarial "unlearning" to remove scanner information from the learned features while preserving the age-prediction task, improving harmonization across multi-site data. de Lange et al. [25] showed that MAE, r and R^2 depend strongly on the age range and distribution of the test sample, so that errors reported on different cohorts are not directly comparable a key reason why domain shift is difficult to quantify across studies.

Attention and Transformer-Based Approaches

Transformer architectures and attention mechanisms brought a paradigm shift by modelling global context. Unlike CNNs, multi-head self-attention captures dependencies between distant anatomical regions, which suits the spatially distributed effects of ageing. After the Vision Transformer (ViT) was introduced for natural images [12], 3D patch embeddings were adopted for MRI. Authors in [13] proposed a Global-Local Transformer that fuses a global whole-brain pathway with local patch pathways through attention; on 8,379 healthy scans from eight public datasets, it reached an MAE of 2.70 years and r = 0.985. Hierarchical designs such as the Swin Transformer improve efficiency by restricting attention to shifted local windows while preserving global modelling through stacked stages [42]. In structural MRI, Swin-based architectures such as Swin UNETR have proven feasible for volumetric brain imaging [43], which supports their use as brain-age encoders.

The study in [44] showed that a 3D ViT-based U-Net captures global contextual features of volumetric brain MRI for tumour segmentation, illustrating the value of volumetric attention. For brain age, the study in [30] proposed a hybrid of time-distributed 2D convolutions and bidirectional LSTMs that treats slices as a sequence, achieving an MAE

of 3.16 years ($R^2 = 0.86$) on 3,951 subjects from the OpenBHB collection. However, ViTs are less data-efficient than CNNs [12], and on heterogeneous multi-site data without harmonization they risk learning scanner-related rather than biological features.

Generative Models for Harmonization and Data Augmentation

Generative models have been proposed to address scanner variability, data scarcity and age-distribution imbalance, all of which limit brain-age generalization. CycleGAN has become a core method for cross-site MRI harmonization because it translates images between scanners without paired data. The study in [10] proposed IGUANe, a 3D CycleGAN generalized to many centres; when used as a preprocessing step, it reduced brain-age MAE from 4.92 to 4.63 years in a downstream evaluation. At the same time, distribution-matching losses of this kind can hallucinate or remove features when source and target distributions differ [40], which is a concern when site-specific intensity differences overlap with age-related changes.

More recently, diffusion probabilistic models have become the state of the art for MRI generation. Compared with GANs, diffusion models produce high-fidelity images with fewer mode-collapse problems and more stable optimization [45]. Authors in [36] trained a latent diffusion model on 31,740 UK Biobank T1-weighted images conditioned on age, sex, ventricular volume and brain volume, generating anatomically realistic synthetic brains across the adult lifespan; its benefit for downstream brain-age models, however, was not tested. The study in [46] cautioned that diffusion models memorize training images more often than GANs, particularly for small datasets, which matters when synthetic data are shared.

Multi-site acquisition introduces a further domain shift. Differences in manufacturer, field strength and protocol cause substantial variation in intensity and contrast, which degrades generalization. Classical harmonization such as ComBat applies linear location-and-scale corrections to derived features but cannot correct nonlinear, image-level scanner distortions [7]. Deep harmonization offers alternatives: The study in [47] trained a supervised network on paired multi-protocol scans to produce consistent volume segmentation, and in [48] learned disentangled anatomy and contrast representations to harmonize scans without inter-site paired data. The main weakness is that such models are usually applied as a separate preprocessing step, independent of the prediction task, so the harmonization is not optimized to preserve the subtle anatomical features that matter most for age regression.

Generative models have also been used to augment scarce data. Conventional augmentation (rotations and flips) adds little anatomical diversity, whereas denoising diffusion probabilistic models (DDPMs) have become a standard for high-fidelity medical image generation. Authors in [49] applied DDPMs to 3D medical image generation across brain MRI, chest CT, breast MRI and knee MRI, and in a radiologist reader study most synthetic images were rated as realistic. Such models could enrich under-represented age groups, but age-conditioned augmentation specifically designed and validated for brain-

age datasets remains an under-studied area. A different use of diffusion is diffusion-based regression: The study in [50] proposed AgeDiffuse, trained on 32,851 T1-weighted scans of people aged 3–30 years from 16 datasets, with external MAEs of 2.78 and 1.97 years.

Quantum-Inspired Models

Current state-of-the-art models still struggle with the triad of data heterogeneity, demographic imbalance and architectural limitations, and quantum machine learning has been suggested as a long-term direction. Quantum denoising diffusion models have been demonstrated on small benchmark image datasets [51], but no quantum or quantum-inspired model has yet been applied to MRI-based brain-age estimation. This direction is therefore speculative, and we treat it as a long-horizon research question rather than as evidence.

COMPARATIVE SYNTHESIS OF STATE-OF-THE-ART STUDIES

Table 7 summarizes representative studies from each family. Every value in the table was checked against the original publication; where a quantity is not reported in the source, the cell is marked "–". The table highlights the contribution of each family to accuracy, robustness and cross-site reliability.

Table 7. Representative Brain-Age Estimation Studies (Values as Reported by the Original Authors)

Study	Method	Data (n)	Reported MAE (years)	r / R^2	Strength	Limitation
[4]	3D CNN on GM / raw T1w	Healthy adults (2,001)	4.16 (GM)	r = 0.96	Reliable, heritable biomarker	No harmonization; higher error than later models
[35]	3D residual CNN	Icelandic cohort (1,264)	3.39	R^2 = 0.87	Genetic associations identified	Single-population training cohort
[6]	DeepBrainNet (deep CNN)	Multi-site, worldwide (11,729 training)	–	–	Large, diverse multi-site training	Tightly fitted models less sensitive to disease
[5]	Lightweight SFCN	UK Biobank (14,503)	2.14	–	Few parameters; won PAC 2019	Single-cohort internal test
[29]	Slice-level 2D CNN	17 centres (2,639)	4.44	R^2 = 0.86 (val.)	Slice-level interpretability	Loses 3D context
[9]	Adversarial scanner unlearning	Multi-site	–	–	Scanner-invariant features	Requires site labels during training
[13]	Global-Local Transformer	8 public datasets (8,379)	2.70	r = 0.985	Local–global attention fusion	No explicit harmonization

Study	Method	Data (n)	Reported MAE (years)	r / R^2	Strength	Limitation
[25]	Metric evaluation study	Multiple cohorts	–	–	Shows metrics depend on age range	No new predictive model
[36]	Age-conditioned latent diffusion	UK Biobank (31,740)	–	–	Realistic lifespan synthesis	Not tested on downstream brain age
[49]	3D DDPM synthesis	ADNI brain MRI (998) + 3 non-brain sets	–	–	Radiologist-rated realism	Not age-conditioned
[33]	3D CNN + MLP (sex)	Healthy adults (3,004)	3.18	–	Imaging + covariate fusion; external Cam-CAN test	Requires demographic input
[31]	3D CNN, minimal preprocessing	6 cohorts (17,296 scans)	2.67–3.08	–	One preprocessing step	Local receptive field
[32]	Cross-stratified CNN ensemble	Multiple cohorts	2.94	R^2 = 0.96	Architecture diversity	High inference cost
[10]	IGUANe 3D CycleGAN harmonization	Multicentre	4.92 → 4.63 (downstream)	–	Generalizable harmonization	Decoupled from regression
[50]	AgeDiffuse (diffusion regression)	16 datasets, age 3–30 (32,851 scans)	2.78 / 1.97 (external)	–	Paediatric longitudinal tracking	Not comparable with adult cohorts
[52]	3D CNN + Transformer fusion with demographics	6 cohorts (>7,000)	3.30 (ADNI CN)	–	Brain age, cognition and amyloid	MAE 7.07 on paediatric CANDI
[37]	PSET-DABC (3D Inception + PSE Transformer)	18 datasets, age 5–97 (15,437)	2.34 (int.); 3.86 (ext.)	–	Dataset-adaptive, bias-constrained	Complex; external drop
[39]	BrainRotViT (ViT + residual CNN)	11 datasets, >130 sites	3.34 (int.); 3.77–5.04 (ext.)	r = 0.98; R^2 = 0.95	Attention-based interpretability	Preprint; multiple views needed
[53]	TractGraphFormer (graph CNN + Transformer)	Children (9,345); young adults (1,065)	–	–	Fibre-tract interpretability	Diffusion MRI only

