An approach to building foundation models for brain image analysis

Davood Karimi

Computational Radiology Laboratory, Department of Radiology
Boston Children's Hospital, and Harvard Medical School, Boston, MA, USA
davood.karimi@childrens.harvard.edu

Abstract. Existing machine learning methods for brain image analysis are mostly based on supervised training. They require large labeled datasets, which can be costly or impossible to obtain. Moreover, the trained models are useful only for the narrow task defined by the labels. In this work, we developed a new method, based on the concept of foundation models, to overcome these limitations. Our model is an attention-based neural network that is trained using a novel self-supervised approach. Specifically, the model is trained to generate brain images in a patch-wise manner, thereby learning the brain structure. To facilitate learning of image details, we propose a new method that encodes high-frequency information using convolutional kernels with random weights. We trained our model on a pool of 10 public datasets. We then applied the model on five independent datasets to perform segmentation, lesion detection, denoising, and brain age estimation. Results showed that the foundation model achieved competitive or better results on all tasks, while significantly reducing the required amount of labeled training data. Our method enables leveraging large unlabeled neuroimaging datasets to effectively address diverse brain image analysis tasks and reduce the time and cost requirements of acquiring labels.

Keywords: foundation models · neuroimaging · deep learning · brain

1 Introduction

Existing machine learning methods for brain image analysis have three critical shortcomings: (1) They are mostly trained with supervised learning, which requires large labeled datasets. It can be costly or impossible to collect such datasets, for example for segmenting complex structures or detecting rare abnormalities. (2) Trained models are restricted to a narrowly-defined task. For example, a model trained to detect stroke lesions may be useless for detecting tumors. (3) They fail to leverage massive unlabeled neuroimaging datasets that are becoming increasingly available. Therefore, the goal of this work is to develop a new approach, based on foundation models, to overcome these limitations.

1.1 Review of related works

Not-fully-supervised methods: There is much interest in unsupervised, self-supervised, and related machine learning methods for medical image analysis [11, 22, 23]. These methods can reduce/eliminate the need for labeled data. For brain abnormality detection, as an example, prior unsupervised works include reconstruction-based and representation-based methods [4, 10]. However, in general, prior works have failed to perform as well as supervised methods [5, 31].

Foundation models: The central idea behind foundation models is to leverage massive unlabeled data to develop powerful "general-purpose" models, which can subsequently tackle a wide range of tasks [6, 46]. Foundation models are trained using self-supervised techniques. They have become dominant in natural language processing (NLP) and computer vision [12, 14] and offer important advantages: (i) They do not require collecting and labeling a large dataset for each new task. (ii) They can learn patterns that are not restricted by a specific label information or tied to a specific task. (iii) They can have better performance on rare events and corrupted or out-of-distribution data [30, 6].

Vision transformers and their limitations: Attention-based models such as vision transformers have emerged as alternatives to CNNs for image analysis [13, 24]. They are superior in learning spatial correlations. However, they also have drawbacks such as higher computational and memory requirements and lower data efficiency. A limitation of these models that we consider in this work is that they are not effective in learning high-frequency information [32, 2]. Hence, they are not good at reconstructing sharp features, i.e., edges and texture, which can impact the performance in image analysis tasks.

1.2 Contributions

(1) We propose a framework for developing foundation models for brain image analysis. Our model is based on a transformer network that is trained in a self-supervised manner to *generate* brain images in a patch-wise fashion, thereby learning the detailed structure of brain images. Our novel training strategy is based on auto-regressive prediction and random input masking, both applied at the level of 3D patches. (2) To facilitate learning of high-frequency detail, we propose a novel approach based on convolutional kernels with random weights. (3) We train the model on 10 unlabeled datasets. We then apply the model on five different tasks and show that it can achieve competitive or better results than the state of the art while significantly reducing the required labeled data.

