Brain Age Prediction via VAE and Transformers on Multi-Pipeline OpenBHB Data

Jason Manassa, Sean Hickey, Kexin Guo, Xiangeng Fang, Yubo Shao

Email: {jmanassa, hicsea, kexinguo, xgfang, ybshao}@umich.edu

Abstract

Predicting brain age from MRI scans is a valuable tool for early detection of neurodegenerative diseases, personalized medicine, and generally understanding the health of an individual's brain. Many approaches have been implemented using machine learning models to predict brain age, but a significant roadblock for the generalization of these methods is the inconsistency in input data for these models. Several methods such as ComBat have been proposed to mitigate site variability, but can be limited by the accuracy of the prediction model paired with these methods. This project aims to improve the integration of the ComBat method with a Variational Autoencoder (VAE) to minimize site variability while maximizing brain age prediction accuracy.

1 Introduction

1.1 Motivation and Background

Predicting chronological age and sex of patients from MRI brain scans remains a longstanding healthcare challenge due to the variety of MRI acquisition sites. Recent advancements in machine learning models, specifically deep learning architectures, often fail with generalization due to differences in the MRI scans themselves due to "site effects", i.e. scanner manufacturer induced variations in the datasets (Baecker et al., 2021; Peng et al., 2021; Jónsson et al., 2019). For example, magnetic field strengths, acquisition times, and subject placement may be slightly different between datasets. By solving this data processing problem, accurate brain age prediction could be utilized to identify neurodegenerative diseases such as Alzheimer's, and neurodevelopmental disorders such as schizophrenia. Furthermore, if this work is successful it could be adapted to improve neuroimaging-based diagnostics and offer insights into how to train on differently collecting health datasets.

This effort was completed in the context of the OpenBHB challenge titled "OpenBHB challenge: predicting brain age with site-effect removal, A data challenge on Healthy Controls." (Dufumier et al., 2022a) At the time of this work, there were 22 participants with 166 submissions. As part of the challenge, the authors provided baselines of CNNs with various architectures, shown in Table 1. These metrics, along with the current leader that reported a challenge metric of 1.468 ± 0.012 and an external mean absolute error of 3.564 ± 0.004, were our reference point for success on this project. The measure of success for the method proposed in this paper is how well the method can perform on these metrics using the same input data.

Table 1: Internal and external MAE, site-prediction balanced accuracy, and final score L_c for DenseNet, ResNet and AlexNet, with and without ComBat de-biasing.

De-biasing	Model (dims)	Int. MAE	Ext. MAE	Site Pred. BAcc (%)	Lc (final score)
–	DenseNet(1024)	2.550 ± 0.009	7.13 ± 0.05	8.0 ± 0.9	3.34
–	ResNet(512)	2.67 ± 0.05	4.18 ± 0.01	6.7 ± 0.1	1.86
–	AlexNet(128)	2.72 ± 0.01	4.66 ± 0.05	8.3 ± 0.2	2.21
ComBat	DenseNet(1024)	5.92 ± 0.01	10.48 ± 0.17	2.23 ± 0.06	5.08
ComBat	ResNet(512)	4.150 ± 0.009	4.76 ± 0.03	4.5	1.88
ComBat	AlexNet(128)	3.37 ± 0.01	5.23 ± 0.12	6.8 ± 0.3	2.33

1.2 Related Work and Baseline Solutions

Several brain age prediction models have been developed, spanning both traditional machine learning and deep learning architectures, such as convolutional neural networks (CNNs) (Jónsson et al., 2019; Baecker et al., 2021; Peng et al., 2021). More recently, specialized models like ResNet and DenseNet have gained popularity for brain age prediction (Yook et al., 2021; Zhang et al., 2024). Despite their effectiveness, these models remain susceptible

to site-related biases. ComBat, a widely used residualization technique for mitigating site variability, requires explicit site information and must be retrained when new sites are introduced (Johnson et al., 2007; Jia et al., 2024). Alternative approaches, such as adversarial training, have also been explored (Gadewar et al., 2023; Usman et al., 2024), but many lack standardized benchmarks, making their accuracy and effectiveness difficult to assess.