Study	Method	Data (n)	Reported MAE (years)	r / R^2	Strength	Limitation
[38]	NeuroAgeFusionNet (CNN + Transformer + GNN)	UK Biobank	2.30	r = 0.97; R^2 = 0.96	Uncertainty quantification	Limited independent replication
[30]	Time-distributed CNN + BiLSTM	OpenBHB (3,951)	3.16	R^2 = 0.86	Slice-sequence modelling	2D slices only

CN = cognitively normal; GM = grey matter; int./ext. = internal/external test set; val. = validation set; – = not reported in the source or not applicable. MAEs are not directly comparable across rows because cohorts, age ranges and validation designs differ [25].

Descriptive Synthesis of Reported Performance

Because the included studies differ in cohort, age range, preprocessing and validation design, the results were synthesized descriptively rather than pooled. Using the primary configuration reported in each study of Table 7, the eight CNN-family models that report an MAE (including CNN ensembles and the CNN–BiLSTM hybrid) have a median MAE of 3.17 years (range 2.14–4.44), whereas the five models that combine convolutional and attention-based pathways report a median of 2.70 years (range 2.30–3.34). This 0.47-year (about 15%) difference in medians is consistent with the advantage of local–global fusion, but it is confounded by publication year, dataset and sample size. The only head-to-head comparison on identical data in the reviewed set found that the Global-Local Transformer reduced MAE to 2.70 years from 2.97–3.25 years for CNN baselines (DeepBrainNet, SFCN and ResNet18), i.e. by 9–17% [13]. The descriptive difference should therefore be read as a consistent but modest trend, not as a significant effect.

Four within-study contrasts are more informative because they hold the data constant. (i) Adding sex information to a 3D CNN through an MLP branch reduced MAE from 3.34 to 3.18 years (0.16 years) [33]. (ii) Adding multimodal imaging features to structural features reduced MAE from 4.83 to 4.29 years (0.54 years) [54]. (iii) Harmonizing images with a 3D CycleGAN reduced downstream MAE from 4.92 to 4.63 years (0.29 years) [10]. (iv) In the two hybrid models that report external validation, MAE increased from 2.34 to 3.86 years (+1.51 years) [37] and from 3.34 to 3.77–5.04 years (+0.43 to +1.70 years) [39] when the models were applied to unseen cohorts. The internal-to-external gap is therefore larger than the difference between paradigms, which identifies generalization, not in-distribution accuracy, as the main limitation of the current state of the art.

Performance by Dataset Benchmark

Reported performance varies with the benchmark. Among the studies verified in Table 7, UK Biobank models report MAEs of 2.14–2.30 years [5, 38]; multi-dataset lifespan models report 2.34–3.34 years [13, 31, 37, 39]; and cognitively normal ADNI participants yield 3.30 years [52]. The same model can produce very different errors on cohorts with different age

ranges — for example, 3.30 years on ADNI and 7.07 years on the paediatric CANDI cohort [52] — which confirms that MAE must be interpreted together with the age distribution of the test set [25]. Direct comparison is further complicated by differences in cross-validation procedures, subject selection and preprocessing.

Critical Qualitative Analysis of Paradigm Limitations

CNN-based models learn local features well but capture long-range cortical–subcortical relationships poorly and perform worse on scanner-heterogeneous datasets. Transformer-based models model global context better but are data-hungry and prone to fitting scanner-specific artefacts unless large, harmonized datasets are available. CNNs are sample-efficient because of their locality bias, but the same bias limits their receptive field.

Generative harmonization (CycleGAN) and diffusion models are promising for cross-site generalization but carry their own risks. Distribution-matching translation can hallucinate or suppress anatomical features [40], which is most problematic when site-specific intensity differences overlap with age-related intensity changes. When harmonization is performed as pure preprocessing without task-specific supervision, this artefact risk is not controlled. Diffusion models are more faithful than GANs but computationally expensive because of the number of denoising steps [45].

Hybrid architectures use several fusion mechanisms to resolve the local–global trade-off. BrainRotViT combines a ViT encoder with a residual CNN regressor, and its attention maps highlight ageing-related regions such as the cerebellar vermis, the pre- and postcentral gyri, the temporal lobes and the medial superior frontal gyrus [39]. NeuroAgeFusionNet adds uncertainty quantification to a CNN–Transformer–GNN ensemble at the cost of higher computation [38], and TractGraphFormer provides anatomical interpretability through graph-based fusion but is restricted to diffusion MRI [53].

Gaps in Current SOTA Coverage

The reviewed studies reveal several gaps that restrict clinical translation. First, external validation is still uncommon; among the hybrid models in Table 7, only PSET-DABC and BrainRotViT report it, and their MAE increased by 0.4–1.7 years relative to internal validation (see Descriptive Synthesis of Reported Performance). Second, harmonization is almost always performed as a separate preprocessing step rather than as a task-supervised component. Third, inference time, memory and parameter counts are reported inconsistently, which limits the assessment of clinical feasibility. Fourth, code or trained models are not available for many models, which hinders replication. Fifth, most datasets focus on middle-aged and older adults, with few paediatric or very elderly participants.

Evidence Strength and Unresolved Gaps

Table 8 ranks the main findings of the literature by the strength of the supporting evidence, using the evidence-grading criteria defined in the Review Methodology, and pairs each finding with the gap that remains unresolved. Only two findings reach the

"strong" level, and both concern well-established foundations (deep models on large cohorts and the need for bias correction). The claims most often used to motivate hybrid generative–Transformer designs are supported only by moderate or limited evidence, and several (diffusion augmentation, federated learning and clinical feasibility) have not yet been tested directly.

Table 8. Ranking of Evidence Strength and Unresolved Gaps

Rank	Finding	Key evidence	Strength	Unresolved gap	Priority
1	Deep models trained on large T1w cohorts reach MAE of about 2–4 years	[4–6, 31, 35]	Strong	Accuracy on small or clinical cohorts	Medium
2	Brain-age gap requires age-bias (regression-to-mean) correction	[2, 17, 25]	Strong	Correction not reported consistently across studies	High
3	MAE and R^2 depend on test-set age range; cross-study comparison is unreliable	[25, 52]	Moderate	No standard benchmark or fixed splits	Critical
4	Local–global (CNN + attention) hybrids give the lowest internal MAEs	[13, 37–39]	Moderate	Head-to-head evidence from one study only [13]; none on external multi-site data	High
5	Performance drops on external cohorts	[37, 39]	Moderate	External validation reported by few studies	Critical
6	Image- or feature-level harmonization improves downstream accuracy	[9–11]	Limited	Harmonization decoupled from the age objective; hallucination risk [40]	Critical
7	Adding non-imaging covariates or modalities improves accuracy	[33, 52, 54]	Limited	Gains of 0.16–0.54 years from single studies; deep multimodal fusion not yet compared with T1w-only baselines	Medium
8	Interpretability maps align with known ageing regions	[28, 39, 53]	Limited	Stability across subjects and sites untested	High
9	Transformers need very large datasets to outperform CNNs	[12, 13]	Limited	No brain-age data-scaling study	Medium

Rank	Finding	Key evidence	Strength	Unresolved gap	Priority
10	Diffusion-based augmentation corrects age imbalance	[36, 49]	Insufficient	Never tested on downstream brain-age bias; memorization risk [46]	High
11	Federated learning preserves centralized accuracy; models are clinically deployable	–	Insufficient	No brain-age federated or prospective clinical studies identified	Medium

HYBRID GENERATIVE–TRANSFORMER ARCHITECTURES: A CRITICAL SYNTHESIS

Recent hybrid architectures combine harmonization or bias control with Transformer-based prediction, moving beyond the decoupled "preprocessing-then-prediction" design by integrating harmonization, augmentation and global context modelling in one pipeline, see Table 9.