2 Methods

As shown in Fig. 1, our method consists of a transformer network that is trained to generate brain images in a patch-wise fashion. It uses a set of 3D patches

as the input context to predict an adjacent patch. This way, the model learns the detailed structure of brain images. Our method is based on a simple but powerful idea: *normal brains are very similar*. In other words, brain structure is characterized by features (e.g., cortical foldings and subcortical structures) that are similar across brains. Hence, a well-devised model can learn these structures after being trained in a self-supervised manner on large datasets to generate these images. In this sense, our approach follows the NLP foundation models that are trained to predict the next word in text. The method works as follows:

1. Given an image, cubes of size D^3 are extracted, where D is in voxels. The cube is divided into patches of size d^3, where $d{=}D/n$, resulting in n^3 patches.
2. The corner patch (red in Fig. 1) is the prediction target. The remaining n^3-1 patches (yellow/orange in Fig. 1) constitute the input context, which the model uses to predict the target. The input patches go through high-frequency feature encoding (block **F** in Fig. 1), described further below.
3. Frequency encoding is followed by embedding into $\mathbb{R}^m$, resulting in a sequence S of length n^3-1, where each element is in $\mathbb{R}^m$. Positional encoding ($\mathbf{P}_2$ in Fig. 1) is added to S and then the elements of S are masked (i.e., removed) at random (block **M**). For positional encoding, we use fixed (i.e., non-learned) encodings as in standard transformers [43, 28]. For masking, we remove each element independently with a probability p. This generates a shorter sequence S^* of length n^*.
4. The reduced sequence S^* is passed to a network (**T** in Fig. 1), consisting of K_{tr} transformer blocks with architecture similar to those in [20].
5. The output of the last transformer block goes through unmasking (**U** in Fig. 1). This simply restores the sequence length to n^3-1 based on the same masking pattern used to generate S^*. The masked elements are given a value of 0 in this stage. The sequence is passed to K_{fc} fully-connected layers.
6. The output of the last fully connected layer is projected back to $\mathbb{R}^{d^3}$ and reshaped (**R** in Fig. 1) to form the model's prediction of the target patch.

High-frequency encoding: To address transformer's limitation in modeling high-frequency information, we encode this information using convolutional kernels with random weights. Neural networks with random weights are highly effective in extracting high-frequency information [35, 9]. Our aim is to design q kernels $\{k_i\}_{i=1:q}$ such that there is low redundancy among the feature maps computed by the set. We first generate a much larger set of candidates $\{k_i^*\}_{i=1:Q}$, where $Q \gg q$, using the initialization method proposed in [15]. We apply $\{k_i^*\}_{i=1:Q}$ on some training images (x) to generate feature maps $f_i^*(x) = k_i^* \circledast x$. We quantify the similarity between pairs of feature maps ($f_i^*(x)$ and $f_j^*(x)$) using projection weighted Canonical Correlation Analysis (pwCCA) [27], which we denote as ρ_{ij}. We select the subset of q kernels in a greedy manner: (step 1) We choose k_1 to be k_1^*; (step 2) We choose k_2 from $\{k_i^*\}_{i=2:Q}$ such that ρ_{1i} is the lowest, i.e., f_i is least similar to f_1; (steps 3-q) We proceed in a similar manner, in every step choosing the kernel that has the lowest maximum similarity (lowest maximum ρ) compared with already-selected kernels. We used $q = 8$ and $Q = 100$. To avoid increasing the dimensionality of the network input, we *add* the high-frequency

