Semi-Supervised Diffusion Model for Brain Age Prediction

Ayodeji Ijishakin[1]*, **Sophie Martin**[1], **Florence Townend**[1], **Federica Agosta**[2], **Edoardo Gioele Spinelli**[2], **Silvia Basaia**[2], **Paride Schito**[2], **Yuri Falzone**[2], **Massimo Filippi**[2], **James Cole**[1], and **Andrea Malaspina**[1]

[1]University College London
[2]San Raffelle Scientific Institute

Abstract

Brain age prediction models have succeeded in predicting clinical outcomes in neurodegenerative diseases, but can struggle with tasks involving faster progressing diseases and low quality data. To enhance their performance, we employ a semi-supervised diffusion model, obtaining a 0.83(p<0.01) correlation between chronological and predicted age on low quality T1w MR images. This was competitive with state-of-the-art non-generative methods. Furthermore, the predictions produced by our model were significantly associated with survival length (r=0.24, p<0.05) in Amyotrophic Lateral Sclerosis. Thus, our approach demonstrates the value of diffusion-based architectures for the task of brain age prediction.

1 Introduction

Brain age prediction refers to the task of predicting an individual's age using neuroimaging data [1]. As individuals age, various brain structures and functions degrade; which lead to negative health outcomes [2]–[7]. These changes do not occur at the same rate across people, such that individuals may have the same chronological age but different brain ages [8]. This disparity is particularly present in neurodegenerative diseases such as: Alzheimer's Disease (AD) and Multiple Sclerosis (MS). In such diseases, an older appearing brain has been associated with: disease status, higher risk of mortality, the speed of disease progression, and disease related genetic markers [9]–[12].

However, the brain age paradigm has seen limited application to amyotrophic lateral sclerosis (ALS) [13]. ALS, is one of the rarest and fastest progressing neurodegenerative diseases; with a median survival time (time from diagnosis to death) of only 3 years [14]. With such a short time-window for the most significant ageing related brain changes to occur, effective brain age models for ALS need to be particularly sensitive to ageing related neuroanatomical changes. One study demonstrated that individuals with cognitive and behavioural symptoms but not motor impairment have an association with a higher brain age; this further outlines the subtlety of the relationship between neuroanatomical ageing and ALS [13]. Additionally, brain age models are often trained on high quality (research grade) data, which reduces their wider applicability in real-world settings where the quality of data is lower (clinical grade); this is particularly true for rare diseases like ALS.

Generative models may offer a solution, as they have demonstrated great capabilities in capturing latent factors within an image (e.g., ageing) and produce representations which are robust to data quality [15]–[22]. In particular, diffusion models have demonstrated image synthesis capabilities well beyond previous approaches [23]–[25].

In this work, we present a novel brain age prediction architecture which leverages the representation learning capabilities of diffusion autoencoders [17]. Our approach produces representations which

---

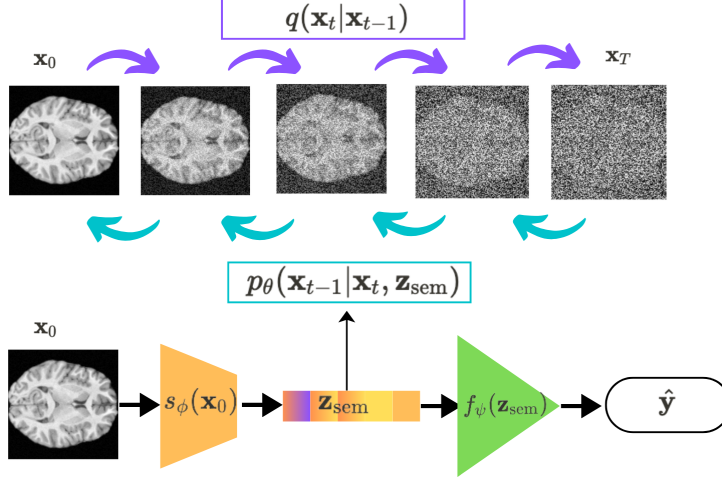*Direct to all correspondence to: ayodeji.ijishakin.21@ucl.ac.uk

Figure 1: The Model architecture. First the image, $\mathbf{x}_0$ goes through the forward process $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$, to produce $\mathbf{x}_T$. The reverse process, $p_\theta(\mathbf{x}_{0:T}|\mathbf{z}_{\text{sem}})$ then reconstructs the image conditioned on the semantic latent, produced by $s_\phi(\mathbf{x}_0)$. Finally, the age predictor, $f_\psi(\mathbf{z}_{\text{sem}})$ predicts an individual's age, $\hat{\mathbf{y}}$.

are robust to data quality, and conditioned on the neuroanatomical effects of ageing. This results in predictions which are accurate on clinical-grade data and associated with mortality in ALS disease at a level which surpasses non-generative approaches.