For our modeling approach, we primarily follow the framework of (Redekop et al., 2024), originally developed for prostate cancer MRI analysis. Other methods applied to similar MRI tasks include 3D ViT (Pachetti et al., 2022) and LoGo MIL (Redekop et al., 2022). In this study, we compare our model against CNN and ResNet-based approaches.

2 Proposed Method

2.1 Dataset and Preprocessing

In our study, we utilize OpenBHB, a large-scale, multi-institutional brain MRI dataset, to support robust and generalizable model training (Dufumier et al., 2022b). OpenBHB consists of 5,330 subjects across a lifespan range of 5 to 88 years and spans multiple international sites across Europe, North America, and China. The dataset integrates T1-weighted (T1w) MRI scans from ten publicly available sources, including ABIDE I, ABIDE II, CoRR, GSP, IXI, Localizer, MPI-Leipzig, NAR, NPC, and RBP. While only T1w images are consistently available across all datasets, some subsets also provide T2-weighted (T2w), diffusion-weighted imaging (DWI), and resting-state fMRI (rs-fMRI) scans.

OpenBHB ensures cross-site harmonization by including only healthy subjects in the current release and standardizing acquisition settings where possible. The dataset encompasses 71 acquisition sites, with nine excluded due to unavailable acquisition details. The median age of participants is 25.3 $\pm$ 15 years, with a sex distribution of 52.1% male.

OpenBHB is openly accessible via IEEE Dataport and serves as a valuable resource for advancing neuroimaging research. OpenBHB provides three levels of data preprocessing: quasi-raw, voxel-based morphometry (VBM) with CAT12, and surface-based morphometry (SBM) with FreeSurfer. These preprocessed datasets enable site-debiased representation learning and facilitate tasks such as brain age prediction. Our study uses VBM data as the input, primarily because it analyzes the entire brain volume voxel by voxel and is less sensitive to artifacts and preprocessing variability. Our experiments show that only the VBM data produced stable results during feature extraction and reconstruction, whereas SBM and quasi-raw data failed to capture meaningful features within our methodological framework.

2.2 Methods

In this project, our goal is to utilize OpenBHB to develop a foundation model (FM) that preserves age-related variability while reducing site-specific biases. Our approach consists of three main components. First, we apply ComBat harmonization (Johnson et al., 2007; Fortin et al., 2017) to address site effects—an inherent challenge in multi-site MRI data collection. Next, we employ a Variational Autoencoder (VAE) (Kingma et al., 2013) to extract representative features from the MRI scans. Finally, we incorporate the trained VAEs into a foundation model using a transformer-based architecture. The latter two steps are inspired by the work of Redekop et al. (2024), which utilized similar structures for disease detection in prostate cancer MRI.

2.3 ComBat Harmonization

A key aspect of our approach is addressing site variability in multi-center MRI datasets. Differences in scanner hardware, acquisition protocols, and preprocessing pipelines introduce non-biological variance. To mitigate these effects, we incorporate ComBat, a statistical harmonization technique originally developed for gene expression studies and later adapted for neuroimaging (Johnson et al., 2007; Fortin et al., 2017).

We follow the formulation introduced by Johnson et al. (2007) and further adapted for brain imaging studies by Fortin et al. (2017, 2018). Let Y_{ijk} represent the observed MRI feature for subject i at site j under preprocessing pipeline k, we fit the model:

$$Y_{ijk} = \alpha_{jk} + X_i\beta_k + \gamma_{jk} + \delta_{jk}\epsilon_{ijk}, \tag{1}$$

where α_{jk} is the site-specific intercept term, X_i represents the biological covariates with regression coefficients β_k, γ_{jk} and δ_{jk} represent the additive and multiplicative batch effects, and the error term ϵ_{ijk}, for which is assumed to follow a Normal distribution with expected value of zero and variance σ_k^2. The batch-adjusted features Y^*_{ijk} would serve as input to our downstream analysis.