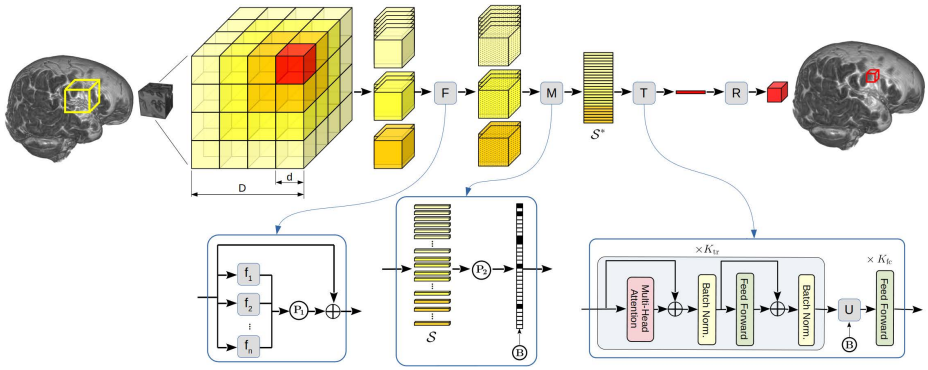

Fig. 1. Proposed method. **F**: Frequency encoding block, where $\boldsymbol{f_1}$-$\boldsymbol{f_n}$ denote different frequency encodings and $\boldsymbol{P_1}$ is positional encoding. **M**: Positional encoding and masking, where $\boldsymbol{P_2}$ is positional encoding and **B** denotes a Bernoulli random process used to mask the sequence. **T**: Vision transformer block. **U**: Unmasking. **R**: Reshaping.

features to the input. Specifically, if x is the input patch to the high-frequency encoding block, the output is $x + \sum_i P_{1i} k_i \circledast x$, where P_1 is fixed "positional" encoding [43]. Note that these kernels are fixed and not updated during training.

Training approach: Our novel training approach is based on two dominant themes in self-supervised learning [12, 1]: (1) Generative methods: they train the model to predict masked/corrupted parts of the input. (2) Invariance-based methods: they train the model to compute similar representations for different *views* of the same input. As shown in Fig. 1, our method is based on auto-regressive prediction of a missing (corner) patch; hence it follows a generative approach. However, we also randomly mask the input patches, thereby giving the model a different view of the input in each training iteration. Each element in S is masked based on a Bernoulli distribution with probability p. The sequence is unmasked to its original length before the fully connected layers, which serve as the decoder section of the network. This approach forces the model to predict the target patch based on a random selection of non-masked patches, which is an invariance-based method and, additionally, serves as data augmentation.

Implementation: We trained our model on a pool of 10 public datasets [16, 29, 39, 8, 7, 33, 36, 40, 37, 41]. From these datasets, we used 11,000 structural MRI and 12,000 diffusion MRI (dMRI) volumes (acquired with different diffusion gradient directions and strengths). Parameter settings were (see Fig. 1): $D = 32$, $d = 8$, $n = 4$, $m = 512$, $p = 0.10$, $K_{\text{tr}} = 5$, $K_{\text{fc}} = 2$. We used the ℓ_1 norm of the difference between the predicted and true voxel intensity of the target patch as the loss function. During training, we sample cubes from random locations in the training images and use them to optimize the model. On a test image, we start at one corner of the image and proceed in a sliding-window fashion to predict the image. To predict all patches that are close to the brain boundaries, we apply the model in different directions (i.e., left/right, superior/inferior, and anterior/posterior).

We trained the model on an Nvidia RTX A4500 GPU for 15 days. We used a batch size of 10, SGD optimizer with a learning rate of 0.001.

Two factors may influence the model's performance and generalizability: (i) Resolution. We account for this factor by resampling all images to an isotropic resolution of 1 mm. (ii) Intensity. We normalize each image such that the voxel intensities have a mean of zero and standard deviation of one.

2.1 Experiments and evaluation strategy

In order to perform *extrinsic evaluation* [6] of our foundation model, we applied the trained model on five different downstream tasks.

Task 1- Cortical plate segmentation. We used 100 T2 images and cortical plate segmentations from the dHCP dataset [3]. We compared the foundation model with nnU-Net [17] and a transformer model [21]. Both these competing networks were trained in a fully-supervised manner.

Task 2- White matter tract segmentation. We used tract segmentation data from the TractSeg project [45]. We focused on three tracts: corticospinal tract (CST), middle cerebellar peduncle (MCP), and optic radiation (OPR). We compared with nnU-Net, trained in a fully-supervised manner.

Task 3- dMRI denoising. We used dMRI scans of 100 subjects from the HCP Development dataset [38]. Compared methods included: MPPCA [44], which is a widely used method based on random matrix theory, and SDnDTI [42], which is a recent deep learning method.