2 Background

2.1 Diffusion Models

Diffusion models allow us to approximate a data distribution, $q(\mathbf{x}_0)$, with a learnt distribution, $p_\theta(\mathbf{x}_0)$, via a latent variable model which takes the form:

$$p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_{0:T})d\mathbf{x}_{1:T}, \quad \text{where} \quad p_\theta(\mathbf{x}_{0:T}) := p_\theta(\mathbf{x}_T)\prod_{t=1}^{T}p_\theta^{(t)}(\mathbf{x}_{t-1}|\mathbf{x}_t) \tag{1}$$

Where the latents $\mathbf{x}_1, ..., \mathbf{x}_T$ share the same dimensions as our training data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ [15], [16], [26]. As denoted above, the transitions between latents are defined as a Markov chain beginning at a prior $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ and ending at our model distribution $p_\theta(\mathbf{x}_0)$. We call this joint distribution, $p_\theta(\mathbf{x}_{0:T})$, the *reverse process*. In contrast to most latent variable models, the approximate posterior $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ is predefined as another Markov chain, that we call the *forward process*. This process destroys the structure of our data, $\mathbf{x}_0$, via the gradual addition of Gaussian noise, defined as:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T}q(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad \text{where} \quad q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{2}$$

Where $\beta_1, ..., \beta_T$ are scalars which enforce a variance schedule. The reverse process, $p_\theta(\mathbf{x}_{0:T})$, denoises our image following the forward process, $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$, to recover our data distribution. Our model, $p_\theta(\mathbf{x}_0)$, may be trained by minimising a variational lower bound on negative log likelihood:

$$\mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] \leq \mathbb{E}_q\left[-\log\frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] = \mathbb{E}_q\left[-\log p_\theta(\mathbf{x}_{0:T}) - \log q(\mathbf{x}_{1:T}|\mathbf{x}_0)\right] \tag{3}$$

At time t, the approximate posterior is another Gaussian, $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I})$, where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$. Thus, the noised image at time t can be expressed as $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\epsilon$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This allows us to simplify our objective, by learning a model, $\epsilon_\theta^{(t)}(\mathbf{x}_t)$, which predicts the sampled noise, ϵ, based on $\mathbf{x}_t$ and t. This holds provided that our noised image at time t

Table 1: Results of brain age prediction. The first three rows are different metrics demonstrating the accuracy of the brain age predictions on an external dataset. The final row is the relationship (again Pearson's r) between the brain-PADs and survival in months (time from scan to death) on ALS patients. Significant relationships are displayed in **bold**.

	OUR MODEL	DENSENET 121[27]	BRAINAGER[28]	DEEPBRAINNET[29]
TEST R	**0.83(P<0.001)**	**0.83(P<0.001)**	**0.46(P<0.001)**	**0.81(P<0.001)**
TEST MAE	5 YEARS	8.63 YEARS	11.7 YEARS	5.55 YEARS
TEST R^2	0.65	0.12	-1.14	0.56
ALS SURVIVAL	**0.24(P<0.05)**	-0.05 (P=0.61)	0.17 (P=0.14)	-0.12 (P=0.28)

Table 2: Results of Two-Sample Kolmogorov-Smirnov tests for significance between the PADs produced by our model and those produced by other models. Significant results are displayed in **bold**.