PSET-DABC [37] uses 3D Inception blocks for local feature extraction and a pyramid squeeze-and-excitation Transformer for global dependencies. Self-supervised pre-training and meta-learning provide dataset-adaptive generalization, and a dataset-adaptive and bias-constrained (DABC) regularization term penalizes dataset-specific bias during training. On 15,437 T1-weighted scans from 18 datasets (age 5–97 years) it reported an MAE of 2.34 years, rising to 3.86 years on external datasets.

BrainRotViT [39] pre-trains a ViT encoder on auxiliary age and sex classification. The frozen encoder processes sagittal slices into embedding matrices that a residual CNN regressor maps to age, with subject sex injected at the final fully connected layer. Separating feature extraction from regression lets the Transformer learn robust global representations while the CNN refines local detail. It reported an MAE of 3.34 years ($r = 0.98$, $R^2 = 0.95$) across 11 datasets from more than 130 sites and 3.77–5.04 years on four independent cohorts.

TractGraphFormer [53] combines a graph CNN stream that encodes local anatomical properties of white-matter fibre tracts with a Transformer stream that learns global information about the structural connectome through self-attention. An attention module identifies the most predictive white-matter connections, linking prediction accuracy with neurobiological interpretability. It was evaluated on children ($n = 9,345$) and young adults ($n = 1,065$).

NeuroAgeFusionNet [38] fuses CNN, Transformer and GNN representations and adds uncertainty quantification, representing local morphology, global anatomical dependencies and topological connectivity simultaneously. It reported an MAE of 2.30 years ($R^2 = 0.96$) on the UK Biobank, with per-prediction uncertainty estimates.

Table 9. Architectural and Performance Characteristics of Published Hybrid Architectures

Method	Architecture components	Training objective	Data	MAE (years)	Key contribution
PSET-DABC [37]	3D Inception + pyramid squeeze-and-excitation Transformer	Age loss + DABC bias regularization; self-supervised pre-training; meta-learning	15,437 scans, 18 datasets	2.34 (int.); 3.86 (ext.)	Dataset-adaptive, bias-constrained training
BrainRotViT [39]	ViT encoder + residual CNN	Auxiliary age/sex pre-training; regression with sex input	11 datasets, >130 sites	3.34 (int.); 3.77–5.04 (ext.)	Attention-based interpretability
TractGraphFormer [53]	Graph CNN + Transformer + attention module	Age and sex prediction from tractography	9,345 children; 1,065 young adults	–	Fibre-tract interpretability
NeuroAgeFusionNet [38]	CNN + Transformer + GNN	Age loss with uncertainty estimation; feature fusion	UK Biobank	2.30	Uncertainty quantification; multi-stream fusion

Four common design principles emerge. First, all successful hybrids separate local feature extraction from global context modelling and fuse them at an intermediate representation level. Second, bias-constrained or harmonization-aware loss terms are associated with better cross-site generalization. Third, explicit interpretability mechanisms (attention maps or gated modules) improve clinical acceptability. Fourth, auxiliary tasks (sex prediction) and self-supervised pre-training improve representation quality and, consequently, age regression. Table 10 links these principles to their technical formulation and to the gap they address.

Table 10. SOTA-Informed Design Principles for Hybrid Brain-Age Architectures

Design principle	Implementation strategy	Technical formulation	Evidence from SOTA	Addressed gap
Self-supervised pre-training	Feature learning from unlabelled MRI	Contrastive or masked-autoencoder loss	PSET-DABC [37]	Data scarcity
Meta-learning adaptation	Few-shot adaptation to new datasets	MAML-type episodic training	PSET-DABC [37]	Multi-site generalization
Bias-constrained training	Penalize dataset-specific bias	Bias regularization term (Eq. 13)	PSET-DABC [37]	Scanner-domain shift; age bias
Auxiliary task grounding	Age/sex pre-training	Multi-task loss $\mathscr{L}_{\text{age}} + \mathscr{L}_{\text{aux}}$	BrainRotViT [39]	Representation quality
Feature fusion	CNN + Transformer (+ GNN)	Cross-attention (Eqs. 7–8)	NeuroAgeFusionNet [38]; GLT [13]	Local–global trade-off
Anatomically guided attention	Attention over tracts or regions	Gate $g = \sigma(\mathbf{W}_g[\cdot] + b)$ (Eq. 9)	TractGraphFormer [53]	Interpretability
Uncertainty quantification	Confidence for each prediction	Gaussian negative log-likelihood (Eq. 11)	NeuroAgeFusionNet [38]	Clinical trust

A critical reading of the literature also shows persistent shortcomings. Joint CNN–Transformer training is computationally demanding, and none of the four hybrids reports a complete accounting of training and inference cost. Generative harmonization modules such as DDPMs require many iterative denoising steps, which hinders real-time deployment. Finally, differences in datasets, preprocessing and cross-validation make results difficult to compare across studies. Future hybrid architectures should therefore integrate these evidence-supported principles within a coherent design strategy rather than combining individual components without a systematic rationale. This provides the basis for the framework presented in the following section.

HAGT: AN EVIDENCE-DERIVED HARMONIZATION-AWARE GENERATIVE–TRANSFORMER DESIGN FRAMEWORK

The synthesis above shows that no published model simultaneously (i) couples harmonization to the age objective, (ii) fuses local and global representations, (iii) constrains age bias and (iv) quantifies uncertainty. We therefore formalize HAGT (Harmonization-Aware Generative–Transformer) as a design framework derived from the evidence in Tables 8 and 10. HAGT is not a new trained model and this review reports no new experiments; it is a specification whose components are each traced to supporting evidence (Table 11) and whose predictions are testable through hypotheses H1–H6.

Formal Definition

Let x be a T1-weighted volume acquired at site $s \in \{1, \ldots, S\}$ from a subject with chronological age y. HAGT is the composition of four modules. M1 – task-coupled harmonizer. A generator G_φ maps the input to a site-normalized volume:

$$\tilde{x} = G_\varphi(x, s) \tag{6}$$

M2 – local encoder. A 3D CNN converts $\tilde{x}$ into T_L local tokens F_L. M3 – global encoder. A (windowed) Transformer converts non-overlapping patches of size p^3 into T global tokens F_G using scaled dot-product attention [55]:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{7}$$

M4 – gated cross-attention fusion and regression head. Local tokens query the global context, and a learned gate decides, per token, how much global information to admit [13, 53]:

$$F_{L\to G} = \mathrm{Attn}\left(F_L W_Q, F_G W_K, F_G W_V\right) \tag{8}$$

$$g = \sigma\left(W_g[F_L; F_{L\to G}] + b\right), \quad F = g \odot F_L + (1 - g) \odot F_{L\to G} \tag{9}$$

