# Modeling Brain Aging with Explainable Triamese ViT: Towards Deeper Insights into Autism Disorder

Zhaonian Zhang, Richard Jiang, Plamen Angelov and Vaneet Aggarwal

***Abstract*— Machine learning, particularly through advanced imaging techniques such as three-dimensional Magnetic Resonance Imaging (MRI), has significantly improved medical diagnostics. This is especially critical for diagnosing complex conditions like Alzheimer's disease. Our study introduces Triamese-ViT, an innovative Tri-structure of Vision Transformers (ViTs) that incorporates a built-in interpretability function, it has structure-aware explainability that allows for the identification and visualization of key features or regions contributing to the prediction, integrates information from three perspectives to enhance brain age estimation. This method not only increases accuracy but also improves interoperability with existing techniques. When evaluated, Triamese-ViT demonstrated superior performance and produced insightful attention maps. We applied these attention maps to the analysis of natural aging and the diagnosis of Autism Spectrum Disorder (ASD). The results aligned with those from occlusion analysis, identifying the Cingulum, Rolandic Operculum, Thalamus, and Vermis as important regions in normal aging, and highlighting the Thalamus and Caudate Nucleus as key regions for ASD diagnosis.**

***Index Terms*— Deep Learning, Natural Aging, ASD Analysis, Triamese-ViT.**

## I. Introduction

THE biological aging process is characterized by accumulating adverse changes, leading to progressive declines in physiological functions. Brain aging, in particular, is closely linked to diseases like Alzheimer's disease [1], psychosis [2], mild cognitive impairment [3], and depression [4]. Understanding brain aging is crucial for enhancing health.

Aging significantly reduces brain volume, notably in gray matter regions such as the prefrontal cortex, and insular cortex, critical for memory, planning, and decision-making [5]–[8]. White matter integrity, vital for neural connectivity, also deteriorates [9], potentially slowing processing speeds and impairing cognition. Additionally, aging increases ventricular and cerebrospinal fluid volumes [5], [6], [9], and in diseases like Alzheimer's, amyloid-beta plaques [10] and tau protein tangles [11] accumulate, leading to neuronal degeneration.

Recent deep learning advances have revolutionized brain diagnostics [12]–[18]. Deep learning-based brain age estimation from MRI (Figure 1) aids in identifying age-related diseases [1], [3], [12]. The 'brain age gap' (BAG), the difference between estimated and chronological brain age, serves as a valuable biomarker [12]. A younger-appearing brain typically indicates health, while an older-appearing brain may signal Alzheimer's disease [1], psychosis [2], mild cognitive impairment [3], or depression [4]. Therefore, refining brain age estimation algorithms is essential for aging analysis and early disease detection.

Current brain age estimation primarily utilizes convolutional neural networks (CNNs) trained on 3D MRI scans or 2D slices [12], [19], [20]. Although CNNs excel at detailed image processing by analyzing local pixel groups [21], they often neglect global structural information critical for comprehensive brain analysis. Furthermore, the opacity of CNNs poses difficulties for their integration into Explainable AI, hindering interpretability in medical diagnostics [22]. Conversely, Vision Transformers (ViTs) offer advantages by segmenting images into patches and employing attention mechanisms to capture complex inter-patch relationships [23], enabling detailed feature extraction and improved transparency through attention maps [24]. However, since ViTs are primarily tailored for 2D data [25]–[27], they may inadequately exploit the full depth and context provided by 3D MRI scans, potentially missing essential depth-related information in brain age estimation.

Explainability is crucial for brain age estimation, highlighting key regions for prediction and aiding in disease diagnosis. Model interpretability typically involves either post-hoc explanations or inherently interpretable models. Post-hoc methods provide explanations for black-box models through feature attribution, often using perturbations [28], [29] or gradients [30], [31]. However, perturbation-based techniques may yield unreliable explanations due to assumptions about feature independence [32]. Inherently interpretable models, like linear models, decision trees, GLMs, GAMs [33], JAMs [34], prototype-based models [35], and weight-aligned models [36], provide more transparent explanations but often compromise on predictive accuracy.

Our study introduces 'Triamese-ViT,' a deep-learning model designed to achieve high accuracy and interpretability in brain aging analysis. Trained on MRI data from 1,351 cognitively

This work was supported in part by the UK EPSRC under Grant EP/P009727/2, and the Leverhulme Trust under Grant RF-2019-492. (Correspondent author: Richard Jiang, e-mail: r.jiang2@lancaster.ac.uk).

Zhaonian Zhang, Plamen Angelov and Richard Jiang are with LIRA Center in Lancaster University,Lancaster, Lancashire, LA1 4WA, UK.

Vaneet Aggarwal is with Edwardson School of Industrial Engineering, Purdue University, West Lafayette IN 47907, USA.

Manuscript received Sept 26th, 2024; revised xxx xxx, 2025.

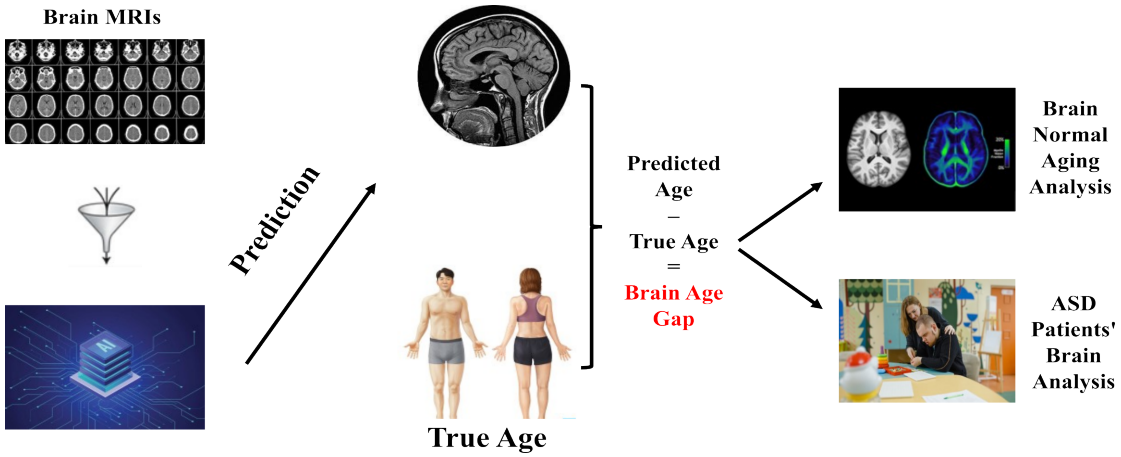


Fig. 1: It illustrates the brain age estimation process. MRIs serve as input to the deep learning models, which then predict the subjects' ages based on these images. These predicted ages are compared with the subjects' actual chronological ages to calculate key indicators, notably the brain age gap (predicted age minus chronological age). In this paper, we used this brain age gap to analyze brain normal aging and ASD patients.

healthy individuals (ages 6–80), Triamese-ViT uses Vision Transformers to capture distinct features from three different orientations (Figure 3). These features are integrated through a Tri Multi-Layer Perceptron to predict age. This novel Tri-Structure demonstrated superior predictive accuracy, fairness, and interpretability compared to existing state-of-the-art methods, achieving an MAE of 3.85, a Spearman correlation of 0.94 between predicted and chronological ages, and a -0.3 correlation between chronological age and brain age gap. Moreover, by combining multi-view data, the model generates 3D-like attention maps, enhancing its value for assessing normal brain aging and diagnosing brain diseases.

Compared to recent works such as [25], [37], our proposed Triamese-ViT achieves higher prediction accuracy while also demonstrating improved fairness in brain age estimation. Additionally, it incorporates a built-in interpretability mechanism, generating explainable result maps with clearer structural insights, which is not available in these previous approaches.Furthermore, in comparison to the high-accuracy 3D ViT model proposed in [38], our model offers substantial advantages in terms of computational efficiency, reduced memory usage, and simplified implementation. By leveraging a multi-view axis-wise processing strategy, Triamese-ViT significantly lowers the computational burden while maintaining strong predictive performance, making it more scalable for large-scale 3D medical imaging applications.

Furthermore, we leveraged the interpretability of Triamese-ViT to perform an analysis of natural brain aging, to track the significance trends of various brain regions over the course of aging. The interpretability of Triamese-ViT was also applied to the diagnosis of ASD, enabling the identification of key regions as perceived by the AI. To accomplish this, we employed 3D occlusion analysis—a conventional technique in Explainable AI (XAI)—to pinpoint regions of high correlation during prediction, thereby validating the alignment with the attention map generated by our model.

Through attention map analyses across different age cohorts, we investigated the changes in brain structure during aging from a machine-learning perspective, identifying regions that exhibit age-specific characteristics. Notable regions identified include the Rolandic Operculum, Cingulum, Thalamus, and Vermis, which are closely associated with various prevalent brain diseases. In the case of patients with ASD, the model highlighted the Thalamus and Caudate Nucleus, emphasizing their relevance in the disorder's pathology.

