

FORECASTING FUTURE ANATOMIES: LONGITUDINAL BRAIN MRI-TO-MRI PREDICTION

Ali Farki, Elaheh Moradi, Deepika Koundal, Jussi Tohka

A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio, Finland

ABSTRACT

Predicting future brain state from a baseline magnetic resonance image (MRI) is a central challenge in neuroimaging, and has important implications for studying neurodegenerative diseases such as Alzheimer’s disease (AD). Most existing approaches predict future cognitive scores or clinical outcomes, such as conversion from mild cognitive impairment to dementia. Instead, here we investigate longitudinal MRI image-to-image prediction that forecasts a participant’s entire brain MRI several years into the future, intrinsically modeling complex, spatially distributed neurodegenerative patterns. We implement and evaluate five deep learning architectures (UNet, U²-Net, UNETR, Time-Embedding UNet, and ODE-UNet) on two longitudinal cohorts (ADNI and AIBL). Predicted follow-up MRIs are directly compared with the actual follow-up scans using metrics that capture global similarity and local differences. The best performing models achieve high-fidelity predictions, and all models generalize well to an independent external dataset, demonstrating robust cross-cohort performance. Our results indicate that deep learning can reliably predict participant-specific brain MRI at the voxel level, offering new opportunities for individualized prognosis.

Index Terms— Longitudinal MRI prediction, deep learning, Alzheimer’s disease, brain magnetic resonance imaging, image-to-image prediction

1. INTRODUCTION

Understanding how the human brain changes over time is a central goal in neuroimaging research. Longitudinal magnetic resonance imaging (MRI) provides rich insights into brain development, aging, and structural alterations associated with neurological and psychiatric conditions. In addition, predicting future MRI patterns is valuable because anticipated brain changes can serve as early indicators of disease progression or cognitive decline, offering a data-driven pathway toward forecasting future clinical outcomes. However, longitudinal datasets are often limited in temporal coverage and sample size, constraining our ability to model individual trajectories of brain change. To address these challenges, recent research has begun exploring computational approaches that can generate future brain MRI scans, effectively trying to predict how an individual’s brain structure will evolve over time [1, 2].

Accurate prediction of future brain morphology could open new avenues for studying brain aging and disease progression, improving patient-level prognosis.

Traditionally, machine learning in neuroimaging has focused on predicting diagnostic [3] or cognitive outcomes from imaging features [4, 5]. Although valuable for clinical applications, these approaches do not offer predictions on how brain anatomy itself changes over time. In contrast, image-to-image prediction seeks to model the full spatial evolution of brain structure, learning how an individual’s MRI transforms over time. This task requires capturing both spatial anatomy and temporal dynamics, making it a challenging and highly informative problem in computational neuroimaging.

A rapidly expanding body of research has explored deep learning approaches for medical image-to-image translation. In neuroimaging, this work has focused primarily on cross-modality synthesis, for example, generating a pseudo-FDG positron emission tomography (PET) image from an MRI or vice versa [6, 7]. Many MRI synthesis methods using deep learning translate between MRI pulse sequences [8], most commonly T1w→T2w/FLAIR, showing that one sequence can often predict another’s contrast. However, they do not address longitudinal prediction within the same modality. Another related line of research involves MRI-to-MRI image synthesis, in which generative models simulate brain scans under different conditions, such as aging or disease effects, but do not predict the actual future MRI of each individual. Ravi et al. introduced degenerative adversarial neuroimage nets to produce aged brain images constrained by disease patterns [2], and Xia et al. (2021) proposed an autoencoder-based method to generate aged MRIs from a single baseline scan [1]. Although these models generate visually convincing progressions, they typically capture average or prototypical changes rather than individualized, subject-specific evolution. More recently, Pombo et al. (2023) combined conditional generative models with diffeomorphic transformations to create “counterfactual” MRIs under altered conditions [9, 10]; however, such approaches are not trained to reproduce each participant’s true future brain image.