A pooled representation of **F** is passed, with optional covariates (e.g., sex), to a head that outputs a predicted mean age μ_i and variance σ_i^2. The whole system is trained end to end with

$$\mathcal{L} = \mathcal{L}_{\text{age}} + \lambda_1 \mathcal{L}_{\text{cyc}} + \lambda_2 \mathcal{L}_{\text{site}} + \lambda_3 \mathcal{L}_{\text{bias}} + \lambda_4 \mathcal{L}_{\text{aux}} \tag{10}$$

$$\mathcal{L}_{\text{age}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(y_i-\mu_i)^2}{2\sigma_i^2} + \frac{1}{2}\log\sigma_i^2\right] \quad (11)$$

$$\mathcal{L}_{\text{cyc}} = \left\|G_{t\to s}(G_{s\to t}(x)) - x\right\|_1 \quad (12)$$

$$\mathcal{L}_{\text{bias}} = |\rho(\mu - y, y)| \quad (13)$$

$$\mathcal{L}_{\text{site}} = -\frac{1}{N}\sum_{i=1}^{N}\log D_\psi(s_i|F_i) \quad (14)$$

Equation 11 is the heteroscedastic Gaussian negative log-likelihood, which yields per-subject uncertainty [38, 56]. Equation 12 is the cycle-consistency term that keeps the harmonizer anatomy-preserving. Equation 13 penalizes the correlation between the brain-age gap and chronological age, i.e. the regression-to-the-mean bias that otherwise has to be removed post hoc [2, 17, 37]. Equation 14 is the cross-entropy of a site discriminator D_ψ; the encoders are trained to maximize it through gradient reversal, so the fused features become site-invariant, as in adversarial scanner unlearning [9]. $\mathscr{L}_{\text{aux}}$ is an auxiliary sex-classification loss [39].

Design Rules Derived from the Evidence

- R1 – Couple harmonization to the task. G_φ receives gradients from $\mathscr{L}_{\text{age}}$ and $\mathscr{L}_{\text{site}}$, not only from $\mathscr{L}_{\text{cyc}}$; this addresses the decoupling gap and constrains hallucination [10, 40].
- R2 – Make harmonization removable at inference. Because site invariance is enforced in feature space (Eq. 14), G_φ can be dropped or replaced by a single forward pass at inference, avoiding the τ-step cost of diffusion harmonizers (Table 12).
- R3 – Bound the token budget. Choose p and the attention window so that the attention cost (Eq. 15) fits clinical hardware.
- R4 – Correct and report bias. Train with Eq. 13 and still report post-hoc corrected gaps (Eq. 5) and age-stratified MAE.
- R5 – Validate externally. Report internal and external MAE on at least one cohort unseen during training, together with cohort, preprocessing, bias-correction, code and computational-cost details.

Table 11. Evidence-to-Component Mapping of the HAGT Framework

HAGT component	Design decision	Supporting evidence	Evidence strength (Table 8)	Testable prediction
M1 harmonizer	Task-coupled, train-time harmonization with cycle and site losses	Downstream MAE gain 0.29 years [10]; unlearning [9]; hallucination risk [40]	Limited (rank 6)	Smaller internal-to-external MAE gap than decoupled harmonization (H4)
M2 local encoder	Lightweight 3D CNN	Strong CNN baselines [5, 31]	Strong (rank 1)	Retains CNN accuracy on small cohorts

HAGT component	Design decision	Supporting evidence	Evidence strength (Table 8)	Testable prediction
M3 global encoder	Windowed Transformer on 3D patches	Global-local gains [13]; windowed attention [42]	Moderate (rank 4); limited for data size (rank 9)	Hybrid beats CNN on identical splits (H1); threshold lower than 10,000 scans (H3)
M4 fusion	Gated cross-attention (Eqs. 8–9)	[13, 38, 53]	Moderate (rank 4)	Gate values indicate regional reliance on global context
Head	Heteroscedastic, bias-constrained (Eqs. 11, 13)	Bias correction [2, 17]; uncertainty [38]	Strong (rank 2)	Lower age-stratified bias at age extremes
Auxiliary inputs	Sex classification / covariate input	Covariate gain 0.16 years [33]; multimodal gain 0.54 years [54]; [39]	Limited (rank 7)	Small but consistent gain (H5)
Optional augmentation	Age-conditioned latent diffusion for sparse ages	[36, 49]	Insufficient (rank 10)	Reduced regression-to-mean bias (H2)

Computational Complexity of HAGT

For an input of $H \times W \times D$ voxels and patch size p, the global encoder operates on $T = HWD/p^3$ tokens. Each of its L layers costs $O(T^2d)$ for the attention matrix plus $O(Td^2)$ for the projections and feed-forward layers [55]:

$$C_{\text{attn}} = O\left(L\left(T^2 d + Td^2\right)\right), \quad T = \frac{HWD}{p^3} \tag{15}$$

The total inference cost of HAGT is the sum of its modules:

$$C_{\text{HAGT}}^{\text{inf}} = C_G + O\left(\sum_l V_l k^3 C_{l-1} C_l\right) + C_{\text{attn}} + O(T_L T d) \tag{16}$$

where the second term is the 3D CNN cost (V_l voxels, kernel size k and C_l channels in layer l) and the last term is cross-attention fusion. The harmonizer cost C_G is zero if harmonization is used only during training (rule R2), one generator pass for a GAN, and τ denoiser passes for a diffusion model, where τ is typically up to 1,000 steps for a standard DDPM [45]. Training additionally requires the site discriminator, the reverse generator for Eq. 12 and backpropagation through all modules.

Worked example. For a 160 × 192 × 160 volume, a patch size of 16 gives $T = 10 \times 12 \times 10 = 1{,}200$ tokens and an attention matrix of 1.44×10^6 entries per head and layer (about 5.8 MB in 32-bit precision). Halving the patch size to 8 gives $T = 9{,}600$ tokens and 9.22×10^7 entries (about 369 MB) — a 64-fold increase. Restricting attention to 7 × 7 × 7 windows (343 tokens) [42] reduces this to $9{,}600 \times 343 \approx 3.3 \times 10^6$ entries, a 28-fold saving that makes fine patches feasible on clinical GPUs. Similarly, replacing a 1,000-step diffusion harmonizer by

train-time-only harmonization removes about 1,000 generator evaluations per scan at inference. These figures explain why computational cost, rather than accuracy, is likely to decide which hybrid designs reach clinical deployment. Table 12 summarizes the cost of each module.

Table 12. Computational Complexity of HAGT Modules and Mitigation Strategies

Module	Time complexity	Memory	Dominant factor	Mitigation
3D CNN encoder (M2)	$O(\Sigma\, V_l k^3 C_{l-1} C_l)$	$O(\Sigma\, V_l C_l)$	Input resolution, channels	Lightweight design [5, 34]
Global self-attention (M3)	$O(L(T^2 d + T d^2))$	$O(T^2)$ per head	Token count $T = HWD/p^3$	Larger patches; windowed attention
Windowed attention (M3)	$O(L(TM^3 d + T d^2))$	$O(TM^3)$	Window size M	Shifted windows [42]
Cross-attention fusion (M4)	$O(T_L T d)$	$O(T_L T)$	Number of local tokens	Pool local tokens before fusion
GNN branch (optional)	$O(\|E\| d + \|V\| d^2)$	$O(\|E\| + \|V\| d)$	Graph size	Parcellation-level graphs
GAN harmonizer (M1)	One generator pass	Generator activations	Generator depth	Train-time-only use (R2)
Diffusion harmonizer (M1)	τ denoiser passes	Denoiser activations	Steps τ (up to 1,000)	Latent diffusion [36]; train-time only
Site discriminator	$O(Td)$, training only	Small	–	Not used at inference

Compared with the authors' earlier NeuroAgeFusionNet [38], HAGT adds task-coupled harmonization (M1), an explicit site-adversarial objective and a bias constraint, and it specifies how to keep inference cost bounded; compared with PSET-DABC [37] and BrainRotViT [39], it adds image-level harmonization and uncertainty-aware regression in a single objective.