Task 4- Lesion detection. We used a dataset of 30 acute stroke patients [34]. We compared with an unsupervised technique based on variational autoencoders (VAE) [5].

Task 5- Brain age estimation. We used 300 T2 images from the dHCP dataset [3]. We compared our foundation model with a residual CNN (ResCNN) [19].

Fine-tuning and application to the target tasks: For segmentation (Tasks 1 and 2) and brain age estimation (Task 5), we fine-tuned the foundation model on small numbers of labeled data from the target tasks (details given below). For these tasks, the last layer of the network had to be modified to output the correct size (For Tasks 1 and 2: $\mathbb{R}^{d^3 \times 2}$ to represent the foreground and background segmentation predictions; For Task 5: a scalar to represent the brain age). For Tasks 1 and 2 we fine-tuned the model using a cross-entropy loss; for Task 5 we fine-tuned with an ℓ_2 loss. Fine-tuning was performed on the output layer and fully-connected layers; we did not fine-tune the transformer blocks. For Tasks 3 and 4 no fine-tuning was performed. For Task 3 (lesion detection) we expected that the model trained on normal brain images would display large reconstruction errors on brain lesions due to their deviation from normal brain appearance. This is the common assumption in abnormality detection [4, 5]. Hence, we applied our model on a test image, computed the reconstruction error, and used an empirical threshold of 1.50 to detect the lesions. For Task 4, following a similar argument, we expected that the trained model learned to reconstruct the true dMRI signal and not the random noise. This is the rationale behind

deep learning-based image denoising [26, 25]. Hence, we applied the foundation model on the test images and used the model predictions as the denoised image.

Evaluation metrics: For segmentation (Tasks 1 and 2) we used Dice Similarity Coefficient (DSC), 95 percentile of the Hausdorff Distance (HD95), and Average Symmetric Surface Distance (ASSD). We assessed the denoising performance (Task 3) in two ways: (i) We used all 94 measurements in the b=1500 shell to estimate a "ground truth" diffusion tensor. Then, we chose 15 measurements from these 94 measurements, denoised them, estimated the diffusion tensor using the denoised measurements, and computed fractional anisotropy (FA) and mean diffusivity (MD) from the tensor and compared these with FA and MD computed from the ground truth diffusion tensor. We denote the difference with the ground truth as ΔFA and ΔMD. (ii) We divided the 186 dMRI measurements (in b=1500 and b=3000 shells) for each subject into two subsets of 93 measurements. We used MSMT-CSD [18] to compute the fiber orientation distribution from each subset and extracted the peak orientation direction for voxels in white matter. We computed the angle between the peaks estimated from the two subsets, denoted as $\Delta\theta$, as a measure of disagreement that is caused, in part, by residual noise. For lesion detection (Task 4) we used F1 score. For Task 5, we computed the error as the difference between the predicted age and the true age.

3 Results and discussion

Intrinsic evaluation [6]: As shown in Fig. 2, our model can accurately reconstruct test images. A neuroradiologist reviewed pairs of true and reconstructed images in a blind fashion and confirmed that the images were indistinguishable even in minute details. Quantitatively, on 25 test images from the HCP dataset our method achieved a voxel-wise reconstruction error of 0.098 ± 0.024, compared with 0.132 ± 0.026 for an auto-encoder model [4]. The difference was statistically significant ($p < 0.001$, computed with a paired t-test) *despite the fact that*, unlike the auto-encoder, our model did not use the target patch in its context input.

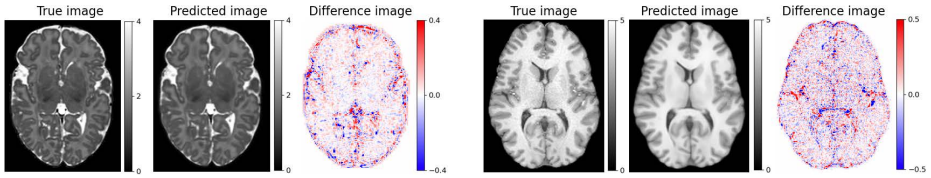

Fig. 2. Example test images reconstructed by the proposed method. Left: a T2 image from the dHCP dataset; Right: a T1 image from the HCP dataset.