	OUR MODEL	DENSENET 121[27]	BRAINAGER[28]	DEEPBRAINNET[29]
MEAN	-6.19	-1.60	-19.17	-1.13
STD	27.29	7.74	10.97	8.06
KS STATISTIC		**0.32(P<0.05)**	**0.47(P<0.001)**	**0.31(P<0.05)**

is assumed to be Gaussian with a fixed variance according to α_t and a learnable mean admitting the following loss (see [16] for the full derivation):

$$L := \sum_{t=1}^{T} \mathbb{E}_{\mathbf{x}_0 \sim q(x_0), \epsilon_t \sim \mathcal{N}(\mathbf{0},\mathbf{I})} \left[\|\epsilon_\theta^{(t)}(\sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\epsilon_t) - \epsilon_t\|_2^2 \right] \tag{4}$$

2.2 Diffusion Autoencoders

Diffusion autoencoders modify diffusion models by conditioning the reverse process on a latent, $\mathbf{z}_{\text{sem}} \in \mathbb{R}^d$ where $\mathbf{x}_0 \in \mathbb{R}^D$ and $d \ll D$. This latent is output by a semantic encoder, $s_\phi(\mathbf{x}_0) : \mathbb{R}^D \to \mathbb{R}^d$, that is distinct from our noise predictor, $\epsilon_\theta^t(\mathbf{x}_t)$. The modified reverse process takes the form:

$$p(\mathbf{x}_{0:T}|\mathbf{z}_{\text{sem}}) = p(\mathbf{x}_T) \prod_{t=1}^{t=T} p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{z}_{\text{sem}}) \tag{5}$$

By conditioning the reverse process on a semantic latent, diffusion autoencoders force the stochasticity contained within the image to be modelled by our forward process, $q(\mathbf{x}_t|\mathbf{x}_0)$ [17]. Whilst the semantically meaningful information is forced into our latent $\mathbf{z}_{\text{sem}}$, this leads to rich representations which allow for accurate predictions in the present context.

3 Method

Our model fuses a diffusion autoencoder with an age predictor network, $f_\psi(\mathbf{z}_{\text{sem}}) : \mathbb{R}^d \to \mathbb{R}^+$ which predicts an individual's age based on the latent representation, $\mathbf{z}_{\text{sem}}$ of their brain image, $\mathbf{x}_0$. The model takes as input a batch of images sampled from a dataset $q(\mathcal{X}) = \{\mathbf{x}_{0,i}\}_{i=1}^{N}$ where $\mathbf{x}_{0,i} \in \mathbb{R}^D$. The images are accompanied by age labels, $\mathcal{Y} = \{\mathbf{y}_i\}$, where $\mathbf{y}_i \in \mathbb{R}^+$. First the image, $\mathbf{x}_0$, goes through the forward process, $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$, to produce $\mathbf{x}_T$. The reverse process, $p_\theta(\mathbf{x}_{0:T}|\mathbf{z}_{\text{sem}})$, then reconstructs the image conditioned on the semantic latent, $\mathbf{z}_{\text{sem}}$, produced by $s_\phi(\mathbf{x}_0)$. Finally, the age predictor, $f_\psi(\mathbf{z}_{\text{sem}})$ predicts an individual's age; see Figure 1 for visual intuition.

Our objective is a sum of two MSE losses. The first loss is modified version of Equation 4. With the addition that the noise predictor ϵ_θ also takes as input $\mathbf{z}_{\text{sem}}$, at every time step. Whilst the second loss is the standard MSE between predicted age and actual age:

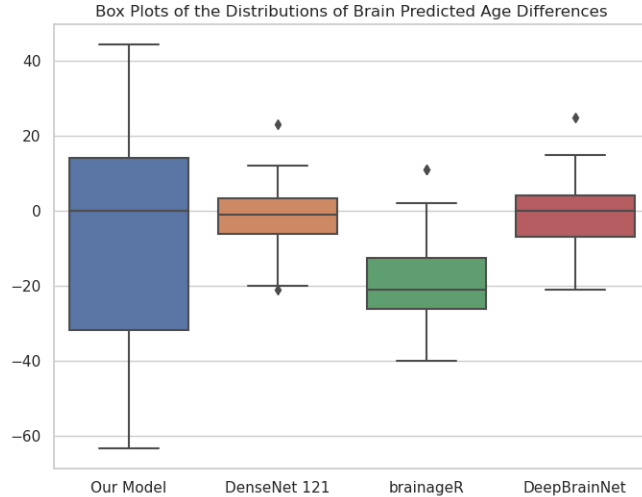$$L_{AGE} = \frac{1}{N} \sum_{i=0}^{N} (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2 \tag{6}$$

3

Figure 2: A Boxplot of the distribution of brain-PADs produced by different models on our ALS dataset. Our model produces more variation within the brain-PADs of the ALS patients. Two-Sample Kolmogrov Smirnov tests (Table 2) demonstrated that our brain-PADs were significantly different from brain-PADs from the other three approaches .