RESEARCH CHALLENGES AND OPEN ISSUES

Despite major advances in CNNs, Transformers, generative models and hybrid architectures, brain-age estimation is still limited in reliability, generalizability and clinical deployability. Building reliable, interpretable and clinically useful systems requires tackling scanner heterogeneity, data imbalance, interpretability gaps and computational cost.

Multi-Site Variability and Scanner-Domain Shift

Differences in scanner manufacturer, field strength, acquisition protocol and preprocessing pipeline change MRI images systematically. These variations skew voxel-intensity distributions and can lead deep models, particularly Transformers, to learn scanner signatures instead of biological ageing patterns [9, 11]. Performance drops by 0.4–1.7 years when hybrids are applied to unseen cohorts (see Descriptive Synthesis of

Reported Performance), and no harmonization framework (ComBat, CycleGAN or diffusion-based) has been adopted as a standard.

Inadequate Harmonization and Domain-Adaptation Pipelines

Although GAN- and diffusion-based harmonization can improve cross-site robustness, these techniques are used inconsistently and lack standardized evaluation. Many studies apply minimal preprocessing, which can inflate MAE on different scanners; GAN-based translation across sites can introduce hallucinated anatomical patterns [40]; and very few studies systematically compare harmonized and non-harmonized performance on the same data [10]. Conventional statistical approaches such as ComBat correct only distribution-level differences in derived features and do not operate on high-dimensional voxel data [7].

Limited Lifespan Coverage and Weak Augmentation

Neuroimaging datasets are typically skewed towards middle-aged adults, with few very young or very old participants. This imbalance strengthens regression to the mean at the age extremes [17]; traditional flips and rotations add little anatomical diversity; and generative augmentation, although more realistic, is complex and unstable to train and has not yet been shown to reduce brain-age bias [36, 49].

Trade-Off Between Local and Global Feature Representation

CNNs capture local morphology well but cannot model long-range cortical–subcortical interactions, whereas Transformers capture global dependencies but need large, balanced data to avoid overfitting [12]. Combining both pathways increases the complexity of fusion, normalization and interpretation, and some hybrids yield uneven results because of differences in implementation quality, patching schemes and training data.

Computational Cost and Resource Efficiency

Whole-brain 3D CNNs, Swin and pyramid Transformers and diffusion models require substantial GPU memory and training time, which exceeds the hardware available in many clinical centres. Diffusion inference requires many denoising steps [45], and few architectures are designed to run on hospital PACS or low-resource systems. The complexity analysis of HAGT (Eqs. 15–16, Table 12) quantifies these costs and shows how they can be bounded.

Explainability and Clinical Trustworthiness

Clinicians need to understand why a model judges a brain to be older or younger than expected. CNN saliency maps are approximate, Transformer attention matrices are dense and difficult to read, and gradient-based explanations such as DeepSHAP vary more than SHAP explanations of morphometric models [28]. Few studies correlate model-derived regions with established ageing biomarkers such as hippocampal atrophy and ventricular expansion, and non-explainable systems are difficult to qualify for clinical approval.

Dataset Bias, Limited Diversity and Lack of Standard Benchmarks

Brain-age datasets often have limited diversity in ethnicity, sex and comorbidity, which can reduce fairness and robustness. Studies use different preprocessing pipelines (FreeSurfer, FSL and ANTs), many state-of-the-art models are not publicly available, and inconsistent use of MAE, bias correction and calibration makes results difficult to compare [25].

Real-World Deployment Barriers

Even the strongest research models rarely transition into clinical workflows. Integration with PACS/RIS requires strict latency, robustness and compatibility; current architectures are too heavy for on-scanner inference; and lifespan brain-age biomarkers risk misuse without clear ethical and privacy guidelines.

FUTURE RESEARCH DIRECTIONS

Brain-age prediction is maturing into a neuroimaging biomarker whose clinical and translational potential has not yet been realized. Based on the limitations identified in the preceding sections, Figure 2 maps each major gap to the current SOTA response, the reason this response is insufficient and the breakthrough required, together with an indicative time horizon.

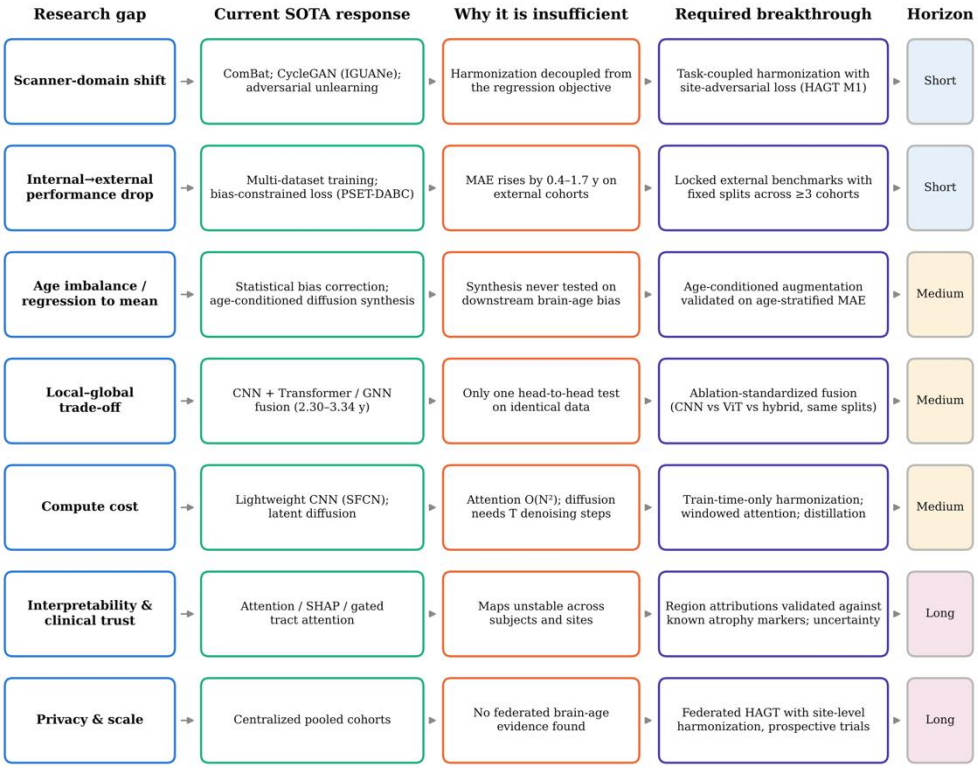


Figure 2. Gap-to-Solution Roadmap. Each Row Links a Research Gap to the Current State-of-the-Art Response, the Evidence of Why It Is Insufficient (Tables 7 and 8) and the Breakthrough Required. Horizon: Short (1–2 Years), Medium (2–4 Years), Long (>4 Years).

Harmonization-Aware Learning Pipelines

CycleGAN, ComBat and diffusion-based harmonization improve cross-site robustness but are applied as external preprocessing. Next-generation systems should be trained end to end so that scanner-invariant features, domain adaptation and age prediction are learned jointly (HAGT rule R1). Such pipelines would reduce domain shift and artefacts and support deployment across heterogeneous clinical scanners.

Lifespan-Balanced Generative Modelling

Because most MRI samples come from middle-aged adults, predictions at the extremes of the age range are less reliable. Age-conditioned diffusion or GAN-based synthesis should be used to balance lifespan distributions and increase morphological variability, and must be validated by its effect on age-stratified MAE and regression-to-the-mean bias, while guarding against training-data memorization [46]. This is most valuable for paediatric, adolescent and geriatric populations, where brain-age deviations are most clinically informative.

Lightweight and Deployable Hybrid Architectures

Future hybrids should preserve local–global integration while reducing memory and compute. Options include windowed or low-rank attention, MobileViT-style 3D variants and distillation-based compression, guided by the complexity analysis of HAGT (Table 12). Such designs would make on-scanner or near-real-time inference possible.