In this paper, our contributions include:

- We propose Triamese-ViT, a high-accuracy and fair model for brain age estimation that outperforms state-of-the-art (SOTA) models.
- Compared to high accuracy 3D ViT model, our model offers significant improvements in computational efficiency, reduced memory usage, and simplified implementation.
- Our model exhibits strong interpretability, producing attention maps with clearer structural insights compared to existing interpretable models. Additionally, its interpretability is validated through occlusion analysis.
- The interpretability of our model identifies crucial brain regions associated with normal aging, providing valuable insights into the aging process.
- Our model also highlights key brain regions relevant to ASD diagnosis, offering potential clinical significance for autism research.

All the above contributions are proved in the Results section.

## II. Method

### A. Data and Code Availability

We used MRI scans from the IXI[1] and ABIDE[2] datasets here. We collected a dataset of healthy participants to train the model and analyze normal brain aging, as well as a dataset of ASD patients to identify crucial brain regions for ASD detection. All the MRIs' types are T1-weighted. The dataset

[1]https://brain-development.org/ixi-dataset/
[2]https://fcon_1000.projects.nitrc.org/indi/abide/

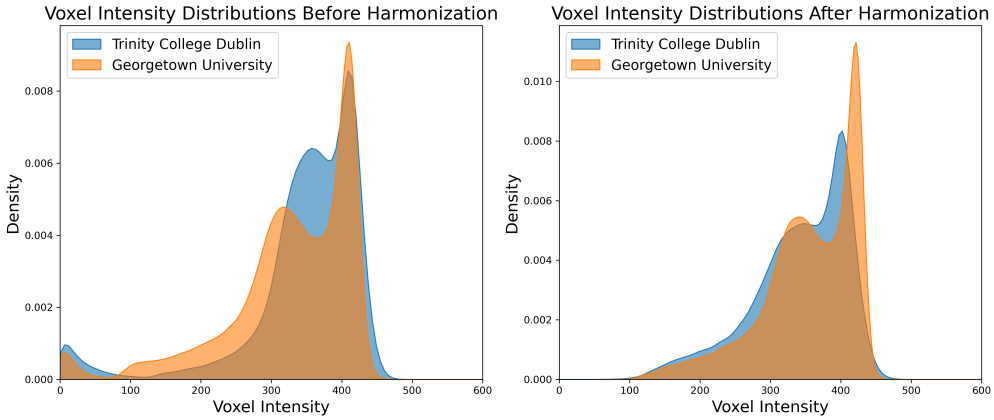


Fig. 2: It illustrates the effect of harmonization, we visualized the voxel intensity distributions from two different sites within our dataset—Trinity College Dublin and Georgetown University—before and after applying ComBat harmonization.

of healthy participants includes 1,351 scans from individuals aged 6 to 80 years, with a mean age of 30.5 years and a standard deviation of 19.95 years. This dataset consists of 872 male and 479 female participants. Based on previous research indicating that gender does not significantly influence brain age estimation [39], we have not included gender analysis here.

The age distribution across different groups within the healthy population is shown in Table I.

| **Age** | 6-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 |
|---|---|---|---|---|---|---|---|---|
| **Samples** | 142 | 420 | 257 | 138 | 112 | 104 | 120 | 58 |

TABLE I: Healthy participants' dataset age distribution.

| **Age** | 6-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
|---|---|---|---|---|---|---|---|
| **Samples** | 82 | 112 | 48 | 6 | 12 | 18 | 2 |

TABLE II: ASD participants' dataset age distribution.

As for the healthy samples split, the dataset is divided into eight age groups: 6-10, 10-20, 20-30, 30-40, 40-50, 50-60, 60-70 and 70-80. For each age group, 70% of the samples were allocated to the training set, 15% to the validation set, and 15% to the test set, ensuring a representative distribution across all subsets.

Regarding the ASD patients' dataset, there are 280 samples included in the experiment. The participants range in age from 6 to 62 years, with a mean age of 18.8 years and a standard deviation of 13.78 years. The detailed age distribution is presented in Table II.

To ensure compatibility and mitigate the potential effects of protocol variability for the different datasets, we applied a standardized preprocessing protocol using FSL 5.10 [40] to the MRI scans. This protocol included several steps: brain extraction [41], bias field correction, nonlinear registration to the MNI standard space, and normalization of voxel values within the brain area by subtracting the mean and dividing by the standard deviation. We also used ComBat harmonization on the datasets to adjust for scanner and site-specific effects while preserving biological variability. After preprocessing, all MRI scans were resized to a voxel dimension of $91 \times 109 \times 91$ with an isotropic spatial resolution of 2mm.

To illustrate the effect of harmonization, we visualized the voxel intensity distributions from two different sites within our dataset—Trinity College Dublin and Georgetown University—before and after applying ComBat harmonization (Figure 2). Prior to harmonization, substantial differences in intensity distributions across datasets were evident, indicating scanner and site-specific variability. After harmonization, intensity distributions became well-aligned, demonstrating the effectiveness of ComBat in reducing unwanted scanner-induced variability while preserving biologically relevant variations.

The code we used in this project is uploaded to Github[3].

### *B. Proposed Triamese-ViT*

In this section, we present our Tri architecture named Triamese-ViT. Our approach is inspired by [42], which highlights that different views of a 3D image contain unique and independent information that can be leveraged in machine learning models. As illustrated in Fig. 3, the structure of Triamese-ViT is based on the Vision Transformer (ViT) [23]. Triamese-ViT processes 3D MRIs, denoted as $M \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and the slice number, respectively. The MRI $M$ is then reshaped into three distinct viewpoints, represented as $M \rightarrow (M_x, M_y, M_z)$, with $M_x \in \mathbb{R}^{H \times W}$ (C channels), $M_y \in \mathbb{R}^{H \times C}$ (W channels), and $M_z \in \mathbb{R}^{W \times C}$ (H channels).

Focusing initially on $M_x$, the MRI is divided into a sequence of flattened 2D squares, denoted as $M_{x,s} \in \mathbb{R}^{N \times (S^2 \cdot C)}$, where the side length of square is $S$, and the number of squares is $N = \frac{H \times W}{S^2}$.

In the transformer encoder layers, the vectors processed are of dimension D. Thus, $M_x$ needs to be mapped to D dimensions using a trainable linear projection. The process is formulated as follows:

$$t_{x,0} = \text{Concat}(M_{x,class}; M^1_{x,s}E; M^2_{x,s}E; \ldots; M^N_{x,s}E) + E_{pos} \quad (1)$$

In Equation 1, $M_{x,class}$ is a learnable token (or class token) added to ViT, akin to the method used in [43]. This class

[3]https://github.com/zhangz59/Triamese-ViT

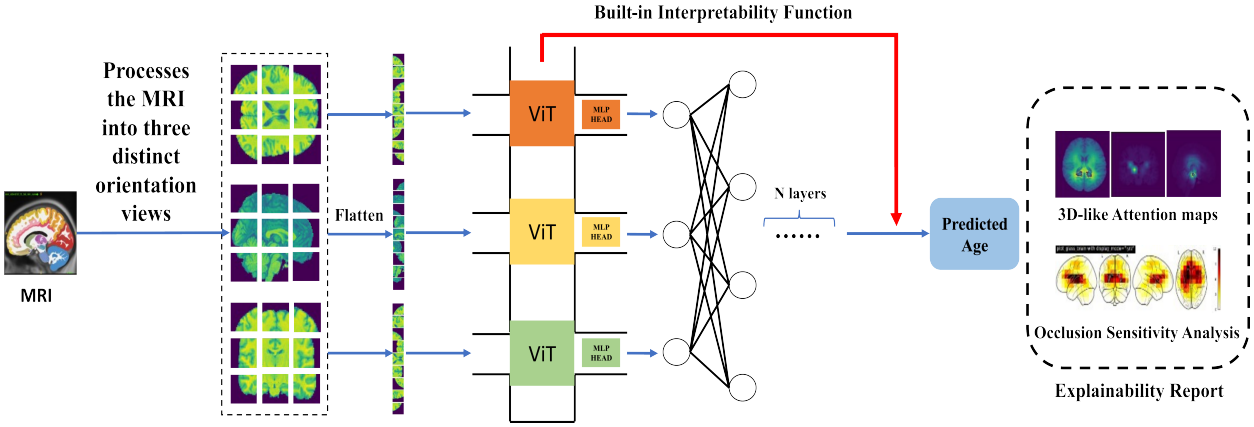


Fig. 3: It depicts the architecture of our model, 'Triamese-ViT'. This model processes brain MRI images from three distinct perspectives utilizing the Vision Transformer (ViT) to extract unique features. These features are then integrated within a Tri Multi-Layer Perceptron (MLP) framework to generate age predictions. And built-in interpretability function generates 3D-like images to explain different brain regions influence during prediction.

token, $M_{x,class}$, is eventually output from the Transformer Encoder as $t^0_{x,L}$, representing the image representation $P$ (Equation 7). Here, $E \in \mathbb{R}^{(S^2 \cdot C) \times D}$ is the linear projection matrix, Concat denotes token concatenation, and $E_{pos} \in \mathbb{R}^{(N+1) \times D}$ is the positional encoding added to each token embedding. $t_{x,0}$ represents the input sequence to the 0-th (first) Transformer encoder layer. The same preprocessing steps are applied to $M_y$ and $M_z$, resulting in $t_{y,0}$ and $t_{z,0}$.