Task 1: With only 15 labeled training images in the target domain, our foundation model achieved segmentation metrics on par with the competing methods trained with 50-250 labeled images (example plot shown in Fig. 3(a)). Fig. 3(b)

shows that the foundation model can achieve highly accurate segmentation with few labeled images in the target domain. Paired t-tests for DSC, HD95, and ASSD showed that for every number of labeled training images the foundation model achieved significantly ($p < 0.001$) more accurate results than the two compared methods. Moreover, DSC, HD95, and ASSD for the foundation model fine-tuned with 15 labeled images were not different from the results achieved by the two compared methods trained with 250 labeled images ($p > 0.10$).

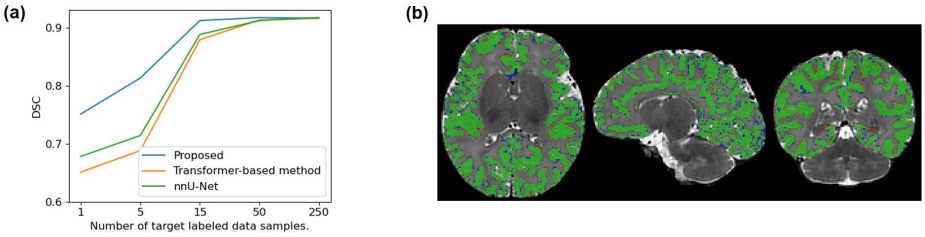

Fig. 3. (a) Plots of DSC for our foundation model and competing methods on Task 1. (b) Example results for the foundation model fine-tuned with 15 labeled images. (Green voxels: correct segmentation, blue: over-segmentation, red: under-segmentation.)

Task 2: As Table 1 shows, with fewer labeled images in the target domain, the foundation model achieved segmentation performance that was comparable with or better than nnU-Net. For all three tracts and all three numbers of labeled images, the foundation model achieved significantly higher DSC than nnU-Net ($p < 0.001$, computed with paired t-tests). For all three tracts, the DSC achieved with foundation model using 15 labeled images was not different ($p > 0.10$) than that of nnU-Net with 50 labeled images. As shown in Fig. 4, the foundation model can segment diverse and complex structures from few labeled images.

Table 1. DSC for the proposed method and nnU-Net on Task 2. m denotes the number of labeled images used to fine-tune our foundation model and train nnU-Net.

	$m = 5$			$m = 15$			$m = 50$		
Method	CST	MCP	OPR	CST	MCP	OPR	CST	MCP	OPR
Foundation model	0.871	0.831	0.786	0.874	0.840	0.797	0.882	0.851	0.810
nnU-Net	0.814	0.807	0.718	0.850	0.822	0.790	0.864	0.833	0.800

Task 3: Examples in Fig. 5 show that our foundation model, without any fine-tuning on the target task data, can effectively reconstruct the genuine dMRI signal while suppressing the noise. Table 2 shows that the foundation model achieves comparable results with the other two methods. All three methods achieved ΔFA and ΔMD that were statistically not different ($p > 0.10$). In terms of $\Delta\theta$, our model achieved a statistically smaller error than SDnDTI ($p < 0.001$).

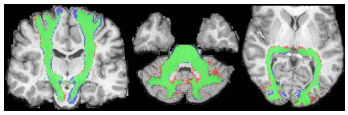

Fig. 4. Example CST, MCP, and OPR segmented with the foundation model fine-tuned on 15 images. Green: correct; blue: false positive; red: false negative. (Higher magnification image in supp. material.)

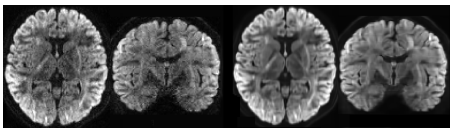

Fig. 5. Example dMRI denoising results by the foundation model in Task 3. Left: noisy; right: denoised.(Higher magnification image in supp. material.)