In this study, we investigate longitudinal MRI-to-MRI prediction, enabling voxel-level modeling of individual brain changes over time. Unlike prior work focused on regional measures [11] or population-average trajectories[1], our approach predicts the entire future MRI of each individual, capturing the

complex spatial pattern of anatomical evolution. In particular, we implement and compare five deep learning architectures within a unified framework, evaluate their predictions against follow-up scans using both global and regional metrics, and test their generalizability on an independent cohort.

2. METHODS

2.1. Datasets and Preprocessing

Data used in this work were obtained from the ADNI (<http://adni.loni.usc.edu>) and the AIBL study group (<https://aibl.org.au>). For up-to-date information, see (www.adni-info.org). AIBL study methodology has been reported previously [12].

We used longitudinal T1-weighted MRI from ADNI and AIBL, including data from all the participants who had undergone at least two scans. Images were preprocessed using CAT12 [13] with segmentation, spatial normalization, modulation, smoothing (4 mm FWHM), and downsampling to 4 mm^3 , yielding gray matter density maps. Two ADNI training sets were generated: BigDataset (2746 pairs from 1255 participants with 24-month intervals at varying starting timepoints) and SmallDataset (1191 baseline-to-month-24 pairs). Data were split by participant (40 validation, 140 test) to prevent leakage. The ADNI test set consisted of baseline-to month-24 pairs. AIBL served as external validation with 18-month scans extrapolated to 24 months (173 participants).

2.2. Deep Learning Architectures, Training and Evaluation

Let x_t^i be 3D MRI of participant i at time t expressed in months from baseline. The general problem we solve is as follows: Given a training set $\{(x_{t_{k_1}}^i, x_{t_{k_2}}^i) : i \in \text{Train}, t_{k_1} < t_{k_2}\}$, learn a model g that predicts $x_{t_2}^j$ based on $x_{t_1}^j$, $t_1 < t_2$. In practice, we set a constraint that $t_{k_2} = t_{k_1} + 24$, $t_2 = t_1 + 24$.

We implemented five architectures predicting 3D MRI at t_{k+24} from t_k : **UNet** [14] with encoder-decoder and skip connections; **U²-Net** [15] with nested residual U-blocks for multi-scale features; **UNETR** [16] using Vision Transformer encoder with convolutional decoder; **Time-Embedding UNet** (TEUNet)[17] with learnable time embeddings; **ODE-UNet** [18, 19] models continuous latent dynamics via $\dot{Z} = f_\theta(Z, t)$; integrating this neural ODE from baseline to $t+24$ months yields the predicted future scan. Here $Z(t)$ is the latent state, t is time (months), and f_θ is the learned vector field.

UNet, U²-Net, UNETR, and TEUNet used Mean Squared Error (MSE) loss. ODE-UNet employed $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} + \lambda_{\text{ODE}}\mathcal{L}_{\text{ODE}}$ following ImageFlowNet [17], where $\mathcal{L}_{\text{feat}}$ is a perceptual/feature-matching term (using a frozen encoder) to preserve anatomy/texture, $\mathcal{L}_{\text{ODE}}$ penalizes deviations from the latent ODE $\dot{Z} = f_\theta(Z, t)$ along the 24-month

trajectory, and λ_{feat} , λ_{ODE} are scalar weights set to 0.1 in our experiments. All models used the Adam optimizer (learning rate 10^{-4} , weight decay 10^{-5} , batch size 1 3D MRI) with batch normalization and early stopping on A100 GPUs. Each model was trained with a fixed maximum training time of 36 hours. Evaluation metrics were MSE, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM). To study how well the models predict changes, we used Δ -Pearson correlation for longitudinal change. Define predicted image by $\hat{x}_t^i$ and vectorized in-mask change maps $\Delta^i = \text{Vec}(x_{t+24}^i - x_t^i)$ and $\hat{\Delta}^i = \text{Vec}(\hat{x}_{t+24}^i - x_t^i)$. Then participant-wise (Global) Δ -Pearson correlation is the average Pearson correlation $(1/|\text{Test}|) \sum_{j \in \text{Test}} r(\hat{\Delta}^j, \Delta^j)$.