Interpretability and Biological Alignment

Biologically grounded explanations are needed, for example attention maps validated against cortical-thinning curves, white-matter degeneration and ventricular enlargement. Quantitative region attribution, counterfactual generative analysis and saliency stabilization can improve trustworthiness and regulatory readiness.

Multi-Modal Integration and Trajectory Modelling

Most models rely on T1-weighted MRI alone and ignore the complementary information in diffusion MRI, resting-state fMRI, perfusion imaging and genetics [53, 54]. Multi-modal fusion Transformers and cross-modal generative alignment could yield more holistic ageing biomarkers. In addition, existing methods produce a single estimate at one time point; temporal Transformers, longitudinal generative models and sequence-based hybrids could instead model individual ageing trajectories and the onset of neurodegeneration [50].

Graph Neural Networks for Structural and Functional Topology

The brain is naturally represented as a graph. Applying GNNs to cortical-thickness networks, structural connectomes or fMRI connectivity enables topology-aware brain-age estimation that aligns predictions with known patterns of network-level degeneration [53].

Federated and Continual Learning

Future brain-age systems will need to learn across geographically distributed sites without a central data repository. Federated learning can adapt models to new demographics and scanners while preserving privacy, and continual learning can maintain stability in evolving longitudinal cohorts with minimal catastrophic forgetting. No federated brain-age study was identified in this review, so H6 remains untested.

Automated Quality Control and Sustainable Pipelines

Automated quality control, anomaly detection and continuous performance monitoring could detect scanner drift and trigger recalibration, leading to self-monitoring brain-age systems. Sustainability-focused training strategies can reduce compute use and carbon footprint during large-scale training.

Quantum-Inspired Learning

Quantum-inspired feature encodings and hybrid quantum–classical models have been proposed as a way to represent high-dimensional morphology [51]. Because no MRI-based brain-age application exists yet, this remains a long-term, exploratory direction.

DISCUSSION

Synthesis of Findings

This review synthesized 52 studies on brain-age estimation with deep-learning, generative and Transformer approaches. Because designs were heterogeneous, the evidence was summarized descriptively and no pooled statistical test was performed. Models that fuse convolutional and attention-based pathways report the lowest internal errors (median MAE 2.70 years, range 2.30–3.34) compared with CNN-family models (median 3.17 years, range 2.14–4.44), a difference of about 15% that is confounded by dataset and publication year. Within-study comparisons, which hold the data constant, show smaller but more reliable effects: adding multimodal imaging features reduced MAE by 0.54 years [54], integrating sex information by 0.16 years [33] and CycleGAN harmonization by 0.29 years [10]. In contrast, external validation increased MAE by 0.43–1.70 years in the hybrids that reported it [37, 39]. The dominant source of error in current systems is therefore the shift between training and deployment populations, not the choice between CNN and Transformer backbones.

Hypothesis Verification

Table 13 classifies each hypothesis using the decision rules defined in the Review Methodology and the evidence grades of Table 8. Overall, none of the six hypotheses is fully supported at its stated threshold. H1 and H5 are partially supported, H3 is not supported as a strict threshold, and H2, H4 and H6 cannot be evaluated with the published evidence. These hypotheses should therefore be regarded as research questions for future

controlled studies rather than as established conclusions; Table 11 shows how each can be tested within the HAGT framework.

Table 13. Verification of Hypotheses H1–H6 Against the Reviewed Evidence

Hypothesis	Evidence from the review	Assessment	Decision
H1: hybrid CNN–Transformer reduces MAE by ≥20% vs CNN on multi-site data	Median internal MAE 2.70 vs 3.17 years (about 15% lower) across different datasets (see Descriptive Synthesis of Reported Performance); head-to-head on identical data: 2.70 vs 2.97–3.25 years (9–17% lower) [13]	Direction consistent; reductions of 9–17% fall short of the 20% threshold; evidence moderate	Partially Supported
H2: DDPM augmentation reduces regression-to-mean bias by >15% vs GAN augmentation	Diffusion models generate realistic, age-conditioned images [36, 49], but no study measured downstream bias against GAN augmentation	No direct test	Insufficient Evidence
H3: Transformers need ≥10,000 training scans to outperform CNNs	A Global-Local Transformer outperformed the CNN baselines evaluated in the same study with 8,379 scans [13]; Transformers are data-hungry in general vision [12]	Evidence contradicts a strict 10,000-scan threshold, but no data-scaling study exists	Not Supported (limited evidence)
H4: diffusion-based harmonization improves cross-site MAE by ≥0.5 years	CycleGAN harmonization improved downstream MAE by 0.29 years [10]; no diffusion-based harmonization study reported brain-age MAE	Threshold not met by GAN evidence; diffusion untested	Insufficient Evidence
H5: multi-modal integration improves MAE by ≥0.3 years over T1w	Adding resting-state functional connectivity features to structural (T1w-derived) features reduced MAE from 4.83 to 4.29 years (0.54 years) [54]; adding sex information gave 0.16 years [33]	Threshold met in one study with a matched baseline; not yet replicated (limited evidence)	Partially Supported
H6: federated learning retains ≥95% of centralized performance	No federated brain-age study identified	Untested	Insufficient Evidence

Why the Current State of the Art Remains Insufficient

Five observations explain why recent gains have not yet translated into clinically usable brain-age biomarkers.

- In-distribution accuracy is saturating, generalization is not. Large-cohort models already reach internal MAEs of about 2.1–2.7 years [5, 37, 38], whereas external

errors are 0.4–1.7 years higher [37, 39]. Further architectural refinements that lower internal MAE by a few tenths of a year do not address the larger external error.

- A lower MAE is not automatically a better biomarker. Bashyam et al. reported that moderately fitted brain-age models discriminated disease groups significantly better than tightly fitted models in two of the four disease groups tested [6], and MAE depends on the age range of the test sample [25]. Optimizing MAE alone can therefore remove the very deviations that make brain age clinically informative.
- Harmonization is decoupled from the prediction objective. Harmonizers are trained to match image distributions, not to preserve age information, and can hallucinate or remove features [40]; the only quantified downstream gain is modest (0.29 years) [10].
- Bias, uncertainty and cost are rarely reported. Without age-bias correction, uncertainty estimates and computational cost, clinicians cannot judge whether an individual brain-age gap is meaningful or whether a model can run in their setting.
- Evidence for generative augmentation, diffusion-based harmonization and federated learning is missing. Three of the six hypotheses could not be evaluated at all, which shows that the field's most-cited future directions are still largely untested.

Accordingly, the following concrete breakthroughs are required: (i) a shared benchmark with fixed subject-level splits across at least three cohorts that include the age extremes, with mandatory external testing; (ii) task-coupled harmonization that is shown to reduce the internal-to-external MAE gap, for example to below 0.5 years; (iii) evaluation of brain-age gaps by their effect size in clinical groups, not only by MAE in healthy controls; (iv) calibrated per-subject uncertainty; (v) hybrid architectures with bounded inference cost (Table 12); and (vi) prospective, longitudinal clinical validation.

Critical Comparison with the State of the Art

NeuroAgeFusionNet (MAE 2.30 years) and PSET-DABC (MAE 2.34 years) report the lowest internal errors among the hybrids, both combining convolutional and attention-based components; PSET-DABC also reports external performance. BrainRotViT (MAE 3.34 years) shows that interpretability can be obtained with a moderate accuracy cost, and TractGraphFormer provides anatomically informed prediction but is currently limited to diffusion MRI. Hybrids work because CNN local feature extraction, Transformer global context modelling and GNN topological reasoning capture complementary aspects of brain-ageing morphology, while harmonization or bias-constrained training reduces scanner-specific artefacts and thereby strengthens the biological ageing signal.