This harmonization is particularly relevant for OpenBHB, which aggregates MRI data worldwide with diverse acquisition settings. Prior studies have validated ComBat's effectiveness in harmonizing MRI-derived features, demonstrating its ability to reduce unwanted sources of scan variability (Fortin et al., 2018). We first loaded and flattened the 3D brain volumes into 1D feature vectors and applied ComBat using site as the batch variable. After

harmonization, we reshaped the data back into 3D form so that it could be used as input for downstream VAE-based representation learning. To evaluate its contribution, we will conduct an ablation experiment comparing models trained on raw (non-harmonized) MRI features and features harmonized using ComBat. This baseline comparison will help verify the extent to which ComBat reduces multi-site differences while preserving meaningful variability in brain aging patterns.

2.4 Variational Autoencoder (VAE)

We utilize a VAE framework to learn compact and meaningful latent representations from high-dimensional 3D VBM brain MRI data, where each input volume has dimensions of $121 \times 145 \times 121$. The encoder consists of a 3D convolutional neural network (CNN) with two convolutional layers (kernel size 3^3, padding 1), each followed by a 2^3 max-pooling layer, reducing the input resolution to $30 \times 36 \times 30$ and increasing the number of feature channels to 32. The use of 3D convolutions enables the model to capture structural patterns across all anatomical planes, preserving volumetric context that would be lost with slice-wise 2D processing. The resulting feature maps are flattened and passed through two fully connected layers to output the mean $\boldsymbol{\mu} \in \mathbb{R}^{128}$ and log-variance $\log \boldsymbol{\sigma}^2 \in \mathbb{R}^{128}$, which parameterize the approximate posterior distribution. To enable backpropagation through the sampling process, we apply the standard reparameterization trick: a latent code is sampled as $\tilde{\boldsymbol{z}} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$. The decoder maps $\tilde{\boldsymbol{z}}$ back to the original image space by first projecting it through a fully connected layer, reshaping to the intermediate spatial dimensions, and then applying two transposed convolutional layers. ReLU activations and dropout regularization (with a rate of 0.1) are used during decoding. The model is trained by minimizing a loss function comprising a reconstruction term (mean squared error between input and output volumes) and a regularization term that penalizes divergence from the prior:

$$\mathcal{L}_{\text{VAE}} = \text{MSE}(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \cdot D_{\text{KL}}(q(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})), \tag{2}$$

We employ early stopping based on validation loss, with a patience of 5 epochs to promote generalization and prevent overfitting.

Note that our original objective was to train the VAE model on all three preprocessing pipelines. However, attempts to train VAEs on the quasi-raw and CAT12 pipelines were unsuccessful, as the models failed to converge. As a result, we chose to proceed exclusively with the latent features extracted from the VBM pipeline for all subsequent modeling and analysis.

2.5 Multi-head Transformer

From prior processing using a VAE, each subject's brain MRI is encoded into a 128-dimensional latent vector. This vector serves as input to a Transformer-based regression model designed to predict chronological brain age.

The model interprets the 128 features as a one-dimensional sequence of scalar tokens. Each scalar is independently projected into a higher-dimensional embedding space via a shared linear transformation ($1 \to d_{model}$, where $d_{model} = 64$). To incorporate positional information across feature dimensions, fixed sinusoidal positional encodings are added to the token embeddings, following the original Transformer formulation (Vaswani et al., 2017).

The encoded sequence is then passed through a stack of Transformer encoder layers. Each layer consists of multi-head self-attention (with 4 heads) followed by a position-wise feedforward network. Residual connections and layer normalization are applied at each sub-layer. In our experiments, we used 2 encoder layers with a feedforward dimension of 128 and a dropout rate of 0.1.

After processing the sequence, we apply mean pooling across the sequence dimension to aggregate feature-wise embeddings into a fixed-length vector. This pooled representation is passed through a linear output layer to generate a scalar prediction of chronological brain age.