To define *Voxel-wise Δ -Pearson correlation* for each voxel v , we compute $r(\{\hat{\Delta}_v^j\}_j, \{\Delta_v^j\}_j)$ across participants and average the resulting r over in-mask voxels.

3. RESULTS

3.1. ADNI Test Set Performance

Table 1 presents comprehensive results. On BigDataset, U²-Net achieved best reconstruction (MSE= 3.27×10^{-4} , PSNR=31.32 dB, SSIM=0.990), while ODE-UNet achieved highest change prediction (Δ -Pearson=0.253). When trained on SmallDataset, UNETR achieved lowest MSE (3.35×10^{-4}), while ODE-UNet excelled at change prediction (Δ -Pearson=0.215).

3.2. AIBL External Validation

Models generalized robustly to AIBL (Table 1). ODE-UNet trained on SmallDataset achieved exceptional results (MSE= 2.21×10^{-4} , PSNR=33.33 dB, SSIM=0.994, Δ -Pearson=0.205), surpassing ADNI performance. Figure 1 illustrates consistent cross-dataset performance.

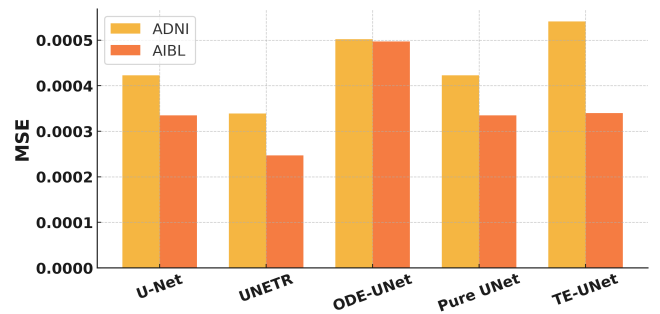


Fig. 1: MSE comparison between ADNI and AIBL test sets.

3.3. Error Analysis

Figure 2 shows U²-Net exhibited most concentrated errors, with all models showing higher errors in rapid-change regions.

Model	Train	ADNI Test Set					AIBL Test Set				
		MSE	PSNR	SSIM	Δ -Pearson (Global)	Δ -Pearson (Voxel)	MSE	PSNR	SSIM	Δ -Pearson (Global)	Δ -Pearson (Voxel)
UNet	Big	4.23	30.34 ± 2.99	0.987	0.202	0.166 ± 0.091	3.35	31.40 ± 3.09	0.991	0.209	0.201 ± 0.086
U ² -Net	Big	3.27	31.32 ± 2.48	0.990	0.241	0.185 ± 0.131	2.37	33.25 ± 2.64	0.993	0.169	0.144 ± 0.123
UNETR	Big	3.39	30.98 ± 2.46	0.989	0.231	0.197 ± 0.099	2.47	32.27 ± 2.62	0.992	0.174	0.169 ± 0.091
TEUNET	Big	5.41	29.47 ± 3.18	0.988	0.126	0.161 ± 0.102	3.40	31.80 ± 3.23	0.992	0.180	0.219 ± 0.092
ODE-UNet	Big	5.02	29.09 ± 2.51	0.987	0.253	0.227 ± 0.111	4.97	29.16 ± 2.63	0.989	0.186	0.191 ± 0.104
UNet	Small	4.31	30.39 ± 2.96	0.987	0.100	0.093 ± 0.092	2.94	32.24 ± 3.11	0.991	0.170	0.160 ± 0.084
U ² -Net	Small	3.44	31.04 ± 2.41	0.989	0.171	0.167 ± 0.117	2.66	32.64 ± 2.61	0.993	0.133	0.123 ± 0.110
UNETR	Small	3.35	31.01 ± 2.47	0.989	0.194	0.173 ± 0.101	2.49	32.31 ± 2.65	0.992	0.175	0.156 ± 0.095
TEUNET	Small	5.26	29.04 ± 2.86	0.986	0.190	0.181 ± 0.114	4.98	29.04 ± 2.47	0.989	0.140	0.169 ± 0.101
ODE-UNet	Small	3.53	31.01 ± 2.60	0.990	0.215	0.226 ± 0.107	2.21	33.33 ± 2.73	0.994	0.205	0.205 ± 0.095

Table 1: Performance metrics on ADNI and AIBL test sets. **MSE values are reported as $\times 10^{-4}$** (i.e., actual MSE = shown value $\times 10^{-4}$). Values for PSNR and voxel-wise Δ -Pearson are reported as mean $\pm$ standard deviation across participants and in-mask voxels, respectively.