The transformed matrices $t_{x,0}$, $t_{y,0}$, and $t_{z,0} \in \mathbb{R}^{(N+1) \times D}$ are fed into the transformer encoder. Each encoder consists of multiple layers, where each layer sequentially processes the input through Layer Normalization (LN), Multi-Head Attention (MSA), another Layer Normalization, and a Multi-Layer Perceptron (MLP). The MSA performs parallel attention calculations across multiple 'heads', allowing for diverse representation and richer understanding of the input data.

$$[Q, K, V] = FC(t_{x,0}) \tag{2}$$

Here, $Q \in \mathbb{R}^{(N+1) \times d}$, $K \in \mathbb{R}^{(N+1) \times d}$, and $V \in \mathbb{R}^{(N+1) \times d}$ represent the Query, Key, and Value matrices, respectively. Assuming the MSA has $n$ heads and $D = n \times d$, each head independently processes the input:

$$\text{head}_i = \text{softmax}\left(\frac{Q_i K_i^{\mathsf{T}}}{\sqrt{d}}\right) V_i \tag{3}$$

$$\text{MSA}(z_{x,0}) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_n) \tag{4}$$

Let $t_{x,0}$ be the input to the first layer of the Transformer Encoder. The feedforward calculations in the encoder are given:

$$t'_{x,l} = \text{MSA}(\text{LN}(t_{x,l-1})) + t_{x,l-1} \tag{5}$$

$$t_{x,l} = \text{MLP}(\text{LN}(t'_{x,l})) + t'_{x,l} \tag{6}$$

where $l \in [1, 2, \ldots, L]$. The outputs from each Transformer Encoder are then passed to an MLP head, consisting of a hidden layer and an output layer, to generate the final prediction for each view. The prediction from the first view, $M_x$, is denoted as $P_x$. By applying the same procedure to $M_y$ and $M_z$, we obtain two additional predictions, $P_y$ and $P_z$.

Finally, these three view-based predictions ($P_x$, $P_y$, and $P_z$) are fed into the MLP, which integrates the information from all three views to produce the final comprehensive prediction:

$$P_{Tri} = \text{MLP}(P_x, P_y, P_z) \tag{7}$$

Here, $P_{Tri}$ denotes the final prediction.

The motivation for adopting an axis-wise ViT instead of a 3D ViT for brain age estimation lies in several key advantages:

- Lower Computational Cost: Triamese-ViT circumvents the high computational demands of processing an entire 3D volume by decomposing it into three orthogonal 2D views. This significantly reduces the computational complexity, as each view is treated as a 2D input to a standard ViT, which scales linearly with input size. Consequently, Triamese-ViT enables faster training and requires substantially less GPU memory compared to a full 3D ViT, making it more feasible for large-scale 3D medical imaging datasets.
- Model Simplicity and Implementation: By leveraging the well-established 2D Vision Transformer framework, Triamese-ViT maintains a straightforward implementation that requires minimal adaptation. This simplifies model design, debugging, and fine-tuning while allowing the integration of pre-trained weights and existing tools developed for 2D ViTs. In contrast, 3D ViTs necessitate extensive architectural modifications, such as 3D tokenization and positional encoding, which introduce additional computational and technical complexities.
- Higher Predictive Accuracy: Empirical evaluations demonstrate that Triamese-ViT achieves superior prediction accuracy compared to 3D ViTs. This improvement stems from its ability to integrate multiple 2D views, capturing diverse and complementary spatial features from different anatomical perspectives, thereby enhancing the robustness and precision of brain age estimation.

## III. Results

| Algorithm | MAE | $r$ | $\|rp\|$ | $R^2$ | Memory |
|---|---|---|---|---|---|
| ResNet [12] | 4.11 | 0.84 | 0.33 | 0.70 | 958 MB |
| VGG19 [44] | 4.09 | 0.7 | 0.49 | 0.68 | 2.27 GB |
| VGG16 [44] | 5.32 | 0.6 | 0.41 | 0.64 | 2.18 GB |
| 5-layer CNN [39] | 4.55 | 0.79 | 0.47 | 0.71 | 2.46 MB |
| Global-Local Transformer [27] | 4.68 | 0.77 | 0.32 | 0.73 | 617 MB |
| Two-Stage-Age-Network [45] | 3.93 | 0.91 | 0.38 | **0.81** | 1.52 GB |
| Efficient Net [46] | 4.55 | 0.88 | 0.4 | 0.77 | 72 MB |
| Multiple Instance Neuroimage Transformer [38] | 3.90 | 0.9 | 0.36 | 0.77 | 4.62 GB |
| ITSVR [47] | 4.21 | 0.75 | 0.35 | 0.71 | 3.57 GB |
| 3D-TDR [37] | 3.97 | 0.85 | 0.42 | 0.80 | 2.15 GB |
| Our Triamese-ViT | **3.85** | **0.94** | **0.3** | **0.81** | 3.99 GB |

TABLE III: The details of tested algorithms' performance. Since the input of Global-Local Transformer should be a 2D image, we extract 2D slices around the center of the 3D brain volumes in the axial as input, which is the same process method as [27]. Other algorithms' input are 3D MRIs with dimensions (91,109,91). Our Triamese-ViT has consistently achieved the best among all measures.

### *A. Comparison With State-of-the-Art Algorithms for Brain Age Estimation*

We employed Triamese-ViT to estimate brain ages based on MRI scans from a cohort of 1,351 healthy individuals aged between 6 and 80 years. The dataset was divided into 70% for training, 15% for validation, and 15% for testing, allowing for a rigorous assessment of model performance. We evaluated the model using three principal metrics: Mean Absolute Error (MAE), the Spearman correlation coefficient between the predicted and chronological ages ($r$), the absolute value of the Spearman correlation coefficient between chronological age and the Brain Age Gap (BAG) ($|rp|$), and R-Squared ($R^2$) between the predicted and chronological ages. The MAE, $r$, and $R^2$ measure the model's accuracy and the degree of correlation between the predicted and chronological ages, while $|rp|$ quantifies the model's fairness, with a higher $|rp|$ indicating a more pronounced age bias. We compared our Triamese-ViT model against other state-of-the-art algorithms to demonstrate its superior performance in brain age estimation. The results are shown in Table III.

Table III presents a comprehensive comparison of the Triamese-ViT model against eight other models, encompassing both classic and state-of-the-art (SOTA) approaches in brain age estimation. This comparison includes four established 3D CNN-based models: a 5-layer CNN, ResNet, VGG16, and VGG19. Additionally, our model was benchmarked against six other SOTA methodologies: the Two-Stage-Age-Network, which features a two-stage cascade network architecture where the first-stage network estimates a rough brain age and the second-stage network refines this estimate based on the discretized brain age provided by the first-stage network; the Global-Local Transformer, which utilizes 2D brain slices to predict; EfficientNet, known for its ensemble architecture; the Multiple Instance Neuroimage Transformer, which is a 3D transformer architecture that changes 2D patches to 3D blocks in ViT; the ITSVR, an improved twi support vector regression; and the 3D-TDR, a tensor-distribution-regression model based on 3D conventional neural networks.

Triamese-ViT achieves the lowest Mean Absolute Error (MAE) at 3.85, followed closely by the Multiple Instance Neuroimage Transformer (MAE 3.90) and the Two-Stage-Age-Network (MAE 3.93), with VGG16 performing the worst (MAE 5.32). For Spearman correlation ($r$) between predicted and chronological ages, Triamese-ViT leads at 0.94, followed by the Two-Stage-Age-Network (0.91) and Multiple Instance Neuroimage Transformer (0.90); VGG16 again ranks lowest (0.60). In terms of fairness, Triamese-ViT achieves the best (lowest) brain age gap correlation ($|rp|$) at -0.3, while ResNet (0.33) and Global-Local Transformer (0.32) also perform well. Regarding R-squared ($R^2$), both Triamese-ViT and the Two-Stage-Age-Network show strong performances (0.81), with the 3D-TDR following closely (0.80); VGG16 has the weakest result (0.64).

Regarding memory consumption, the 5-layer CNN has the fewest parameters, requiring only 2.46 MB of memory. EfficientNet follows, utilizing 72 MB, as it trains on only a slice of the MRI data. In contrast, the 3D ViT has the highest memory requirement at 4.62 GB. Although Triamese-ViT also demands substantial memory at 3.99 GB, its consumption remains lower than that of the 3D ViT.

This comparative analysis underscores the Triamese-ViT model's superior performance in brain age estimation, highlighting its advantages in accuracy and fairness compared to other leading models in the field.

### *B. Ablation Study*

In this part, we conduct ablation experiments to explore and justify the design choices in the structure of Triamese-ViT. First, we provide the rationale behind selecting the hyperparameter **S** as 7 in ViT. Experimental evaluations revealed that a smaller patch size increases sensitivity but results in overly detailed attention maps, which may introduce noise and hinder interpretability. Conversely, a larger patch size encompasses multiple brain regions within a single patch, reducing the granularity of the attention maps and potentially obscuring critical structural information. The choice of **S** = 7 represents an optimal balance, ensuring sufficient sensitivity while preserving meaningful spatial features for brain structure analysis.