The model is trained using the Mean Squared Error (MSE) loss function. Hyperparameters such as the number of attention heads, number of layers, and hidden dimensions were selected based on 5-fold cross-validation over the training set. Early stopping based on validation loss was employed to prevent overfitting.

3 Results

3.1 Combat

We visualized the mean voxel intensity per subject across several imaging sites in the training dataset, before and after ComBat adjustment. See Figure 1. Before harmonization, subject-level mean intensities showed noticeable differences between sites, reflecting scanner-induced variability. After applying ComBat, these site-wise distributions became more consistent, with reduced inter-site variability and more overlapping interquartile ranges. This demonstrates that ComBat successfully removed global intensity shifts across sites while preserving within-site variance, making the data more suitable for downstream analysis.

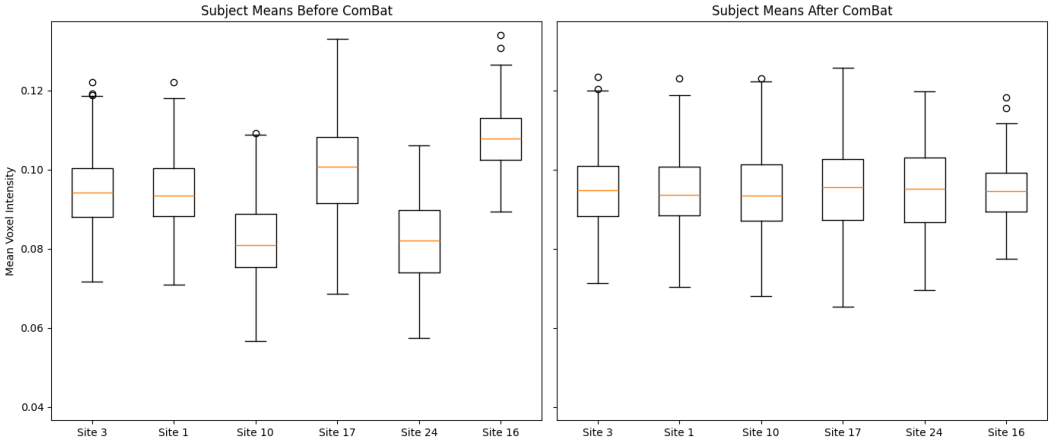


Figure 1: Comparison of subject mean across different imaging sites before and after ComBat.

3.2 VAE

From Figure 2 we observe that both the training and validation losses decrease with epochs. Since we used dropout during the training process, the training loss is higher than the validation loss during the initial few epochs. After approximately 47 epochs, the validation loss stabilizes, and based on our previously mentioned early stopping criterion, we conclude that the model has converged. The extracted 128-dimensional features are then fed into a transformer for brain age prediction.