Figures 3 and 4 demonstrate ODE-UNet’s preservation of intensity distributions and capture of longitudinal change patterns.

4. DISCUSSION

This work presents a comprehensive comparison among five deep learning architectures for longitudinal, voxel-level MRI prediction across participants with varying levels of cognitive impairment (none to dementia). Given a single baseline, models predict future MRI scans that capture spatially complex patterns of potential neurodegeneration. U²-Net provided the best overall image fidelity in three of the four Train $\times$ Dataset blocks (highest PSNR and lowest/near-lowest MSE), while ODE-UNet was competitive—being best on AIBL/Small (PSNR 33.33 ± 2.73 dB, SSIM 0.994). UNETR achieved the lowest MSE only on ADNI/Small. On the other hand, ODE-UNet performed best on change prediction, with Δ -correlation values as high as 0.253, underlining the strength of continuous-time modeling with respect to disease evolution. Taken together, these results indicate that MSE alone may not be a reliable indicator of predictive quality: models can obtain lower MSE yet produce less realistic or less temporally consistent longitudinal predictions. For example, although U²-Net and UNETR sometimes achieved lower MSE, ODE-UNet yielded higher Δ -Pearson correlation (global and voxel-wise), indicating better agreement with longitudinal change.

On average, under a fixed 36 h training budget, models trained on the more diverse BigDataset performed best especially for change prediction, where ODE-UNet achieved higher Δ -Pearson scores both globally and at the voxel level. In contrast, SmallDataset training sometimes resulted in lower MSE, largely because its smaller size allowed models to fully converge within the time limit, whereas BigDataset runs remained under-optimized for some models. Thus, the observed MSE difference primarily reflects the constrained optimization time rather than any inherent limitation of BigDataset. All

models generalized very well to the independent AIBL cohort, indicating that learned patterns reflect core aspects of brain aging. Among them, U²-Net and ODE-UNet demonstrated the smallest errors on AIBL, thus supporting robustness across populations.

Voxel-level prediction allows advantages over regional biomarker modeling, such as preserving spatial dependencies and enabling individualized downstream analyses. Corresponding error maps and histograms confirmed low reconstruction error and effective modeling of both subtle and gross changes.

We made certain methodological choices that warrant commenting. We chose to use spatially normalized and modulated gray matter density (GMD) maps rather than raw MRIs as GMD can be considered to be quantitative in the sense that GMDs from two individuals can be compared. We trained models on a fixed 24-month prediction for both methodological and conceptual simplicity. We chose to downsample images for computational reasons. Prediction of raw images, longer and variable horizons, multi-modal inputs, e.g. integration of demographic and genetic information, and uncertainty estimation are left for future work. Our results suggest that deep learning has the potential to provide individualized forecasting of brain atrophy. Provided that such models are further refined and validated, they will have significant implications for prognosis, improvement in clinical trial design, and personalized planning of interventions for neurodegenerative disease.

Compliance with Ethical Standards

This retrospective study used data from ADNI (<http://adni.loni.usc.edu>) and AIBL (<https://aibl.org.au>), which received ethical approval with participant consent. Secondary analysis required no additional approval per data use agreements.