Then, we focus on the number of layers in the Tri-MLP. While keeping all other variables constant, we vary the number of MLP layers and observe their impact on the model's performance. The findings depicted in Fig. 4 show a distinctive trend in the Mean Absolute Error (MAE) relative to the MLP layers in Triamese-ViT MLP. The MAE initially rises when increasing layers from 4 to 6, then decreases after 6 layers, reaching a minimum of 9 layers before rising again at 10 layers. This indicates an optimal layer count for balancing model complexity and accuracy. The observed MAE variation with different layer counts underscores the intricate relationship between model depth and performance, emphasizing the need for precise architectural tuning in the model.

Then we turned our focus to the backbone of Triamese-ViT. To assess the impact of different backbone architectures,

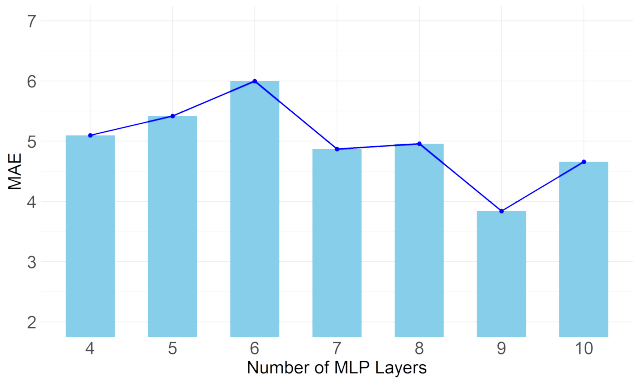


Fig. 4: The impact of the number of MLP layers in Triamese-ViT.

we substituted the original ViT with alternative models like ResNet, a 5-layer CNN, and VGG19. These were then integrated with the Tri-MLP to evaluate how they influenced overall performance. The results of this experiment are detailed in Table IV. According to our findings, the original ViT backbone proves to be the most effective for the Triamese structure. The 5-layer CNN also shows commendable adaptability, registering an MAE of 4, a Spearman correlation (r) of 0.85, $\|rp\|$ of 0.45 and $R^2$ of 0.72. In stark contrast, ResNet and VGG19 appear significantly less suited for the Triamese framework. Both these architectures yielded MAEs exceeding 10, which are highly unfavorable outcomes for brain age estimation. This experiment underscores the importance of selecting an appropriate backbone model for the Triamese structure to ensure optimal performance.

We also investigated alternative fusion strategies for combining the outputs from the three ViT branches in Triamese-ViT model. Specifically, we compared original MLP-based fusion layer against two alternative fusion methods: convolutional attention (using Convolutional Block Attention Module—CBAM) and self-attention mechanisms. Following each fusion method, a four-layer MLP was employed to generate the final predictions. Our experimental results indicate that the CBAM-based fusion strategy achieves promising performance, yielding an MAE of 4.23, a $r$ of 0.81, a $\|rp\|$ of 0.35, and a $R^2$ of 0.78, indicating good accuracy and fairness. Conversely, the self-attention fusion approach demonstrated inferior performance, with an MAE of 6.57, a $r$ of 0.52, $\|rp\|$ of 0.41, and a $R^2$ of 0.64. Nonetheless, both alternative fusion methods underperformed compared to our original Triamese-ViT architecture. These comparative results justify our choice of the MLP-based fusion layer and enrich the robustness and comprehensiveness of our methodological analysis.

Next, we explore the unique structures within our Triamese-ViT model, particularly focusing on the individual contributions of the three Vision Transformers (ViTs) oriented along different axes of the MRIs. These are the $ViT_x$ with dimensions (91,109,91), $ViT_y$ with dimensions (91,91,109), and $ViT_z$ with dimensions (109,91,91). The performance of each of these orientation-specific ViTs is crucial in understanding the efficacy of the combined Triamese-MLP structure.

| Algorithm | MAE | r | $\|rp\|$ | $R^2$ |
|---|---|---|---|---|
| VGG-Backbone | 10.31 | 0.30 | 0.31 | 0.29 |
| ResNet-Backbone | 10.36 | 0.45 | 0.25 | 0.37 |
| CNN-Backbone | 4 | 0.85 | 0.45 | 0.72 |
| CBAM-fusion layer | 4.23 | 0.81 | 0.35 | 0.78 |
| self-attention-fusion layer | 6.57 | 0.52 | 0.41 | 0.64 |
| $ViT_x$ | 4.42 | 0.78 | 0.33 | 0.71 |
| $ViT_y$ | 4.99 | 0.92 | 0.29 | 0.79 |
| $ViT_z$ | 5.29 | 0.73 | 0.37 | 0.7 |
| $ViT_{map}$ | 5.04 | 0.61 | 0.55 | 0.65 |
| Our Triamese-ViT | **3.85** | **0.94** | **0.3** | **0.81** |

TABLE IV: The details of the backbone-changed, fusion-layer changed models and unique structures. $ViT_x$, $ViT_y$, and $ViT_z$ are focusing on the individual contributions of the three Vision Transformers (ViTs) oriented along different axes of the MRIs in Triamese-ViT. $ViT_{map}$ also utilizes three ViTs on different viewpoints but each ViT in $Triamese_{map}$ outputs a feature map from the Transformer Encoder, rather than a direct prediction from the MLP Head. Then the MLP in this variant takes as input the concatenated feature maps from the three ViTs to make the final prediction.

Additionally, we tested a variant model, $Triamese_{map}$, which also utilizes three ViTs on different viewpoints. However, unlike the standard Triamese-ViT, each ViT in $Triamese_{map}$ outputs a feature map from the Transformer Encoder, rather than a direct prediction from the MLP Head. The Triamese MLP in this variant then takes as input the concatenated feature maps from the three ViTs to make the final prediction.

The comparative performance of these models, including each individual orientation ViT and the $Triamese_{map}$ variant, is presented in Table IV. It says Triamese MLP supports a great improvement of performance, for MAE, $ViT_x$ is the second best with 4.42, and $ViT_z$ is the worst with 5.29. As for $r$, $ViT_y$ has the highest value with 0.92, This is closely followed by the combined Triamese-ViT model. Notably, $ViT_{map}$, which uses concatenated feature maps for prediction, shows the lowest correlation value at 0.61. Regarding the aspect of fairness, only $ViT_{map}$ displays a strong negative correlation. This suggests a significant reduction in age bias. Conversely, the other models, including the individual orientation-specific ViTs, exhibit minimal ageism in their predictions. As for $R^2$, $ViT_y$ has the highest value with 0.79, and $ViT_{map}$ has the worst performance with 0.65.

Overall, the data in Table IV strongly supports the efficacy of the Triamese-ViT in enhancing both the accuracy and fairness of brain age estimation, validating its design.

### C. Explainable Results for Brain Age Estimation

As is often the case, the prediction process in deep learning models can resemble a 'black box', where complex architectures and numerous parameters obscure the decision-making process. In this section, we aim to elucidate the predictive strategy of the Triamese-ViT model and enhance its interpretability using two distinct methods. The first method involves the use of 3D-like attention maps generated from the Triamese-ViT, which is a built-in method of the model. Since we input 3D MRIs into the ViTs from three different viewpoints (as

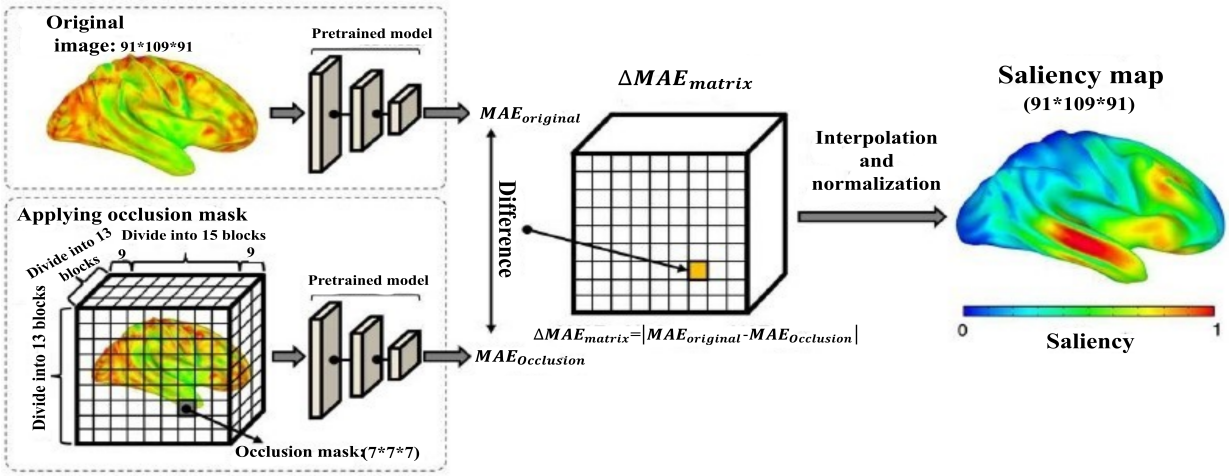


Fig. 5: Illustration of the framework for occlusion analysis. In this work, occlusion analysis systematically obscures regions in brain MRI images using a 7×7×7 voxel mask to assess their impact on model predictions. By measuring changes in Mean Absolute Error (MAE) as the mask moves across the brain, a saliency map is generated, highlighting critical regions for age estimation.

depicted in Figure 3), we obtain three distinct 2D attention maps corresponding to these perspectives. These 2D maps are then expanded into 3D and combined by averaging them to produce a composite 3D attention map.