Thus, the model receives direct supervision from the labels learnt by $f_\psi(\mathbf{z}_{\text{sem}})$ and receives self-supervision via its generative arm, learnt by $p_\theta(\mathbf{x}_{0:T}|\mathbf{z}_{\text{sem}})$ making for a semi-supervised diffusion model. The model is also semi-supervised in the classic sense of the term as unlabelled are used when training to help condition our latent space. In those instances the image is run through the model in the usual fashion without producing an age prediction $f_\psi(\mathbf{z}_{\text{sem}})$.

4 Results

We trained our model on 4631 2D T1w Magnetic Resonance Imaging (MRI) scans from 8 publicly available datasets (mean age = 56 years, std = 20.9 years). For more details concerning the datasets and pre-processing see Appendix B; for details concerning the network design and hardware see Appendix D. Our test set comprised 473 2D T1w MR images (mean age = 68 years, std = 10 years), 100 % of which were clinical grade.

We achieved a test accuracy 0.83(p<0.01) measured as the Pearson's r correlation coefficient between predicted and chronological age (MAE = 5 years, $R^2 = 0.65$). Table 1 displays our performance compared to state-of-the-art brain age prediction methods, which demonstrates that we are competitive with these approaches. The final row of Table 1 shows the correlation between the brain predicted age difference (measured as predicted age minus chronological age) and survival in months in our ALS patient group. This quantity is the biomarker typically used to characterise accelerated ageing [1], [2].

The brain predicted age differences (brain-PADs) produced by our model are significantly correlated with survival in months for 72 patients with ALS disease (mean age = 61 years, std = 10.9 years). The other standard approaches did not produce brain-PADs which were significantly associated with this outcome. Table 2 shows that the brain-PADs produced by our method were significantly different from the rest of the techniques; see Figure 2 for visual intuition.

5 Discussion

In this work, we present a novel diffusion based brain age prediction architecture. To our knowledge, this is the first application of a class of diffusion model to this task. Our results demonstrate that our approach is competitive with state-of-the-art brain age prediction models on a dataset completely

composed of clinical-grade data. It exhibited a modest improvement over existing techniques, achieving the joint highest Pearson correlation coefficient (r = 0.83, p < 0.01), the highest coefficient of determination ($R^2 = 0.65$), and the lowest mean absolute error (MAE) of 5 years.

Most importantly, the ALS brain-PADs produced by our method also capture more variance than non-generative approaches in a statistically significant way; as displayed in the box plots shown in Figure 2. This variation in brain-PADs was also significantly associated with survival time in ALS, unlike non-generative approaches. This demonstrates that our method is sensitive to the subtle relationship between ALS derived neuroanatomical changes and the ageing process. We posit that our improved performance can be attributed to the capacity of diffusion based architectures, to produce representations which are well conditioned on latent factors within an image, (e.g., the neuroanatomical effects of ageing) and robust to data quality.

6 Future Work and Limitations

Although the association between the ALS brain-PADs and survival was significant, the results would be better if the effect size were larger. However, the small sample size of ALS patients ($n = 72$) certainly contributed to this. It is also interesting to note that the mean of the brain-PADs was negative (-6.19 years), which is the opposite direction of what is expected. Future work should use larger cohorts of ALS patients to see if the effect size of the association between PADs and survival increases; as well as whether the brain-PADs remain mainly negative. The model may also be augmented by including a separate ALS survival objective; but this of course requires more patient data. It may also be extended to process 3D data and should be tested on other diseases such as: AD and MS. We leave this to future work, and treat the present study as a preliminary exposition of our approach to brain age prediction.

7 Related Work

Choi et al. [30] used a variational autoencoder to model age conditioned brain topography using positron emission tomography data. The authors found that as they aged individuals with the model, their predicted regional metabolic changes, were associated with their actual follow-up changes. However, this approach was not able to make age predictions as age was required in-order to use the model.

Mouches et al. [31] used an age conditioned autoencoder coupled with a brain age regressor, to produce a model which can produce synthetic brain images of a particular age and perform brain age predictions.

Zhao et al. [32] similarly combined an age regressor with a variational autoencoder, to model age conditioned latent's whilst producing brain age predictions. Although both approaches also used MRI data, they were not applied to a patient dataset, so the significance of their approaches for disease groups is unknown. Additionally they were trained solely on research-grade data so are unlikely to generalise to clinical-grade scans.