Acknowledgments

Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI)

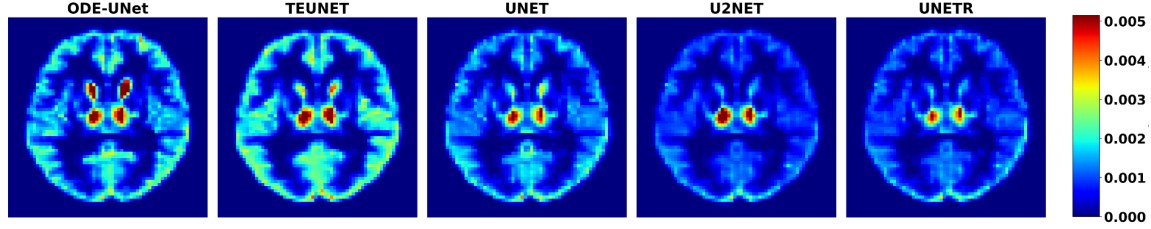


Fig. 2: Voxel-wise MSE maps averaged across ADNI test participants. Axial slices at $z = 14mm$ of the MNI space are shown.

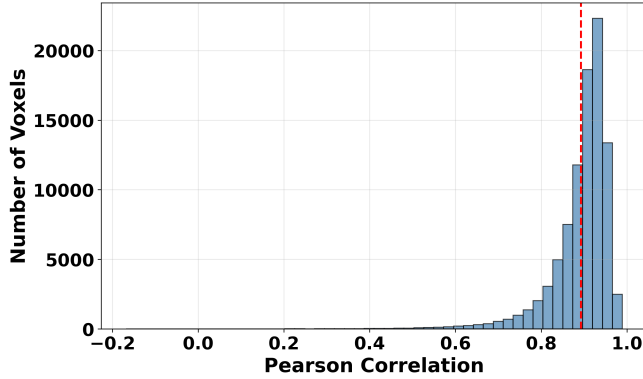


Fig. 3: Histogram of voxel-wise correlations between the predicted 24-month image and the actual 24-month image across participants in ADNI test set (ODE-UNet trained on BigDataset).

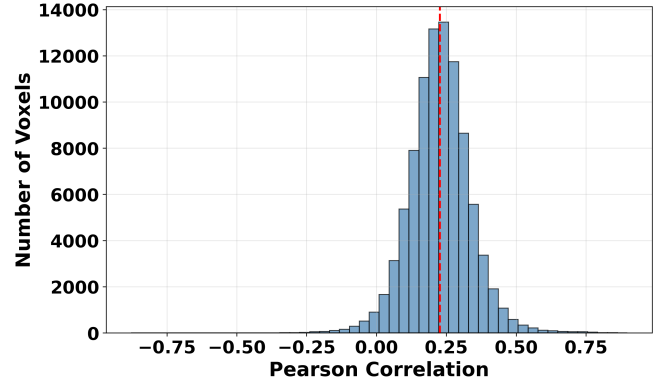


Fig. 4: Histogram of voxel-wise Δ -Pearson correlations on ADNI test set (ODE-UNet trained on BigDataset).

database (<http://adni.loni.usc.edu>) and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO), which was made available at the ADNI database. The investigators within ADNI and AIBL contributed to the design and implementation of ADNI/AIBL and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. AIBL researchers are listed at <http://www.aibl.csiro.au>

Data collection and sharing for ADNI were supported by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd;

Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Supported by The Research Council of Finland grant 351849 (under the frame of ERA PerMed (“Pattern-Cog”), 346934 (PRIMAL), and 358944 (Flagship of Advanced Mathematics for Sensing Imaging and Modeling) and by the Finnish Ministry of Education and Culture’s Pilot for Doctoral Programmes (Pilot project Mathematics of Sensing, Imaging and Modelling). We acknowledge CSC – IT Center for Science, Finland and UEF Bioinformatics Center.

We used OpenAI’s GPT-5 for language editing and general readability during manuscript preparation. The authors reviewed and revised the text subsequently, so they are solely

responsible for the final content.