The other method is a classic XAI method called Occlusion Sensitivity Analysis. Its process is shown in Figure 5. It is a technique that systematically obscures different parts of the input data to evaluate their influence on the model's output. In our case, specific regions within brain MRI images are obscured. This is achieved by applying a cube-shaped occlusion mask, sized at 7×7×7 voxels, which sets the encompassed voxels to zero. We methodically move this mask throughout the entire brain volume, ensuring there is no overlap between successive positions. As the mask traverses the brain, it enables us to observe variations in the model's predictions. These alterations, measured in terms of Mean Absolute Error (MAE), compare the prediction accuracy with and without the occlusion. The degree of change in MAE indicates the relative importance of each brain region. The compilation of these changes forms a saliency map, effectively highlighting the areas that the model predominantly relies on for making its age estimations.

Figure 6 displays the results of our interpretability analyses. Figure 6.a illustrates the results from built-in interpretation compared to original brain, while Figure 6.b shows the outcomes from Occlusion Sensitivity Analysis. Appendix[4] records the detailed values of these two methods. From Figures 6.a and 6.b, it is evident that the brightest spots, indicating regions of highest relevance for age estimation, are centrally located, possibly pointing to deep brain structures. The symmetry in highlighted areas across both hemispheres aligns with the mirrored nature of many brain processes and structures.

Appendix highlights the consistency between the two explainable AI (XAI) methods. For the attention maps, regions with attention values above 3 are considered key, while for Occlusion Sensitivity Analysis, regions with values above 4 are deemed critical. Both methods identify the Rolandic Operculum, Cingulum, and Thalamus as important for brain age prediction. Additionally, attention maps also emphasize the significance of the Vermis, and Occlusion Sensitivity Analysis highlights the Insula, Caudate Nucleus, Putamen, and Heschl's gyrus.

To prove the outstanding interpretability of Triamese-ViT, we compared the explanation results with another inherently interpretable model's. Figure 7 represents the interpretability results from the Global-Local Transformer [27] on brain age estimation. By comparing Figure 7 (Global-Local Transformer) with Figure 6 (Triamese-ViT), it becomes evident that the interpretability of the Triamese-ViT is more informative. The large area of color coverage makes Global-Local Transformer challenging to identify specific highlighted brain structures, and the results are limited to a single top-down view, providing partial information about brain regions.

In contrast, the 3D-like attention maps generated by Triamese-ViT clearly associate attention values with specific brain structures, enabling a more precise understanding of the regions influencing predictions. Additionally, Triamese-ViT provides attention maps from three distinct orientations, offering comprehensive 3D information about the brain.

### *D. Normal Aging Analysis*

The experimental results presented above demonstrate that our Triamese-ViT model not only excels in predicting brain age for healthy samples compared to classic and SOTA algorithms but also offers superior interpretability through its attention maps, compared to traditional explainable AI (XAI) methods. Therefore, in this section, we will apply the Triamese-ViT model to analyze the normal aging process in the human brain.

[4]https://github.com/zhangz59/Triamese-ViT/blob/main/JBHI_Appendix.pdf

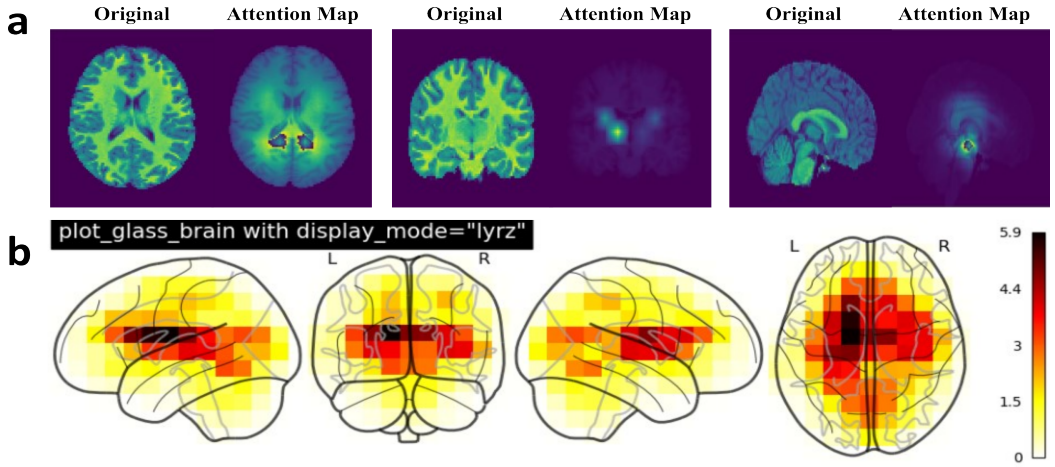


Fig. 6: Comparison between the Triamese-ViT's attention map and occlusion analysis for healthy people. Figure 6.a presents the results from built-in interpretation compared to the original brain, while Figure 6.b shows the outcomes of the occlusion analysis. Together, these sections identify the specific brain regions that the Triamese-ViT model finds most crucial for age prediction.

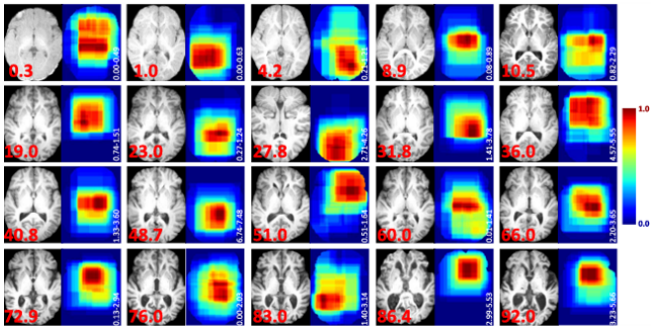


Fig. 7: This figure is from [27]. It shows the interpretability results from the Global-Local Transformer on brain age estimation.

Figure 8 shows attention maps from the Triamese-ViT model across three axes (x, y, z) for predicting brain age in healthy individuals aged 6 to 80 years. Each map represents attention values averaged by decade. Bright regions highlight areas significantly influencing age predictions, predominantly appearing centrally and symmetrically, suggesting deep brain structures are crucial. The decreasing peripheral attention indicates cortical regions may be less critical for age estimation. Notably, high attention towards Thalamus, linked to a relay station for sensory and motor signals, occurs in several age groups [48].

Attention patterns differ by age group. Younger individuals (6-10s) show broad attention distribution, possibly due to rapid brain maturation. From the 10s to 30s, the attention becomes increasingly focused, reflecting stabilized development and emerging age-related structural changes. Starting in the 30s, emphasis on midline structures might relate to aging white matter. In the 40s and 50s, deep brain structures are frequently highlighted. In older adults (60s-70s), attention spreads widely, indicating a wider array of structural changes is becoming more prominent and informative for age estimation.

Overall, these patterns align with established knowledge about brain development and aging—dynamic changes during youth, specific structural shifts in middle age, and widespread changes in later life.

Since the MRIs are in standard MNI space, highlighted regions in Figure 8 can be matched to specific brain structures. We have global-normalized the attention values across all age groups, ensuring that the same intensity in different maps during normal aging analysis corresponds to the same numerical attention value. Figure 9 shows attention trends in brain regions with high attention values during natural aging, derived from the Triamese-ViT attention maps. Appendix details the attention values. For the attention values for the brain regions from Triamese-ViT, we first extracted attention maps from three different views as shown in Figure 8, and then expanded each into a 3D map with dimensions of 91×109×91. By calculating the average values of these 3D attention maps, we obtained the final 3D attention values from Triamese-ViT.

Based on Figure 9, we can observe distinct patterns from machine learning's eyes in how different regions are highlighted during natural aging.