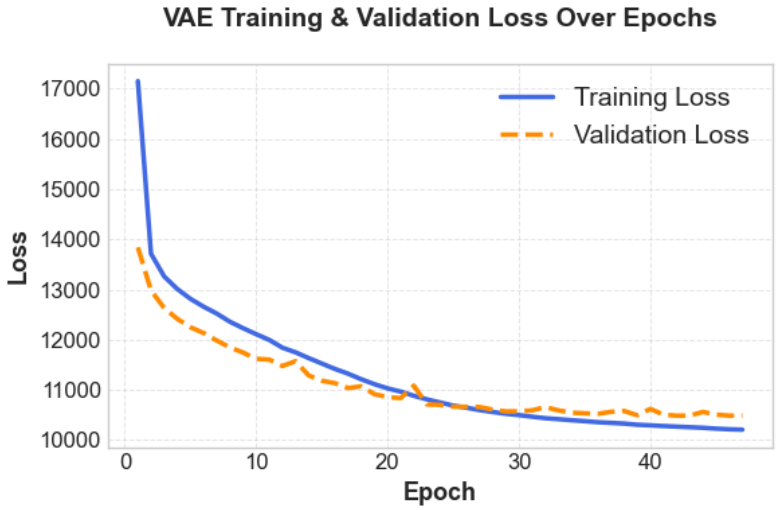


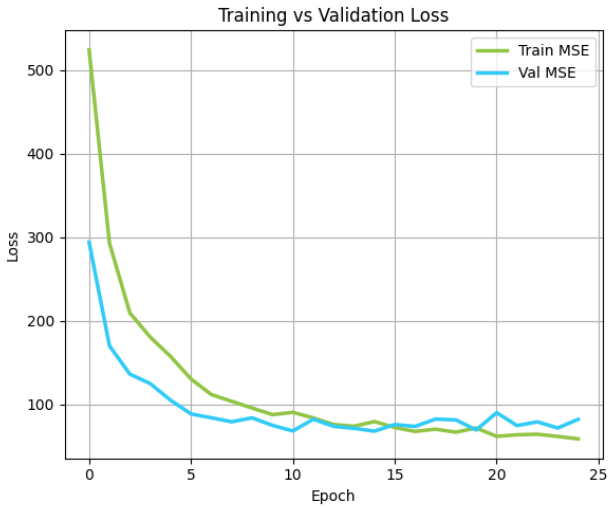
Figure 2: The training and validation loss of the Variational Autoencoder (VAE) model over epochs. The blue curve represents the training loss, while the orange dashed curve represents the validation loss.

3.3 Transformer

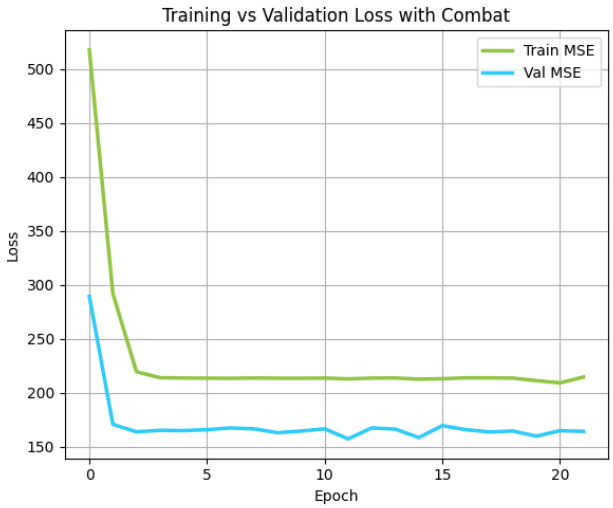
Figure 3a and Figure 3b compare the training and validation loss curves of the Transformer model trained on VBM features without and with ComBat harmonization, respectively. The model trained without ComBat correction achieves lower validation loss and exhibits stable convergence within the first 20 epochs, suggesting effective learning and good generalization to the validation set. In contrast, the model trained on ComBat-harmonized features converges more slowly and plateaus at a higher loss level, indicating suboptimal learning performance. This suggests that, in this setting, ComBat harmonization may inadvertently remove informative within-site variance or distort the latent space learned by the VAE, ultimately degrading model performance.

Figure 4 further illustrates the effect of ComBat harmonization by comparing true versus predicted age on the validation set for Transformer models trained with and without harmonization. In the left panel (Figure 4a), the model trained on unharmonized VBM features exhibits a strong linear relationship between predicted and true age, with predictions generally aligned along the identity line. While some underestimation is observed for older individuals, the overall predictions are well-calibrated and show good dynamic range.

In contrast, the right panel (Figure 4b) shows the predictions from the model trained on ComBat-harmonized features. Here, the predicted ages are largely compressed within a narrow range (approximately 25–45 years), with little variation across different true age groups. This leads to severe underestimation of older subjects and

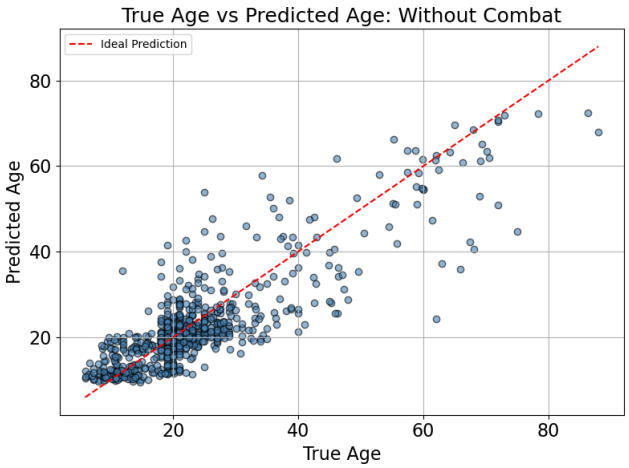


(a) Without Combat harmonization.