8 Acknowledgements

References

[1] L. Baecker, R. Garcia-Dias, S. Vieira, C. Scarpazza, and A. Mechelli, "Machine learning for brain age prediction: Introduction to methods and clinical applications," *eBioMedicine*, vol. 72, p. 103 600, Oct. 2021, ISSN: 2352-3964. DOI: 10.1016/J.EBIOM.2021.103600.

[2] J. H. Cole, R. E. Marioni, S. E. Harris, and I. J. Deary, "Brain age and other bodily 'ages': Implications for neuropsychiatry," *Molecular psychiatry*, vol. 24, no. 2, pp. 266–281, 2019.

[3] H. M. Aycheh, J.-K. Seong, J.-H. Shin, *et al.*, "Biological brain age prediction using cortical thickness data: A large scale cohort study," *Frontiers in aging neuroscience*, vol. 10, p. 252, 2018.

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

[5] J. H. Cole, R. Leech, D. J. Sharp, and A. D. N. Initiative, "Prediction of brain age suggests accelerated atrophy after traumatic brain injury," *Annals of neurology*, vol. 77, no. 4, pp. 571–581, 2015.

[6] A.-M. G. De Lange, M. Anatürk, S. Suri, *et al.*, "Multimodal brain-age prediction and cardio-vascular risk: The whitehall ii mri sub-study," *NeuroImage*, vol. 222, p. 117 292, 2020.

[7] M. A. Eckert, "Slowing down: Age-related neurobiological predictors of processing speed," *Frontiers in neuroscience*, vol. 5, p. 25, 2011.

[8] J. H. Cole, S. J. Ritchie, M. E. Bastin, *et al.*, "Brain age predicts mortality," *Molecular psychiatry*, vol. 23, no. 5, pp. 1385–1392, 2018.

[9] M. Ly, G. Z. Yu, H. T. Karim, *et al.*, "Improving brain age prediction models: incorporation of amyloid status in Alzheimer's disease," *Neurobiology of Aging*, vol. 87, pp. 44–48, Mar. 2020, ISSN: 0197-4580. DOI: 10.1016/J.NEUROBIOLAGING.2019.11.005.

[10] J. H. Cole PhD, J. Raffel MD, T. Friede PhD, *et al.*, "Longitudinal Assessment of Multiple Sclerosis with the Brain-Age Paradigm," *Annals of Neurology*, vol. 88, no. 1, pp. 93–105, Jul. 2020, ISSN: 1531-8249. DOI: 10.1002/ANA.25746. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/ana.25746https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.25746https://onlinelibrary.wiley.com/doi/10.1002/ana.25746.

[11] L. C. Löwe, C. Gaser, and K. Franke, "The Effect of the APOE Genotype on Individual BrainAGE in Normal Aging, Mild Cognitive Impairment, and Alzheimer's Disease," *PLOS ONE*, vol. 11, no. 7, e0157514, Jul. 2016, ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0157514. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0157514.

[12] 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, e01020, Aug. 2018, ISSN: 2162-3279. DOI: 10.1002/BRB3.1020. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/brb3.1020https://onlinelibrary.wiley.com/doi/abs/10.1002/brb3.1020https://onlinelibrary.wiley.com/doi/10.1002/brb3.1020.

[13] A. Hermann, G. N. Tarakdjian, A. G. M. Temp, *et al.*, "Cognitive and behavioural but not motor impairment increases brain age in amyotrophic lateral sclerosis," *Brain Communications*, vol. 4, no. 5, fcac239, Sep. 2022, ISSN: 2632-1297. DOI: 10.1093/braincomms/fcac239. eprint: https://academic.oup.com/braincomms/article-pdf/4/5/fcac239/46472826/fcac239.pdf. [Online]. Available: https://doi.org/10.1093/braincomms/fcac239.

[14] B Swinnen and W Robberecht, "The phenotypic variability of amyotrophic lateral sclerosis," *NATURE REVIEWS | NEUROLOGY*, vol. 10, pp. 661–670, 2014. DOI: 10.1038/nrneurol.2014.184. [Online]. Available: www.nature.com/nrneurol.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *ArXiv*, vol. abs/2006.11239, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219955663.

[16] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[17] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 609–10 619, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244729224.

[18] C. Luo, "Understanding diffusion models: A unified perspective," *ArXiv*, vol. abs/2208.11970, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251799923.