- **Early childhood (0s)**: Significant attention in Inferior Frontal Gyrus, Rolandic Operculum, Cingulum, Calcarine, Caudate Nucleus, Cuneus, Thalamus, and Vermis, with highest values in Thalamus and Rolandic Operculum, indicating critical developmental roles.
- **Adolescence (10s)**: Attention generally decreases but remains notable in the Cingulum and Thalamus, reflecting neural network maturation.
- **Young adulthood (20s–40s)**: Attention stabilizes, with Thalamus and Cingulum always keep the highest values, suggesting maintenance roles in cognitive, emotional regulation and executive function.
- **Middle age (40s–50s)**: Attention generally decreases but remains notable in the Cingulum.
- **Middle-aged and elderly (50s–60s)**: Attention values increased across all regions, particularly within the vermis, thalamus, and Rolandic operculum. This heightened

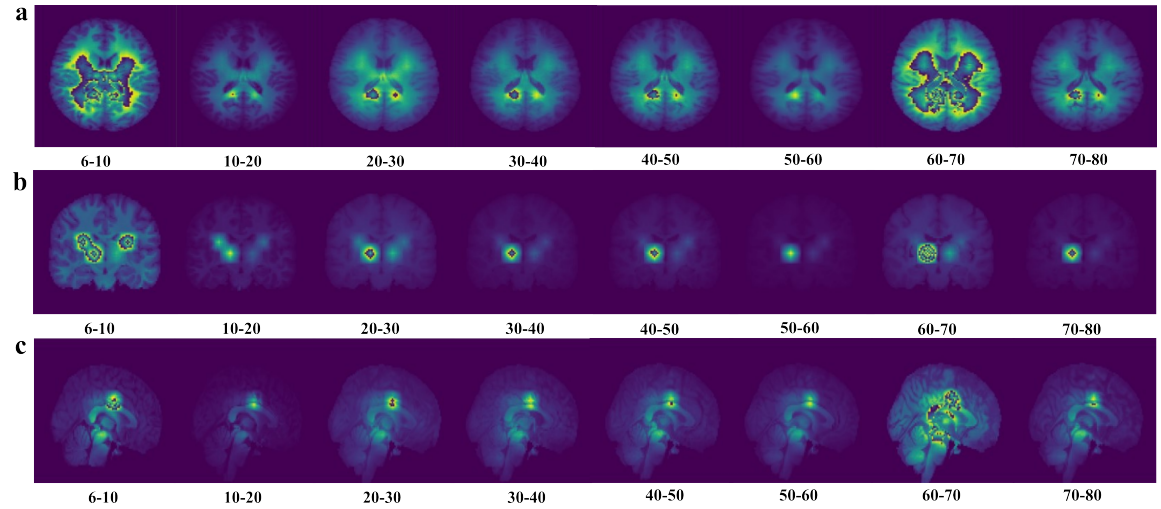


Fig. 8: This figure represents the Triamese-ViT's attention maps from different axes of the MRIs during natural aging from 6 to 80 years old. **a** shows x-axis attention maps, **b** shows y-axis attention maps, and **c** shows z-axis attention maps. Each attention map was calculated by averaging the attention values over each decade.

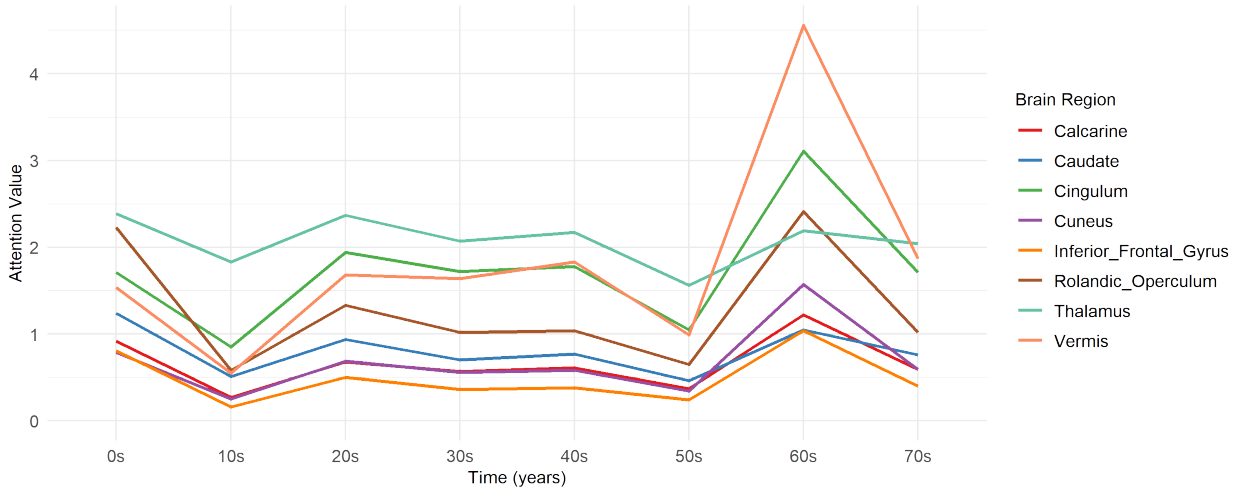


Fig. 9: This figure presents the attention trend lines for the most important regions throughout natural aging based on the Triamese-ViT built-in interpretation.

attention may reflect aging-related changes affecting coordination, balance, and sensorimotor processing.

- **Older adulthood (60s–70s)**: Attention values declined across all regions, suggesting reduced differentiation among individuals within this age group. This decrease may indicate widespread age-related atrophy or reduced functional activity across multiple brain regions.

The consistent prominence of the Thalamus and Cingulum highlights their critical roles throughout the lifespan, supported by existing research on their importance in cognitive networks and vulnerability to pathological aging [49], [50].

### *E. Contribution to ASD Diagnosis*

To demonstrate Triamese-ViT's impact on the disease diagnosis, we applied it to datasets of Autism Spectrum Disorders (ASD) patients, aiming to identify brain regions most affecting ASD. We trained Triamese-ViT on healthy samples, where the attention mechanism was learned during training and remains fixed during prediction. However, attention weights are influenced not only by the learned attention mechanism but also by the input features. If the ASD data differs from the healthy training data (e.g., in brain structure or function), the self-attention mechanism generates different attention maps, as the relationships between input patches differ.

When we apply the Triamese-ViT model trained on healthy samples to ASD data, the resulting attention maps differ from those generated for healthy data. Using the attention maps from healthy samples as a baseline, we can analyze these differences to understand how the identified regions in ASD patients diverge from normal brain aging patterns. This comparison allows us to pinpoint the crucial brain regions associated with ASD, offering valuable insights into the unique characteristics of ASD brains.

Furthermore, we conducted an Occlusion Sensitivity Analysis to benchmark the attention map outcomes of Triamese-ViT.

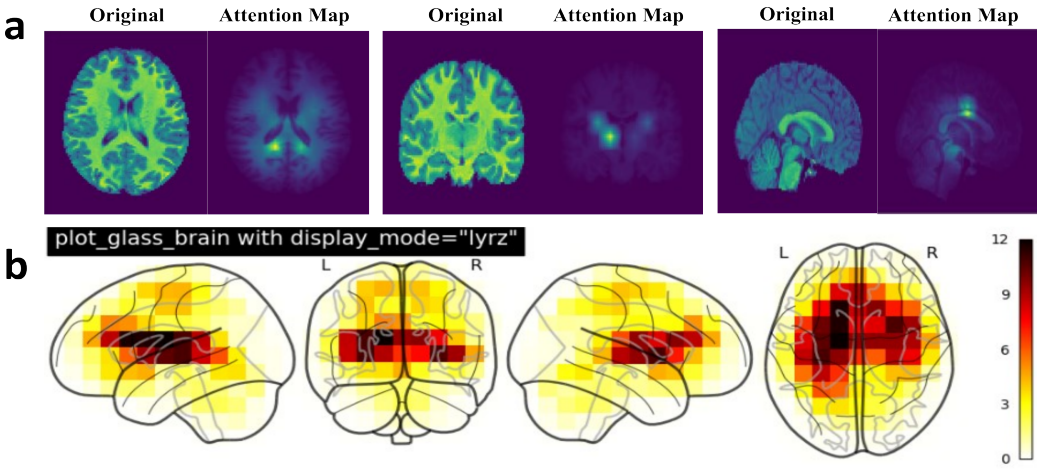


Fig. 10: Comparison between the Triamese-ViT's attention map and occlusion analysis for ASD patients. Figure 10.a presents the attention map results compared to the original brain, while Figure 10.b shows the outcomes of the occlusion analysis. Together, these sections identify the specific brain regions that the Triamese-ViT model finds most crucial for ASD diagnosis.

This analysis involves systematically moving a mask across the brain's entire volume, without overlap, to determine the relative importance of each region. We calculate the importance based on the difference in Brain Age Gap (BAG) before and after occlusion ($BAG_{original}$-$BAG_{occlusion}$). A larger difference indicates the greater importance of a region. It's important to note that positive differences signify significant areas, whereas negative values suggest less crucial regions.

The findings are shown in Figure 10. Figure 10.a presents the built-in interpretation results compared to the original brain, while Figure 10.b shows the outcomes of the occlusion analysis. Appendix says the details of important degrees of different brain regions during diagnosis matching the highlighted areas in Figure 10.a and Figure 10.b.

Both the built-in interpretation and occlusion analysis indicate that the Thalamus plays a significant role in ASD, with occlusion analysis also highlighting the Caudate Nucleus as important for ASD diagnosis. These findings align with existing medical research. For example, [51] reported that individuals with ASD exhibit an expanded surface area in the right posterior thalamus, corresponding to the pulvinar nucleus. They also noted that the shape of the caudal putamen shows a steeper increase in concavity with age in those with ASD. Similarly, [52] examined dynamic functional network connectivity (dFNC) between 51 intrinsic connectivity networks in 170 individuals with ASD and 195 age-matched typically developing (TD) controls using independent component analysis and a sliding window approach. They found that ASD is marked by atypical large-scale subcortical-cortical connectivity, including disrupted resting-state functional connectivity between thalamic and sensory regions. Additionally, [53] compared neuropsychological test scores and caudate volumes in children with ASD, bipolar disorder (BD), and typically developing (TD) children. Their study concluded that children with ASD had larger right and left caudate volumes and modest executive function deficits compared to TD controls.