Limitations of This Review

This review has several limitations. First, heterogeneity in cohorts, age ranges, preprocessing and validation precluded meta-analysis, so the comparisons are descriptive. Second, performance values were taken from the original publications and may be affected by reporting and publication bias. Third, a formal study-level risk-of-bias appraisal was not performed, so the evidence grades rely on the consistency and validation design of the

reported results. Fourth, most included studies use T1-weighted MRI only, so conclusions about other modalities are tentative. Fifth, quantum-inspired models have not yet been applied to MRI. Finally, HAGT is an evidence-derived design framework that has not been empirically validated in this review; its predictions are stated as testable hypotheses.

CONCLUSIONS

Brain-age estimation has progressed from kernel regression on morphometric features to hybrid architectures that combine convolutional, attention-based, graph-based and generative components. This evidence-graded review of 52 studies shows that the progress is real but uneven. Only two findings are supported by strong evidence: deep models trained on large T1-weighted cohorts achieve MAEs of about 2–4 years, and brain-age gaps require bias correction. Hybrid local–global models report the lowest internal errors (median 2.70 versus 3.17 years for CNN-family models), but this advantage is smaller than the 0.4–1.7-year increase in error observed on external cohorts. The central problem of the field has therefore shifted from accuracy to generalization, bias and clinical validity.

None of the six hypotheses was fully supported, and three could not be tested, which shows that much of the field's agenda rests on untested assumptions. To close this gap, we proposed HAGT, an evidence-derived framework in which harmonization is coupled to the age objective, local and global representations are fused through gated cross-attention, age bias is constrained during training and uncertainty is estimated for each prediction, with an explicit complexity analysis showing how inference cost can be bounded. Progress now depends less on new backbones than on shared external benchmarks, task-coupled harmonization, disease-relevant evaluation, calibrated uncertainty and prospective clinical validation. Meeting these requirements is the step that would allow brain age to move from a research metric to a clinically usable biomarker for neurodegenerative disease and population health.

AUTHOR CONTRIBUTIONS

Conceptualization, M.S., and S.R.D.; Methodology, M.S., and S.R.D.; Validation, S.R.D.; Investigation, M.S.; Resources, M.S.; Data Curation, M.S., and S.R.D.; Writing – Original Draft Preparation, M.S.; Writing – Review & Editing, S.R.D.; Visualization, M.S., and S.R.D.; Supervision, S.R.D.; Project Administration, S.R.D.

DATA AVAILABILITY

The data generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

CONFLICT OF INTERESTS

The authors declare that they have no conflict of interest. Sowmya and Dutta are the authors of NeuroAgeFusionNet [38], which is discussed in this review and was assessed with the same criteria as all other included studies.

REFERENCES

1. Cole, J.H., and Franke, K. Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers, *Trends in Neurosciences*, **2017**, *40*(12), 681–690.
2. Smith, S.M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T.E., and Miller, K.L. Estimation of Brain Age Delta from Brain Imaging, *NeuroImage*, **2019**, 200, pp. 528–539.
3. Baecker, L., Garcia-Dias, R., Vieira, S., Scarpazza, C., and Mechelli, A. Machine Learning for Brain Age Prediction: Introduction to Methods and Clinical Applications, *EBioMedicine*, **2021**, *72*, 103600.
4. Cole, J.H., Poudel, R.P.K., Tsagkrasoulis, D., Caan, M.W.A., Steves, C., Spector, T.D., and Montana, G. Predicting Brain Age with Deep Learning from Raw Imaging Data Results in a Reliable and Heritable Biomarker, *NeuroImage*, **2017**, 163, 115–124.
5. Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., and Smith, S.M. Accurate Brain Age Prediction with Lightweight Deep Neural Networks, *Medical Image Analysis*, **2021**, *68*, 101871.
6. Bashyam, V.M., Erus, G., Doshi, J, et al. MRI Signatures of Brain Age and Disease over the Lifespan Based on a Deep Brain Network and 14,468 Individuals Worldwide, *Brain*, **2020**, *143*(7), 2312–2324.
7. Fortin, J.P., Cullen, N., Sheline, Y.I., et al. Harmonization of Cortical Thickness Measurements Across Scanners and Sites, *NeuroImage*, **2018**, *167*, 104–120.
8. Dinsdale, N.K., Jenkinson, M., and Namburete, A.I.L. Unlearning Scanner Bias for MRI Harmonisation in Medical Image Segmentation, In: *Medical Image Understanding and Analysis (MIUA 2020)*, Communications in Computer and Information Science, vol. 1248, Springer, **2020**, pp. 15–25.
9. Dinsdale, N.K., Jenkinson, M., and Namburete, A.I.L. Deep Learning-Based Unlearning of Dataset Bias for MRI Harmonisation and Confound Removal, *NeuroImage*, **2021**, *228*, 117689.
10. Roca, V., Kuchcinski, G., Pruvo, J.P., Manouvriez, D., and Lopes, R. IGUANe: A 3D Generalizable CycleGAN for Multicenter Harmonization of Brain MR Images, *Medical Image Analysis*, **2025**, *99*, 103388.
11. Pomponio, R., Erus, G., Habes, M., et al. Harmonization of Large MRI Datasets for the Analysis of Brain Imaging Patterns Throughout the Lifespan, *NeuroImage*, **2020**, 208, 116450.
12. Spahiu, E., Xhako, D., Hyka, N., & Hoxhaj, S. 3D Magnetic Resonance Image Segmentation Using HD Brain Extraction in 3D Slicer. *Journal of Transactions in Systems Engineering*, **2025**, *3*(1), 340–348.
13. He, S., Grant, P.E., and Ou, Y. Global-Local Transformer for Brain Age Estimation, *IEEE Transactions on Medical Imaging*, **2022**, *41*(1), 213–224.
14. Tanveer, M., Ganaie, M.A., Beheshti, I., et al. Deep Learning for Brain Age Estimation: A Systematic Review, *Information Fusion*, **2023**, *96*, 130–143.
15. Wu, Y., Gao, H., Zhang, C., Ma, X., Zhu, X., Wu, S., and Lin, L. Machine Learning and Deep Learning Approaches in Lifespan Brain Age Prediction: A Comprehensive Review, *Tomography*, **2024**, *10*(8), 1238–1262.
16. Azzam, M., Xu, Z., Liu, R, et al. A Review of Artificial Intelligence-Based Brain Age Estimation and Its Applications for Related Diseases, *Briefings in Functional Genomics*, **2025**, *24*, elae042.