[19] J. M. Tomczak, "Deep generative modeling," *Deep Generative Modeling*, 2022.

[20] S. Mo, Z. Sun, and C. Li, "Representation disentanglement in generative models with contrastive learning," *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1531–1540, 2023.

[21] I. Higgins, L. Matthey, A. Pal, *et al.*, "Beta-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2016.

[22] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[23] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *ArXiv*, vol. abs/2204.06125, 2022. [Online]. Available: `https://api.semanticscholar.org/CorpusID:248097655`.

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 674–10 685, 2021. [Online]. Available: `https://api.semanticscholar.org/CorpusID:245335280`.

[25] "Midjourney," [Online]. Available: `https://www.midjourney.com/`.

[26] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*, PMLR, 2015, pp. 2256–2265.

[27] D. A. Wood, S. Kafiabadi, A. Al Busaidi, *et al.*, "Accurate brain-age models for routine clinical mri examinations," *Neuroimage*, vol. 249, p. 118 871, 2022.

[28] F. Biondo, A. Jewell, M. Pritchard, *et al.*, "Brain-age is associated with progression to dementia in memory clinic patients," *NeuroImage: Clinical*, vol. 36, p. 103 175, 2022.

[29] V. M. Bashyam, G. Erus, J. Doshi, *et al.*, "Mri signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide," *Brain*, vol. 143, no. 7, pp. 2312–2324, 2020.

[30] H. Choi, H. Kang, D. S. Lee, and A. D. N. Initiative, "Predicting aging of brain metabolic topography using variational autoencoder," *Frontiers in aging neuroscience*, vol. 10, p. 212, 2018.

[31] P. Mouches, M. Wilms, D. Rajashekar, S. Langner, and N. Forkert, "Unifying brain age prediction and age-conditioned template generation with a deterministic autoencoder," in *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, M. Heinrich, Q. Dou, M. de Bruijne, J. Lellmann, A. Schläfer, and F. Ernst, Eds., ser. Proceedings of Machine Learning Research, vol. 143, PMLR, 2021, pp. 497–506. [Online]. Available: `https://proceedings.mlr.press/v143/mouches21a.html`.

[32] Q. Zhao, E. Adeli, N. Honnorat, T. Leng, and K. M. Pohl, "Variational autoencoder for regression: Application to brain aging analysis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, Springer, 2019, pp. 823–831.

[33] B. B. Avants, N. Tustison, and H. J. Johnson, "Advanced normalization tools (ants)," 2020.

[34] R. Beare, B. C. Lowekamp, and Z. R. Yaniv, "Image segmentation, registration and characterization in r with simpleitk.," *Journal of statistical software*, vol. 86, 2018.

[35] M Schell, I Tursunova, I Fabian, *et al.*, "Automated brain extraction of multi-sequence mri using artificial neural networks," European Congress of Radiology-ECR 2019, 2019.

[36] W. A. Falcon, "Pytorch lightning," *GitHub*, vol. 3, 2019.

Checklist

1. For all authors...

    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

    (b) Did you describe the limitations of your work? [Yes] see 6.

    (c) Did you discuss any potential negative societal impacts of your work? [Yes] see A.

    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes].

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?[No] This is because although alot of the data is available from public datasets, individuals must apply to the relevant consortiums in order to have access. We have provided much detail concerning the datasets in Appendix B, such that the relevant applications can be made if someone wishes. We have however included code that can be used to train a model, (see the Abstract). This can then be used to make predictions if the data is acquired.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] see Appendix D
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider) [Yes] see Appendix D

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets?[N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A   Potential Negative Societal Impact

Brain age prediction is primarily used in order to make inferences about cognitive and neuroanatomical decline. Although these predictions may aid in: clinical trial design, disease prognostication and classification. One could imagine brain age prediction being used by bad actors to discriminate against an individual based on their brain age. An example would be an insurance company who would offer higher premiums for individuals with older looking brains. With that being said, we anticipate that the clinical utility of this tool outweighs the likelihood of such events occurring.