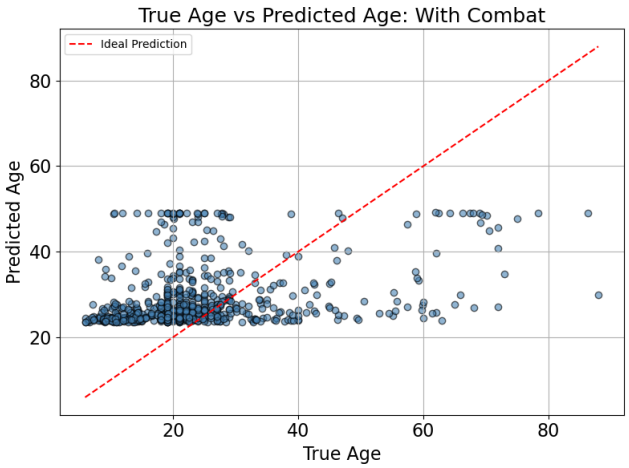


(b) With Combat harmonization.

Figure 3: Training and validation loss of the Transformer model.



(a) Without Combat harmonization.



(b) With Combat harmonization.

Figure 4: True versus predicted brain age on the validation set by the transformer model.

overestimation of younger ones, indicating a collapse in the predictive range. These results suggest that ComBat may have removed biologically relevant variance critical for age prediction.

Although the test set of OpenBHB is not directly accessible, we evaluate the performance of our method on the validation set. The model achieves an MAE of **5.62** when trained on unharmonized VBM features, and an MAE of **8.7** when trained on ComBat-harmonized features. This trend is consistent with the pattern observed in Table 1, where ComBat harmonization generally reduces site-predictive accuracy but at the cost of increased prediction error. However, our method appears less effective overall compared to the CNN-based architectures reported in the table. We hypothesize that this may be partially due to the failure to utilize latent features from all three preprocessing pipelines, as our VAE models did not converge successfully on the quasi-raw and CAT12 variants. Consequently, the Transformer model was limited to learning from a narrower representation space, which may have constrained its predictive performance.

3.4 Comparing our method to an MLP

In order to create a ground truth to compare our pipeline to, we utilized a multi-layer perceptron that tries to predict age off the voxel-based morphometry dataset from earlier. The MLP works by first loading the 3D VBM data and associated age labels, flattening said data, and standardizing it. A custom PyTorch MLP with two hidden layers, dropout regularization, and ReLU activations was trained using mean squared error loss and the Adam optimizer over 200 epochs with validation splitting. The MLP was wrapped in a scikit-learn regressor class which allowed us to integrate it in a larger pipeline that included standardization and principal component analysis for dimension reduction to just 100 features. We validated the model using a 5-fold cross validation technique, leveraging both RMSE and R^2 training vs validation loss curves. Our prediction from our ComBat, VAE, transformer pipeline failed in similar fashion to the MLP. Notably, the MLP was ran utilizing priors about site location, whereas after ComBat in our pipeline we lose this information, therefore to achieve similar results to an MLP that does not remove these features is notable.