5. REFERENCES

- [1] T Xia, A Chartsias, C Wang, and SA Tsafaris, “Learning to synthesise the ageing brain without longitudinal data,” *Med Image Anal*, vol. 73, pp. 102169, 2021.
- [2] D Ravi, SB Blumberg, S Ingala, F Barkhof, DC Alexander, and NP Oxtoby, “Degenerative adversarial neuroimage nets for brain scan simulations: Application in ageing and dementia,” *Med Image Anal*, vol. 75, pp. 102257, 2022.
- [3] Simeon Spasov, Luca Passamonti, Andrea Duggento, Pietro Lio, Nicola Toschi, Alzheimer’s Disease Neuroimaging Initiative, et al., “A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer’s disease,” *Neuroimage*, vol. 189, pp. 276–287, 2019.
- [4] E. Moradi, R. Dahnke, V. Imani, C. Gaser, A. Solomon, and J. Tohka, “Integrating plasma, mri, and cognitive biomarkers for personalized prediction of decline across cognitive domains,” *Neurobiology of Aging*, vol. 155, pp. 53–65, 2025.
- [5] Olivier Colliot, *Machine Learning for Brain Disorders*, Springer Nature, 2023.
- [6] S Dayarathna, KT Islam, S Uribe, G Yang, M Hayat, and Z Chen, “Deep learning based synthesis of mri, ct and pet: Review and analysis,” *Med Image Anal*, vol. 92, pp. 103046, 2024.
- [7] A Sikka, S Peri, JS Virk, U Niyaz, and DR Bathula, “Mri-to-pet cross-modality translation using globally & locally aware gan (gla-gan) for multi-modal diagnosis of alzheimer’s disease,” *J Precis Med Health Dis*, p. 100004, 2025.
- [8] Salman UH Dar, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Cukur, “Image synthesis in multi-contrast mri with conditional generative adversarial networks,” *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2375–2388, 2019.
- [9] G Pombo, R Gray, MJ Cardoso, S Ourselin, G Rees, J Ashburner, and P Nachev, “Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3d deep generative models,” *Med Image Anal*, vol. 84, pp. 102723, 2023.
- [10] J Fu, Y Zheng, N Dey, D Ferreira, and R Moreno, “Synthesizing individualized aging brains in health and disease with generative models and parallel transport,” *Med Image Anal*, p. 103669, 2025.
- [11] M Hadji, E Moradi, and J Tohka, “Predicting future brain atrophy based on longitudinal mri,” *medRxiv*, pp. 2025–04, 2025.
- [12] KA Ellis, AI Bush, D Darby, D De Fazio, J Foster, P Hudson, NT Lautenschlager, N Lenzo, RN Martins, P Maruff, et al., “The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer’s disease,” *Int Psychogeriatr*, vol. 21, no. 4, pp. 672–687, 2009.
- [13] C Gaser, R Dahnke, PM Thompson, F Kurth, and E Luders, “Cat: a computational anatomy toolbox for the analysis of structural mri data,” *Gigascience*, vol. 13, pp. giae049, 2024.
- [14] O Ronneberger, P Fischer, and T Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Int Conf Med Image Comput Comput Assist Interv*. Springer, 2015, pp. 234–241.
- [15] X Qin, Z Zhang, C Huang, M Dehghan, OR Zaiane, and M Jagersand, “U2-net: Going deeper with nested u-structure for salient object detection,” *Pattern Recognit*, vol. 106, pp. 107404, 2020.
- [16] A Hatamizadeh, Y Tang, V Nath, D Yang, A Myronenko, B Landman, HR Roth, and D Xu, “Unetr: Transformers for 3d medical image segmentation,” in *Proc IEEE/CVF Winter Conf Appl Comput Vis*, 2022, pp. 574–584.
- [17] C Liu, K Xu, LL Shen, G Huguette, Z Wang, A Tong, D Bzdok, J Stewart, JC Wang, LV Del Priore, et al., “Imageflownet: Forecasting multiscale image-level trajectories of disease progression with irregularly-sampled longitudinal medical images,” in *ICASSP 2025-2025 IEEE Int Conf Acoust Speech Signal Process (ICASSP)*. IEEE, 2025, pp. 1–5.
- [18] RTQ Chen, Y Rubanova, J Bettencourt, and DK Duvenaud, “Neural ordinary differential equations,” *Adv Neural Inf Process Syst*, vol. 31, 2018.
- [19] M Biloš, J Sommer, SS Rangapuram, T Januschowski, and S Günnemann, “Neural flows: Efficient alternative to neural odes,” *Adv Neural Inf Process Syst*, vol. 34, pp. 21325–21337, 2021.