B   Dataset and Pre-processing

We curated one dataset for the task of standard brain age prediction and a separate dataset for ALS survival analyses. Our brain age prediction dataset consisted of 4621 3D structural T1-weighted MR images for training (mean age=56 years, std=20.9 years) and 473 3D structural T1-weighted MR images for testing (mean age=68 years, std=10 years). All individuals included in training were considered to be radiologically normal (of non-disease status). Our data were drawn from 8 publicly available datasets. These were: the Open Access Series of Imaging Studies-1, the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL), the Southwest University Adult Lifespan Dataset, the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, the Dallas Lifespan Brain Study, the Nathan Kline Insti- tute Rocklands Sample, the National Alzheimer's Coordinating Center and the Cambridge Centre for Ageing and Neuroscience study (Cam-CAN). Our ALS dataset comprised 72 3D T1w MR images (mean age = 61 years, std=10.9 years) from the San Rafaele Hospital in Milan. All images underwent the following pre-processing. The ANTS package was used to conduct affine registrations of the images to the MNI 152 brain template, they were then resampled to $130 \times 130 \times 130$ resolution, Simple ITK was used to perform n4 bias field correcttion and HD-BET was used to skull strip the images [33]–[35]. 2D medial axial slices were then taken from the 3D volumes, which had their pixel values normalised to be between 0 and 1, after being resized to 128×128 resolution.

C   Network Design and Hardware

All model hyperparameters and architectural details were chosen heuristically. Our noise predictor network $\epsilon_\theta^{(t)}$ was a U-Net with the following design. The first convolutional layer expanded the input from a single channel to 32. The tensor stayed at this channel depth for another three convolutional layers. Following this, channel multipliers (2, 3, 4) were applied every 3 convolutional layers making for a channel expansion of $(32 \to 32 \to 32) \to (32 \to 32 \to 32) \to (64 \to 64 \to 64) \to (96 \to 96 \to 96) \to (128 \to 128 \to 128)$. Three attention blocks then followed, which received a flattened version of the output; this formed the middle block of the U-Net. The upward path of the U-Net was simply the exact reverse of the downward path with regard to channel expansion. Where the channels were equal across the two U-Net paths, a residual connection sent information across. Group normalisation followed each convolutional layer as well as the SiLU activation function. Our semantic encoder network, $\mathbf{s}_\phi(\mathbf{x}_0)$ was the downward path of the U-Net and the middle block. The semantic encoder, output latent representations, $\mathbf{z}_{\text{sem}}$ of dimension 512. The age predictor network, $\mathbf{f}_\psi(\mathbf{z}_{\text{sem}})$, was an MLP which consisted of two linear layers of size: $(512 \to 128 \to 32)$ following these layers a final layer mapped to the age. Between each layer was a ReLU, batch normalisation and 0.5 dropout. There were 10.12 million parameters in total. All training was performed on an Nvidia GeForce RTX 4090 graphics card, with the Adam optimizer using PyTorch lightning [36] all model components $\epsilon_\theta^{(t)}(\mathbf{z}_{\text{sem}})$, $s_\phi(\mathbf{x}_0)$ and $f_\psi(\mathbf{z}_{\text{sem}})$ were trained simultaneously.

D   Network Design and Hardware

All model hyperparameters and architectural details were chosen heuristically. Our noise predictor network $\epsilon_\theta^{(t)}(\mathbf{x}_0)$ was a U-Net with the following design. The first convolutional layer expanded the input from a single channel to 32. The tensor stayed at this channel depth for another three

convolutional layers. Following this, channel multipliers (2, 3, 4) were applied every 3 convolutional layers making for a channel expansion of $(32 \to 32 \to 32) \to (32 \to 32 \to 32) \to (64 \to 64 \to 64) \to (96 \to 96 \to 96) \to (128 \to 128 \to 128)$. Three attention blocks then followed, which received a flattened version of the output; this formed the middle block of the U-Net. The upward path of the U-Net was simply the exact reverse of the downward path with regard to channel expansion. Where the channels were equal across the two U-Net paths, a residual connection sent information across. Group normalisation followed each convolutional layer as well as the SiLU activation function. Our semantic encoder network, $\mathbf{s}_\phi(\mathbf{x}_0)$ was the downward path of the U-Net and the middle block. The semantic encoder, output latents, $\mathbf{z}_{\text{sem}}$ of dimension 512. The age predictor network, $\mathbf{f}_\psi(\mathbf{z}_{\text{sem}})$, was an MLP which consisted of two linear layers of size: $(512 \to 128 \to 32)$ following these layers a final layer mapped to the age. Between each layer was a ReLU, batch normalisation and 0.5 dropout. There were 10.12 million parameters in total. All training was performed on an Nvidia GeForce RTX 4090 graphics card, with the Adam optimizer using PyTorch lightning [36].
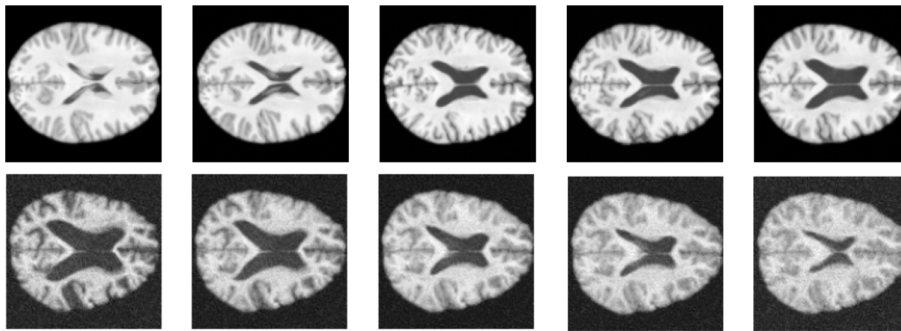
E   Additional Results