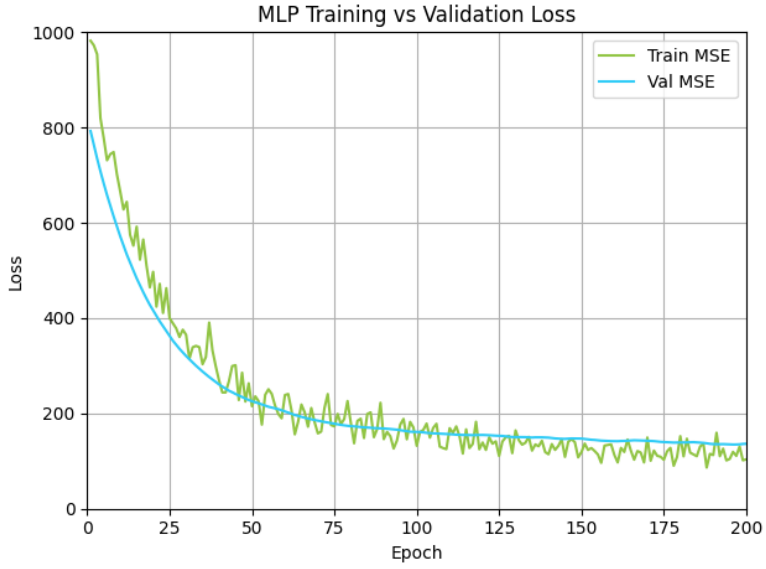


Figure 5: The training and validation loss of the MLP model over epochs. The green curve represents the training loss, while the blue curve represents the validation loss.

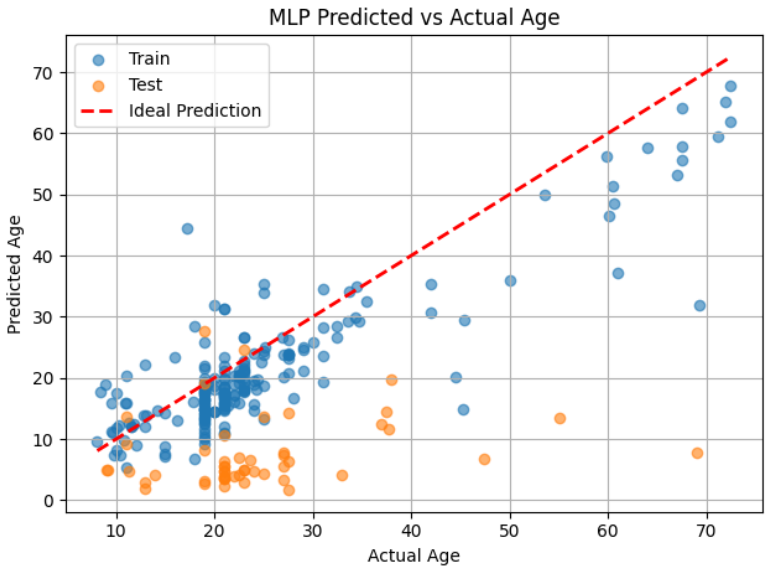


Figure 6: Predicted versus actual age where the ideal prediction matches well with the training data in blue, but fails for the test data in orange.

4 Discussion

Overall, the proposed method did not meet our expectations for improving brain age prediction from MRI data in the presence of site-related bias. We hypothesized that the variational autoencoder (VAE) would be able to capture biologically meaningful latent representations of brain structure, enabling the Transformer model to make more accurate predictions than conventional approaches. Furthermore, we anticipated that applying ComBat harmonization to these latent features would effectively mitigate site bias and further enhance prediction accuracy.

However, our results suggest that the application of ComBat may have inadvertently removed critical variance necessary for accurate age prediction. The harmonized features appeared to lack the dynamic range required for the Transformer to generalize well across the full age spectrum, leading to degraded performance compared to models trained on unharmonized VBM features.

Future work should investigate which specific components of the latent space are altered or suppressed by ComBat. This would help clarify whether harmonization methods can be adapted to preserve age-relevant biological signals while still correcting for site effects. Additionally, alternative strategies for site harmonization—such as domain adaptation, adversarial training, or feature disentanglement—could be explored to better balance bias removal with information retention.