While our method was not fine-tuned on any training data from the target dataset, SDnDTI was trained with 50 images.

Table 2. Results for Task 3.

Method	δFA	δMD	$\delta\theta(^\circ)$
Proposed	0.031	0.039	4.20
SDnDTI	0.031	0.041	4.29
MPPCA	0.030	0.039	4.24

Table 3. Mean absolute error (weeks) in brain age estimation (Task 5).

Method	$m = 100$	$m = 200$	$m = 400$
Proposed	0.87	0.86	0.86
ResCNN	1.03	0.90	0.89

Task 4: Foundation model achieved an F1 score of 0.688 compared with 0.605 for VAE. Wilcoxon signed-rank test showed that the difference was significant ($p < 0.001$).
Task 5: Table 3 shows that our method fine-tuned with 100 images achieved a lower prediction error than ResCNN trained with 400 images. With each of the three different numbers of training images in the target domain, our method achieved significantly ($p < 0.001$) lower prediction errors than ResCNN.
Ablation experiments: Extensive experiments showed a substantial positive impact for the proposed high-frequency encoding approach. For example, disabling the high-frequency encoding in Tasks 1 and 2 reduced the DSC achieved by the method such that it was no longer significantly better than nnU-Net. Disabling the high-frequency encoding in Task 4 reduced our F1 score to 0.550, making it not significantly better than VAE.

4 Conclusions

Our results show that the proposed method can be used to develop foundation models to address a wide range of brain image analysis tasks. The model size and datasets used in this work were much smaller than in current foundation models in NLP and computer vision. Therefore, we believe the results reported in this paper will significantly improve by employing larger models and datasets. Nonetheless, our results show that our method can lead to (1) improvements in

the accuracy of machine learning methods in analyzing brain images; (2) reduction in labeled data requirements for building new machine learning methods to address existing problems and emerging needs; (3) expansion of the range of brain image analysis tasks that can be addressed with machine learning, such as detection of rare abnormalities, where training data can be very scarce.

Acknowledgments. This study was supported in part by the National Institute of Neurological Disorders and Stroke and Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health (NIH) under award numbers R01HD110772 and R01NS128281. The content of this publication is solely the responsibility of the author and does not necessarily represent the official views of the NIH.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Assran, M., et al.: Masked siamese networks for label-efficient learning. In: European Conference on Computer Vision. pp. 456–473. Springer (2022)
2. Basri, R., et al.: Frequency bias in neural networks for input of non-uniform density. In: International Conference on Machine Learning. pp. 685–694. PMLR (2020)
3. Bastiani, M., et al.: Automated processing pipeline for neonatal diffusion mri in the developing human connectome project. NeuroImage **185**, 750–763 (2019)
4. Baur, C., et al.: Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018. pp. 161–169. Springer (2019)
5. Baur, C., et al.: Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. Medical Image Analysis **69**, 101952 (2021)
6. Bommasani, R., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
7. Bookheimer, S.Y., et al.: The lifespan human connectome project in aging: an overview. Neuroimage **185**, 335–348 (2019)
8. Botvinik-Nezer, R., et al.: Paingen placebo. OpenNeuro (2023). https://doi.org/doi:10.18112/openneuro.ds004746.v1.0.1
9. Cao, W., Wang, X., Ming, Z., Gao, J.: A review on neural networks with random weights. Neurocomputing **275**, 278–287 (2018)
10. Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972 (2018)
11. Cheplygina, V., et al.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis **54**, 280–296 (2019)
12. Devlin, J., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)