Figure 3: Age Interpolations. In the first row, the image in the first column is that of a 25-year-old, we then linearly interpolate between them and a 75-year-old (final column). The second row is the same but reversed, from age 92 to age 20. The difference in data quality, across the two rows, exemplifies the robustness of the model to varied data quality.
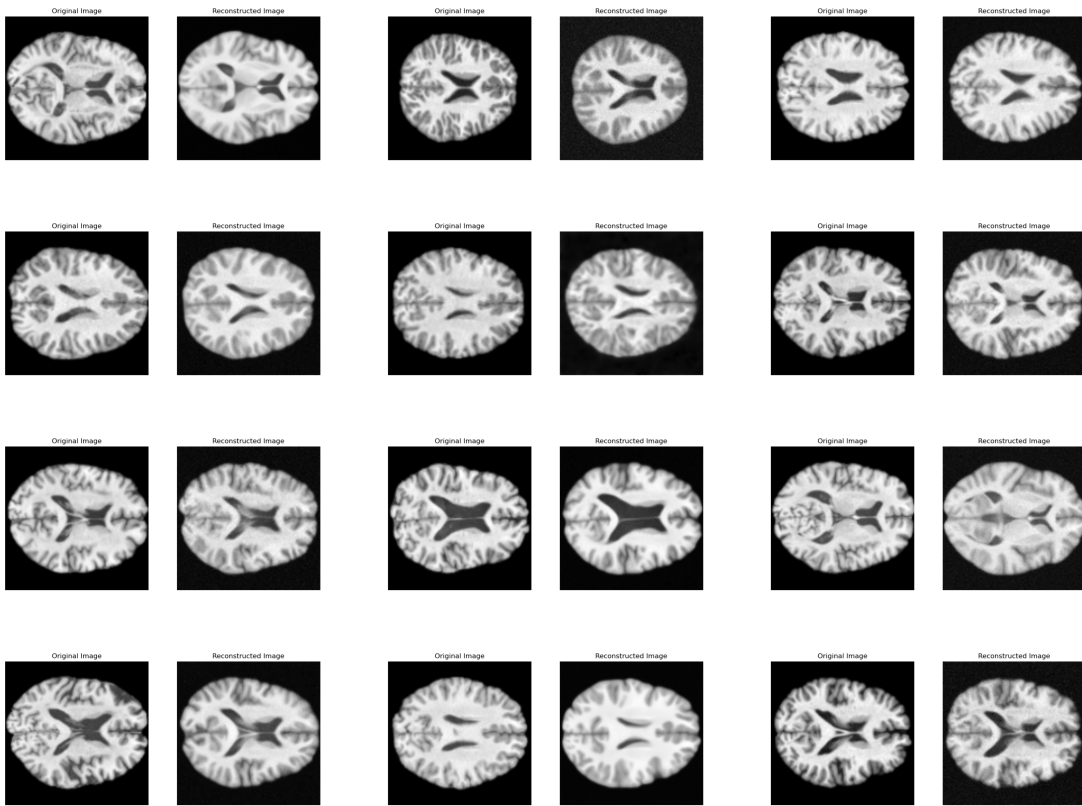
Figure 4: Image Reconstructions. The images on the left columns are the original image, and the images on the right columns are their reconstructed version.