Compared to existing approaches based on convolutional neural networks (CNNs) or other deep architectures, our method demonstrated limited effectiveness in both site bias correction and predictive performance. Nonetheless, the integration of sequence-based models like Transformers with representation learning remains a promising direction, particularly if future improvements can address the sensitivity to harmonization techniques.

References

Baecker, L., Garcia-Dias, R., Vieira, S., Scarpazza, C., and Mechelli, A. (2021). Machine learning for brain age prediction: Introduction to methods and clinical applications. *EBioMedicine*, 72.

Dufumier, B., Grigis, A., Victor, J., Ambroise, C., Frouin, V., and Duchesnay, E. (2022a). Age prediction with site-effect removal: A challenge on the openbhb dataset. `https://ramp.studio/problems/brain_age_with_site_removal`. Accessed: 2025-05-01.

Dufumier, B., Grigis, A., Victor, J., Ambroise, C., Frouin, V., and Duchesnay, E. (2022b). Openbhb: a large-scale multi-site brain mri data-set for age prediction and debiasing. *NeuroImage*, 263:119637.

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., McInnis, M., Phillips, M. L., Trivedi, M. H., Weissman, M. M., and Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120.

Fortin, J.-P., Parker, D., Tunç, B., et al. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, 161:149–170.

Gadewar, S. P., Ramesh, A., Liu, M., Gari, I. B., Nir, T. M., Thompson, P., and Jahanshad, N. (2023). Predicting individual brain mris at any age using style encoding generative adversarial networks. In *18th International Symposium on Medical Information Processing and Analysis*, volume 12567, pages 471–479. SPIE.

Jia, W., Li, H., Ali, R., Shanbhogue, K. P., Masch, W. R., Aslam, A., Harris, D. T., Reeder, S. B., Dillman, J. R., and He, L. (2024). Investigation of combat harmonization on radiomic and deep features from multi-center abdominal mri data. *Journal of Imaging Informatics in Medicine*.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.

Jónsson, B. A., Bjornsdottir, G., Thorgeirsson, T. E., Ellingsen, L. M., Walters, G. B., Gudbjartsson, D. F., Stefansson, H., Stefansson, K., and Ulfarsson, M. O. (2019). Brain age prediction using deep learning uncovers associated sequence variants. *Nature communications*, 10(1):5409.

Kingma, D. P., Welling, M., et al. (2013). Auto-encoding variational bayes.

Pachetti, E., Colantonio, S., and Pascali, M. A. (2022). On the effectiveness of 3d vision transformers for the prediction of prostate cancer aggressiveness. In *International Conference on Image Analysis and Processing*, pages 317–328. Springer.

Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A., and Smith, S. M. (2021). Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis*, 68:101871.

Redekop, E., Pleasure, M., Wang, Z., Sarma, K. V., Kinnaird, A., Speier, W., and Arnold, C. W. (2024). Codebook vq-vae approach for prostate cancer diagnosis using multiparametric mri. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2365–2372.

Redekop, E., Sarma, K. V., Kinnaird, A., Sisk, A., Raman, S. S., Marks, L. S., Speier, W., and Arnold, C. W. (2022). Attention-guided prostate lesion localization and grade group classification with multiple instance learning. In *International conference on medical imaging with deep learning*, pages 975–987. PMLR.

Usman, M., Rehman, A., Shahid, A., Rehman, A. U., Gho, S.-M., Lee, A., Khan, T. M., and Razzak, I. (2024). Multi-task adversarial variational autoencoder for estimating biological brain age with multimodal neuroimaging. *arXiv preprint arXiv:2411.10100*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Yook, S., Miao, Y., Park, C., Park, H. R., Kim, J., Lim, D. C., Joo, E. Y., and Kim, H. (2021). Predicting brain age based on sleep eeg and densenet. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 245–248. IEEE.

Zhang, X., Duan, S.-Y., Wang, S.-Q., Chen, Y.-W., Lai, S.-X., Zou, J.-S., Cheng, Y., Guan, J.-T., Wu, R.-H., and Zhang, X.-L. (2024). A resnet mini architecture for brain age prediction. *Scientific reports*, 14(1):11185.