## IV. Discussion

Our research introduces Triamese-ViT, a deep-learning algorithm designed specifically for brain age estimation. Tested against state-of-the-art (SOTA) models, Triamese-ViT demonstrates superior performance. Its primary innovation is the unique Tri architecture, which integrates comprehensive context understanding with detailed image analysis, effectively capturing complex relationships between image patches. This results in more accurate, precise, and interpretable predictions, significantly advancing clinical applications, especially in early detection of neurodegenerative diseases and personalized brain health assessments.

Evaluated on a public dataset, Triamese-ViT achieved outstanding results, including a Mean Absolute Error (MAE) of 3.85, a Spearman correlation of 0.94 with chronological age, and a favorable -0.3 Spearman correlation between Brain Age Gap (BAG) and chronological age, highlighting reduced age bias. Such accuracy is critical clinically, aiding in detecting deviations from typical aging, potentially indicating early neurodegeneration.

Beyond accuracy, Triamese-ViT's interpretability is a significant advantage. Analyzing attention values across brain regions throughout natural aging revealed distinct patterns: heightened activity in critical developmental regions like the Inferior Frontal Gyrus, Rolandic Operculum, and Thalamus during early childhood; maturation-related decreases during adolescence, except in the Cingulum and Thalamus; cognitive and emotional developments in young adulthood marked by increased Rolandic Operculum and Thalamus activity; stability with slight attention increases in the Thalamus and Vermis during middle age; and a resurgence of attention in various regions, particularly the Vermis, in Middle-aged and elderly adulthood. And in older adulthood, all the regions attention goes down. These findings align well with established medical literature [54]–[61], underscoring Triamese-ViT's practical value in neuroscience and clinical applications.

Triamese-ViT's interpretability significantly enhances disease diagnostics. Conventional diagnosis methods are often

time-intensive and subjective, increasing workload and risk of inaccuracies. The model's clear interpretive outputs, like attention maps, support clinicians by highlighting influential brain regions, thereby improving diagnostic efficiency and accuracy. When tested on an ASD patient dataset, Triamese-ViT identified the Thalamus and Caudate Nucleus as key regions for ASD diagnosis, consistent with existing studies [51]–[53]. Thus, Triamese-ViT provides valuable assistance in early disease detection and targeted interventions, promising enhanced clinical outcomes.

We acknowledge that Triamese-ViT has limitations, including high-frequency fluctuations in its attention maps, which may reduce clarity and interpretability. To address this, future work will explore spatial smoothing and attention regularization to reduce noise and enhance the biological relevance of highlighted patterns.

Integrating multi-modality MRI data—such as T1-weighted, T2-weighted, and diffusion-weighted images—has shown potential to improve model accuracy, robustness, and generalizability [62]. We plan to extend our approach using these modalities to better characterize brain structures and validate the generalization of Triamese-ViT.

This aspect of our findings paves the way for further research and highlights the profound and reliable insights offered by Triamese-ViT. It establishes Triamese-ViT as an invaluable tool for advancing our comprehension of brain aging and related diseases. Future research could focus on validating these findings in clinical trials, exploring the use of Triamese-ViT in personalized treatment plans, and further enhancing its interpretability to better support healthcare professionals in their decision-making processes.

## V. Conclusion

We introduced an innovative deep-learning model for brain age estimation called Triamese-ViT, which surpasses current leading algorithms in terms of accuracy, fairness, and interoperability. We applied Triamese-ViT to the analysis of natural brain aging and ASD diagnosis, yielding meaningful results validated by existing medical research. We believe that Triamese-ViT represents a significant advancement in the integration of AI with medicine, offering promising progress in both brain age estimation and broader medical AI research and development.

## VI. References


[1] I. Beheshti, N. Maikusa, and H. Matsuda, "The association between "brain-age score"(bas) and traditional neuropsychological screening tools in alzheimer's disease," *Brain and Behavior*, vol. 8, no. 8, p. e01020, 2018.

[2] Y. Chung, J. Addington, C. E. Bearden, K. Cadenhead, B. Cornblatt, D. H. Mathalon, T. McGlashan, D. Perkins, L. J. Seidman, M. Tsuang, *et al.*, "Use of machine learning to determine deviance in neuroanatomical maturity associated with future psychosis in youths at clinically high risk," *JAMA psychiatry*, vol. 75, no. 9, pp. 960–968, 2018.

[3] C. Gaser, K. Franke, S. Klöppel, N. Koutsouleris, H. Sauer, and A. D. N. Initiative, "Brainage in mild cognitive impaired patients: predicting the conversion to alzheimer's disease," *PloS one*, vol. 8, no. 6, p. e67346, 2013.

[4] L. K. Han, R. Dinga, T. Hahn, C. R. Ching, L. T. Eyler, L. Aftanas, M. Aghajani, A. Aleman, B. T. Baune, K. Berger, *et al.*, "Brain aging in major depressive disorder: results from the enigma major depressive disorder working group," *Molecular psychiatry*, vol. 26, no. 9, pp. 5124–5139, 2021.

[5] E. Courchesne, H. J. Chisum, J. Townsend, A. Cowles, J. Covington, B. Egaas, M. Harwood, S. Hinds, and G. A. Press, "Normal brain development and aging: quantitative analysis at in vivo mr imaging in healthy volunteers," *Radiology*, vol. 216, no. 3, pp. 672–682, 2000.

[6] C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston, and R. S. Frackowiak, "A voxel-based morphometric study of ageing in 465 normal adult human brains," *Neuroimage*, vol. 14, no. 1, pp. 21–36, 2001.

[7] E. R. Sowell, B. S. Peterson, P. M. Thompson, S. E. Welcome, A. L. Henkenius, and A. W. Toga, "Mapping cortical change across the human life span," *Nature neuroscience*, vol. 6, no. 3, pp. 309–315, 2003.

[8] H. Lemaitre, A. L. Goldman, F. Sambataro, B. A. Verchinski, A. Meyer-Lindenberg, D. R. Weinberger, and V. S. Mattay, "Normal age-related brain morphometric changes: nonuniformity across cortical thickness, surface area and gray matter volume?," *Neurobiology of aging*, vol. 33, no. 3, pp. 617–e1, 2012.

[9] N. Raz and K. M. Rodrigue, "Differential aging of the brain: patterns, cognitive correlates and modifiers," *Neuroscience & Biobehavioral Reviews*, vol. 30, no. 6, pp. 730–748, 2006.

[10] S. Sadigh-Eteghad, B. Sabermarouf, A. Majdi, M. Talebi, M. Farhoudi, and J. Mahmoudi, "Amyloid-beta: a crucial factor in alzheimer's disease," *Medical principles and practice*, vol. 24, no. 1, pp. 1–10, 2015.

[11] L. I. Binder, A. L. Guillozet-Bongaarts, F. Garcia-Sierra, and R. W. Berry, "Tau, tangles, and alzheimer's disease," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1739, no. 2-3, pp. 216–223, 2005.

[12] J. H. Cole and K. Franke, "Predicting age using neuroimaging: innovative brain ageing biomarkers," *Trends in neurosciences*, vol. 40, no. 12, pp. 681–690, 2017.

[13] X. Feng, Z. C. Lipton, J. Yang, S. A. Small, F. A. Provenzano, A. D. N. Initiative, F. L. D. N. Initiative, *et al.*, "Estimating brain age based on a uniform healthy population with deep learning and structural magnetic resonance imaging," *Neurobiology of aging*, vol. 91, pp. 15–25, 2020.

[14] K. Armanious, S. Abdulatif, W. Shi, S. Salian, T. Küstner, D. Weiskopf, T. Hepp, S. Gatidis, and B. Yang, "Age-net: An mri-based iterative framework for brain biological age estimation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 7, pp. 1778–1791, 2021.

[15] Z. Zhang, R. Jiang, C. Zhang, B. Williams, Z. Jiang, C.-T. Li, P. Chazot, N. Pavese, A. Bouridane, and A. Beghdadi, "Robust brain age estimation based on smri via nonlinear age-adaptive ensemble learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2146–2156, 2022.

[16] R. Jiang, P. Chazot, N. Pavese, D. Crookes, A. Bouridane, and M. E. Celebi, "Private facial prediagnosis as an edge service for parkinson's dbs treatment valuation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2703–2713, 2022.

[17] D. Konar, S. Bhattacharyya, T. K. Gandhi, B. K. Panigrahi, and R. Jiang, "3-d quantum-inspired self-supervised tensor network for volumetric segmentation of medical images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 10312–10325, 2024.

[18] Z. Zhang, R. Jiang, D. Crookes, and P. Chazot, "Machine learning-based biological ageing estimation technologies: A survey," in *Recent Advances in AI-enabled Automated Medical Diagnosis*, pp. 96–109, CRC Press, 2022.