17. Liang, H., Zhang, F., and Niu, X. Investigating Systematic Bias in Brain Age Estimation with Application to Post-Traumatic Stress Disorders, *Human Brain Mapping*, **2019**, *40*(11), 3143–3152.
18. Page, M.J., McKenzie, J.E., Bossuyt, P.M., et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews, *BMJ*, **2021**, *372*, 71.
19. McHugh, M.L. Interrater Reliability: The Kappa Statistic, *Biochemia Medica*, **2012**, *22*(3), 276–282.
20. Fjell, A.M., and Walhovd, K.B. Structural Brain Changes in Aging: Courses, Causes and Cognitive Consequences, *Reviews in the Neurosciences*, **2010**, *21*(3), 187–221.
21. Franke, K., and Gaser, C. Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained?, *Frontiers in Neurology*, **2019**, *10*, 789.
22. Glasser, M.F., Smith, S.M., Marcus, D.S., et al. The Human Connectome Project's Neuroimaging Approach, *Nature Neuroscience*, **2016**, *19*(9), 1175–1187.
23. Tustison, N.J., Avants, B.B., Cook, P.A., et al. N4ITK: Improved N3 Bias Correction, *IEEE Transactions on Medical Imaging*, **2010**, *29*(6), 1310–1320.
24. Fortin, J.P., Parker, D., Tunç, B., et al. Harmonization of Multi-Site Diffusion Tensor Imaging Data, *NeuroImage*, **2017**, 161, 149–170.
25. de Lange, A.M.G., Anatürk, M., Rokicki, J., Han, L.K.M., Franke, K., Alnæs, D., Ebmeier, K.P., Draganski, B., Kaufmann, T., Westlye, L.T., Hahn, T., and Cole, J.H. Mind the Gap: Performance Metric Evaluation in Brain-Age Prediction, *Human Brain Mapping*, **2022**, *43*(10), 3113–3129.
26. Han, J., Kim, S. Y., Lee, J., Lee, W. H. Brain age prediction: A comparison between machine learning models using brain morphometric data. *Sensors* **2022**, *22*(20), 8077.
27. Franke, K., Ziegler, G., Klöppel, S., Gaser, C. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage* **2010**, *50* (3), 883–892.
28. De Bonis, M.L.N., Fasano, G., Lombardi, A., et al. Explainable brain age prediction: A comparative evaluation of morphometric and deep learning pipelines. *Brain Informatics* **2024**, *11*, 33.
29. Ballester, P.L., da Silva, L.T., Marcon, M., Esper, N.B., Frey, B.N., Buchweitz, A., Meneguzzi, F. Predicting brain age at slice level: Convolutional neural networks and consequences for interpretability. *Front. Psychiatry* **2021**, *12*, 598518.
30. Mahmoud, E., Elshennawy, N.M., Elkholy, A. Hybrid deep learning model for brain age prediction using time-distributed convolutional and bidirectional LSTM networks. *Sci. Rep.* **2026**, *16*, 18332.
31. Dartora, C., Marseglia, A., et al. A deep learning model for brain age prediction using minimally preprocessed T1w images as input. *Front. Aging Neurosci.* **2024**, *15*, 1303036.
32. Li, X., Hao, Z., Li, D., Jin, Q., Tang, Z., Yao, X., Wu, T. Brain age prediction via cross-stratified ensemble learning. *NeuroImage* **2024**, *299*, 120825.
33. Joo, Y., Namgung, E., Jeong, H., et al. Brain age prediction using combined deep convolutional neural network and multi-layer perceptron algorithms. *Sci. Rep.* **2023**, *13*, 22388.
34. Eltashani, F., Parreno-Centeno, M., Cole, J.H., Papa, J.P., Costen, F. Brain age prediction using a lightweight convolutional neural network. *IEEE Access* **2025**, *13*, 6750–6763.

35. Jonsson, B.A., Bjornsdottir, G., Thorgeirsson, T.E., et al. Brain age prediction using deep learning uncovers associated sequence variants. *Nat. Commun.* **2019**, *10*, 5409.
36. Pinaya, W.H.L., Tudosiu, P.D., Dafflon, J., et al. Brain imaging generation with latent diffusion models. In *Deep Generative Models (DGM4MICCAI 2022)*; Lecture Notes in Computer Science, Vol. 13609; Springer: **2022**; pp. 117–126.
37. Hu, Y., Wang, H., Cao, J., Li, B. Dataset-adaptive and bias-constrained brain age estimation using pyramid squeeze and excitation Transformer. *Neurocomputing* **2025**, *648*, 130705.
38. Sowmya, M., Dutta, S.R. NeuroAgeFusionNet: An ensemble deep learning framework integrating CNN, Transformers, and GNN for robust brain age estimation using MRI scans. *Sci. Rep.* **2025**, *15*, 41996.
39. Jalal, W., Rahman, M.N., Rahman, A.H., Rahman, M.S. BrainRotViT: Transformer-ResNet hybrid for explainable modeling of brain aging from 3D sMRI. *arXiv* **2025**, arXiv:2511.15188.
40. Cohen, J.P., Luck, M., Honari, S. Distribution matching losses can hallucinate features in medical image translation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI 2018)*; Lecture Notes in Computer Science, Vol. 11070; Springer: **2018**; pp. 529–536.
41. Cole, J.H., Ritchie, S.J., Bastin, M.E., et al. Brain age predicts mortality. *Mol. Psychiatry* **2018**, *23*(5), 1385–1392.
42. Liu, Z., Lin, Y., Cao, Y., et al. Swin Transformer: Hierarchical vision Transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021)*; **2021**; pp. 10012–10022.
43. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., Xu, D. Swin UNETR: Swin Transformers for semantic segmentation of brain tumors in MRI images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2021)*; Lecture Notes in Computer Science, Vol. 12962; Springer: **2022**; pp. 272–284.
44. Afridi, S., Jan, A., Irfan, M.A., Khattak, M.I., Khan, T. A. 3D-ViT-UNet: 3D vision Transformer based UNet-like model for volumetric brain tumor segmentation. *PLOS Digit. Health* **2026**, *5*(3), e0001323.
45. Ho, J., Jain, A., Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*; **2020**; pp. 6840–6851.
46. Akbar, M.U., Wang, W., Eklund, A. Beware of diffusion models for synthesizing medical images—A comparison with GANs in terms of memorizing brain MRI and chest X-ray images. *Mach. Learn.: Sci. Technol.* **2025**, *6*(1), 015022.
47. Dewey, B.E., Zhao, C., Carass, A., Oh, J., Calabresi, P.A., van Zijl, P.C.M., Prince, J.L. Deep harmonization of inconsistent MR data for consistent volume segmentation. In *Simulation and Synthesis in Medical Imaging (SASHIMI 2018)*; Lecture Notes in Computer Science, Vol. 11037; Springer: **2018**; pp. 20–30.
48. Zuo, L., Dewey, B.E., Liu, Y., He, Y., Newsome, S.D., Mowry, E.M., Resnick, S.M., Prince, J.L., Carass, A. Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. *NeuroImage* **2021**, *243*, 118569.
49. Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., et al. Denoising diffusion probabilistic models for 3D medical image generation. *Sci. Rep.* **2023**, *13*, 7303.

50. Zapaishchykova, A., Tak, D., Ye, Z., Liu, K.X., Likitlersuang, J., Vajapeyam, S., Chopra, R.B., Seidlitz, J., Bethlehem, R.A.I., Mak, R.H., Mueller, S., Haas-Kogan, D.A., Poussaint, T.Y., Aerts, H.J.W.L., Kann, B.H. Diffusion deep learning for brain age prediction and longitudinal tracking in children through adulthood. *Imaging Neurosci.* **2024**, *2*, 1–14.

51. Kölle, M., Stenzel, G., Stein, J., Zielinski, S., Ommer, B., Linnhoff-Popien, C. Quantum denoising diffusion models. In *2024 IEEE International Conference on Quantum Software (QSW)*; **2024**; pp. 88–98.

52. Wang, C., Zhang, W., Ni, M., Wang, Q., Liu, C., Dai, L., Zhang, M., Shen, Y., Gao, F. Deep-learning based multi-modal models for brain age, cognition and amyloid pathology prediction. *Alzheimer's Res. Ther.* **2025**, *17*(1), 126.

53. Chen, Y., Zhang, F., Wang, M., Zekelman, L.R., Cetin-Karayumak, S., Xue, T., Zhang, C., Song, Y., Rushmore, J., Makris, N., Rathi, Y., Cai, W., O'Donnell, L.J. TractGraphFormer: Anatomically informed hybrid graph CNN-Transformer network for interpretable sex and age prediction from diffusion MRI tractography. *Med. Image Anal.***2025**, *101*, 103476.

54. Liem, F., Varoquaux, G., Kynast, J., et al. Predicting brain-age from multimodal imaging data captures cognitive impairment. *NeuroImage* **2017**, *148*, 179–188.

55. Vaswani, A., Shazeer, N., Parmar, N., et al. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*; **2017**; pp. 5998–6008.

56. Kendall, A., Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*; **2017**; pp. 5574–5584.