15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: IEEE International Conference on Computer Vision (ICCV) 2015 (2015)
16. Howell, B.R., et al.: The unc/umn baby connectome project (bcp): An overview of the study design and protocol development. NeuroImage **185**, 891–905 (2019)
17. Isensee, F., et al.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
18. Jeurissen, B., et al.: Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion mri data. NeuroImage **103**, 411–426 (2014)
19. Jónsson, B.A., et al.: Brain age prediction using deep learning uncovers associated sequence variants. Nature communications **10**(1), 5409 (2019)
20. Karimi, D., Dou, H., Gholipour, A.: Medical image segmentation using transformer networks. IEEE Access **10**, 29322–29332 (2022)
21. Karimi, D., Vasylechko, S.D., Gholipour, A.: Convolution-free medical image segmentation using transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 78–88. Springer (2021)
22. Karimi, D., Warfield, S.K., Gholipour, A.: Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations. Artificial intelligence in medicine **116**, 102078 (2021)
23. Karimi, D., et al.: Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis **65**, 101759 (2020)
24. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM computing surveys (CSUR) **54**(10s), 1–41 (2022)
25. Krull, A., Buchholz, T.O., Jug, F.: Noise2void-learning denoising from single noisy images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2129–2137 (2019)
26. Lehtinen, J., et al.: Noise2noise: Learning image restoration without clean data. arXiv preprint arXiv:1803.04189 (2018)
27. Morcos, A., et al.: Insights on representational similarity in neural networks with canonical correlation. Advances in neural information processing systems **31** (2018)
28. Murphy, K.P.: Machine learning: a probabilistic perspective (2012)
29. Nugent, A.C., et al.: "the nimh healthy research volunteer dataset" (2023). https://doi.org/doi:10.18112/openneuro.ds004215.v1.0.2
30. Orr, L., et al.: Bootleg: Chasing the tail with self-supervised named entity disambiguation. arXiv preprint arXiv:2010.10363 (2020)
31. Pinaya, W.H., et al.: Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. Medical Image Analysis **79**, 102475 (2022)
32. Rahaman, N., et al.: On the spectral bias of neural networks. In: International Conference on Machine Learning. pp. 5301–5310. PMLR (2019)
33. Reynolds, J.E., Long, X., Paniukov, D., Bagshawe, M., Lebel, C.: Calgary preschool magnetic resonance imaging (mri) dataset. Data in brief **29**, 105224 (2020)
34. Rorden, C., Absher, J., Newman-Norlund, R.: "stroke outcome optimization project (soop)" (2024). https://doi.org/doi:10.18112/openneuro.ds004889.v1.1.2
35. Saxe, A.M., Koh, P.W., Chen, Z., Bhand, M., Suresh, B., Ng, A.Y.: On random weights and unsupervised feature learning. In: ICML. vol. 2, p. 6 (2011)
36. Schuch, F., et al.: An open presurgery mri dataset of people with epilepsy and focal cortical dysplasia type ii. Scientific Data **10**(1), 475 (2023). https://doi.org/doi:10.18112/openneuro.ds004199.v1.0.5
37. Snoek, L., et al.: "aomic-id1000". OpenNeuro (2021). https://doi.org/10.18112/openneuro.ds003097.v1.2.1

38. Somerville, L.H., et al.: The lifespan human connectome project in development: A large-scale study of brain connectivity development in 5–21 year olds. Neuroimage **183**, 456–468 (2018)
39. Spreng, R.N., et al.: Neurocognitive aging data release with behavioral, structural and multi-echo functional mri measures. Scientific Data **9**(1), 119 (2022). https://doi.org/doi:10.18112/openneuro.ds003592.v1.0.13
40. Strike, L.T., et al.: "queensland twin adolescent brain (qtab)". OpenNeuro (2022). https://doi.org/doi:10.18112/openneuro.ds004146.v1.0.4
41. Strike, L.T., et al.: Queensland twin imaging (qtim). OpenNeuro (2023). https://doi.org/doi:10.18112/openneuro.ds004169.v1.0.7
42. Tian, Q., et al.: Sdndti: Self-supervised deep learning-based denoising for diffusion tensor mri. Neuroimage **253**, 119033 (2022)
43. Vaswani, A., et al.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
44. Veraart, J., et al.: Denoising of diffusion mri using random matrix theory. Neuroimage **142**, 394–406 (2016)
45. Wasserthal, J., Neher, P., Maier-Hein, K.H.: Tractseg-fast and accurate white matter tract segmentation. NeuroImage **183**, 239–253 (2018)
46. Zhou, C., et al.: A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. arXiv preprint arXiv:2302.09419 (2023)