[19] J. Hong, Z. Feng, S.-H. Wang, A. Peet, Y.-D. Zhang, Y. Sun, and M. Yang, "Brain age prediction of children using routine brain mr images via deep learning," *Frontiers in Neurology*, vol. 11, p. 584682, 2020.

[20] L. Bellantuono, L. Marzano, M. La Rocca, D. Duncan, A. Lombardi, T. Maggipinto, A. Monaco, S. Tangaro, N. Amoroso, and R. Bellotti, "Predicting brain age with complex networks: From adolescence to adulthood," *NeuroImage*, vol. 225, p. 117458, 2021.

[21] Y. Hu, H. Wang, and B. Li, "Sqet: Squeeze and excitation transformer for high-accuracy brain age estimation," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1554–1557, IEEE, 2022.

[22] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, pp. 1–74, 2021.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*,

"An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[25] M. Tanveer, M. Ganaie, I. Beheshti, T. Goel, N. Ahmad, K.-T. Lai, K. Huang, Y.-D. Zhang, J. Del Ser, and C.-T. Lin, "Deep learning for brain age estimation: A systematic review," *Information Fusion*, 2023.

[26] K. Al-Hammuri, F. Gebali, A. Kanan, and I. T. Chelvan, "Vision transformer architecture and applications in digital health: a tutorial and survey," *Visual Computing for Industry, Biomedicine, and Art*, vol. 6, no. 1, p. 14, 2023.

[27] S. He, P. E. Grant, and Y. Ou, "Global-local transformer for brain age estimation," *IEEE transactions on medical imaging*, vol. 41, no. 1, pp. 213–224, 2021.

[28] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[29] S. Lundberg, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[30] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," *Advances in neural information processing systems*, vol. 32, 2019.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[32] U. Bhalla, S. Srinivas, and H. Lakkaraju, "Discriminative feature attributions: bridging post hoc explainability and inherent interpretability," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[33] T. J. Hastie, "Generalized additive models," in *Statistical models in S*, pp. 249–307, Routledge, 2017.

[34] J. Chen, L. Song, M. Wainwright, and M. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," in *International conference on machine learning*, pp. 883–892, PMLR, 2018.

[35] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: deep learning for interpretable image recognition," *Advances in neural information processing systems*, vol. 32, 2019.

[36] M. Böhle, M. Fritz, and B. Schiele, "B-cos networks: Alignment is all we need for interpretability," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10329–10338, 2022.

[37] L. Chen and X. Luo, "Tensor distribution regression based on the 3d conventional neural networks," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 7, pp. 1628–1630, 2023.

[38] A. Singla, Q. Zhao, D. K. Do, Y. Zhou, K. M. Pohl, and E. Adeli, "Multiple instance neuroimage transformer," in *International Workshop on PRedictive Intelligence In MEdicine*, pp. 36–48, Springer, 2022.

[39] B. Couvy-Duchesne, J. Faouzi, B. Martin, E. Thibeau-Sutre, A. Wild, M. Ansart, S. Durrleman, D. Dormont, N. Burgos, and O. Colliot, "Ensemble learning of convolutional neural network, support vector machine, and best linear unbiased predictor for brain age prediction: Aramis contribution to the predictive analytics competition 2019 challenge," *Frontiers in Psychiatry*, vol. 11, p. 593336, 2020.

[40] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.

[41] S. M. Smith, "Fast robust automated brain extraction," *Human brain mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[42] R. Jiang, A. T. Ho, I. Cheheb, N. Al-Maadeed, S. Al-Maadeed, and A. Bouridane, "Emotion recognition from scrambled facial images via many graph embedding," *Pattern Recognition*, vol. 67, pp. 245–251, 2017.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[44] T.-W. Huang, H.-T. Chen, R. Fujimoto, K. Ito, K. Wu, K. Sato, Y. Taki, H. Fukuda, and T. Aoki, "Age estimation from brain mri images using deep learning," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 849–852, IEEE, 2017.

[45] J. Cheng, Z. Liu, H. Guan, Z. Wu, H. Zhu, J. Jiang, W. Wen, D. Tao, and T. Liu, "Brain age estimation from mri using cascade networks with ranking loss," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3400–3412, 2021.

[46] K. M. Poloni, R. J. Ferrari, A. D. N. Initiative, *et al.*, "A deep ensemble hippocampal cnn model for brain age estimation applied to alzheimer's diagnosis," *Expert Systems with Applications*, vol. 195, p. 116622, 2022.

[47] M. Ganaie, M. Tanveer, and I. Beheshti, "Brain age prediction using improved twin svr," *Neural Computing and Applications*, vol. 36, no. 1, pp. 53–63, 2024.

[48] S. N. Shah, M.-E. Dounavi, P. A. Malhotra, B. Lawlor, L. Naci, I. Koychev, C. W. Ritchie, K. Ritchie, and J. T. O'Brien, "Dementia risk and thalamic nuclei volumetry in healthy midlife adults: The prevent dementia study," *Brain Communications*, vol. 6, no. 2, p. fcae046, 2024.

[49] N. Cera, R. Esposito, F. Cieri, and A. Tartaro, "Altered cingulate cortex functional connectivity in normal aging and mild cognitive impairment," *Frontiers in Neuroscience*, vol. 13, p. 857, 2019.

[50] R. Fama and E. V. Sullivan, "Thalamic structures and associated cognitive functions: Relations with age and aging," *Neuroscience & Biobehavioral Reviews*, vol. 54, pp. 29–37, 2015.

[51] M. Schuetze, M. T. M. Park, I. Y. Cho, F. P. MacMaster, M. M. Chakravarty, and S. L. Bray, "Morphological alterations in the thalamus, striatum, and pallidum in autism spectrum disorder," *Neuropsychopharmacology*, vol. 41, no. 11, pp. 2627–2637, 2016.

[52] Z. Fu, Y. Tu, X. Di, Y. Du, J. Sui, B. B. Biswal, Z. Zhang, N. de Lacy, and V. D. Calhoun, "Transient increased thalamic-sensory connectivity and decreased whole-brain dynamism in autism," *Neuroimage*, vol. 190, pp. 191–204, 2019.

[53] G. T. Voelbel, M. E. Bates, J. F. Buckman, G. Pandina, and R. L. Hendren, "Caudate nucleus volume and cognitive performance: Are they related in childhood psychopathology?," *Biological psychiatry*, vol. 60, no. 9, pp. 942–950, 2006.

[54] S. Sutoko, H. Atsumori, A. Obata, T. Funane, A. Kandori, K. Shimonaga, S. Hama, S. Yamawaki, and T. Tsuji, "Lesions in the right rolandic operculum are associated with self-rating affective and apathetic depressive symptoms for post-stroke patients," *Scientific reports*, vol. 10, no. 1, p. 20264, 2020.

[55] I. A. Humbert, D. G. McLaren, K. Kosmatka, M. Fitzgerald, S. Johnson, E. Porcaro, S. Kays, E.-O. Umoh, and J. Robbins, "Early deficits in cortical control of swallowing in alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 19, no. 4, pp. 1185–1197, 2010.

[56] O. Karaman, H. Çakın, A. Alhudhaif, and K. Polat, "Robust automated parkinson disease detection based on voice signals with transfer learning," *Expert Systems with Applications*, vol. 178, p. 115013, 2021.

[57] Y. Zhang, N. Schuff, G.-H. Jahng, W. Bayne, S. Mori, L. Schad, S. Mueller, A.-T. Du, J. Kramer, K. Yaffe, *et al.*, "Diffusion tensor imaging of cingulum fibers in mild cognitive impairment and alzheimer disease," *Neurology*, vol. 68, no. 1, pp. 13–19, 2007.

[58] K. Wiltshire, L. Concha, M. Gee, T. Bouchard, C. Beaulieu, and R. Camicioli, "Corpus callosum and cingulum tractography in parkinson's disease," *Canadian Journal of Neurological Sciences*, vol. 37, no. 5, pp. 595–600, 2010.

[59] J. Xuereb, E. Perry, J. Candy, J. Bonham, R. Perry, and E. Marshall, "Parameters of cholinergic neurotransmission in the thalamus in parkinson's disease and alzheimer's disease," *Journal of the neurological sciences*, vol. 99, no. 2-3, pp. 185–197, 1990.

[60] B. Maiti, K. S. Rawson, A. B. Tanenbaum, J. M. Koller, A. Z. Snyder, M. C. Campbell, G. M. Earhart, and J. S. Perlmutter, "Functional connectivity of vermis correlates with future gait impairments in parkinson's disease," *Movement Disorders*, vol. 36, no. 11, pp. 2559–2568, 2021.

[61] S. Ho Park, M. Kim, D. L. Na, and B. S. Jeon, "Magnetic resonance reflects the pathological evolution of wernicke encephalopathy," *Journal of Neuroimaging*, vol. 11, no. 4, pp. 406–411, 2001.

[62] R. J. Jirsaraie, A. J. Gorelik, M. M. Gatavins, D. A. Engemann, R. Bogdan, D. M. Barch, and A. Sotiras, "A systematic review of multimodal brain age studies: Uncovering a divergence between model accuracy and utility," *Patterns*, vol. 4, no. 4, 2023.