Modern Vision Architectures Meet Alzheimer's Disease Diagnosis in sMRI

Javier Salazar Cavazos
University of Michigan
Ann Arbor, MI, US
javiersc@umich.edu

Neel Dutta
University of Michigan
Ann Arbor, MI, US
nkdutta@umich.edu

Kartikeya Gupta
University of Michigan
Ann Arbor, MI, US
kartigup@umich.edu

Elbert Yi
University of Michigan
Ann Arbor, MI, US
elbertyi@umich.edu

Abstract

Structural magnetic resonance imaging (sMRI) is commonly used for the identification of Alzheimer's disease (AD) because it captures atrophy-induced changes in brain structure. In this work, we utilize sMRI data as a biomarker to classify patients on the AD spectrum using modern vision architectures such as variational auto encoder, vision transformers, MLP-Mixer, and ConvNeXt. Experimental results show interesting conclusions on features learned, self-supervised vs. supervised methods, and the effects of inductive bias in models for our data.

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disease that gradually worsens and cocurrently achieves neural damage leading to impairment of cognitive ability [1]. Once a patient is no longer cognitively normal (CN), mild cognitive impairment (MCI) is the next prodromal stage of AD. Patients are highly likely to develop dementia of the Alzheimer's type (DAT) once diagnosed with MCI [2]. There are 50 million people affected worldwide (in 2020) with projections of 150 million by 2050 [3]. Thus, there is a rising need for computer systems to assist in diagnosis that can learn from multiple biomarkers. For this work, we investigate the use of structural magnetic resonance imaging (sMRI) data as a biomarker for the prediction and diagnosis of AD. We utilize this image modality since there are structural, meaning anatomical, changes in the brain due to atrophy caused by the disease [4].

Clinical diagnosis of AD consists of patients performing neuropsychological tests that check for memory recall and other mental functions. From this test, a cognition score is calculated, and the patient is classified into one of the three categories. Importantly, clinical diagnosis is challenging, with sensitivity values ranging from 70.9% to 87.3% and specificity ranging from 44.3% to 70.8% [5]. Therefore, there is a possibility that some of the "ground truth" labels we have are incorrect due to other potential confounders like chronic stress, severe depression, and others [6]. In addition, even though there is an early buildup of plaque in the brain (leading to MCI symptoms like memory loss), structural changes like atrophy are observed in later stages of the AD spectrum [6]. Meaning, the sMRI features can potentially overlap between some of the classes, like CN and MCI. In summary, AD classification using sMRI data is a challenging problem due to the confounders and reliability of "ground truth" labels. In the next section, we explore published, related work in the domain of computer vision for AD diagnosis.

2. Related Work

AD classification and machine learning have a rich history, starting with more classical methods such as support vector machines (SVMs) [7] and AdaBoost [8]. With the prevalence of deep learning architectures, researchers have explored more recent alternatives such as ResNet [9], 2D slice-based vision transformers [10], and convolutional auto encoders [11]. Since 3D data is computationally expensive, some have explored using derived features of the scans, such as functional and structural connectivity maps (2D data), within a deep learning setting [12].

However, with the recent advances in architectural modeling such as ConvNeXt and MLP-Mixer, there is an opportunity to explore how these methods compare against other established techniques. Moreover, we directly compare the self-supervised VAE against supervised models to learn what advantages exist when incorporating label information into the classification system as opposed to learning a latent space that represents a person's brain. Further, we are interested to see if more general models (with fewer inductive biases such as MLP-Mixer/ViT) are able to achieve better classification or if we are data limited with our dataset size.

3. Data

We utilized the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [13] for this work. Specifically, we select 3T T1 weighted sMRI scans pulse sequenced with MPRAGE or FGSPR. Note that in ADNI Phase 2/GO the MCI class was subdivided into early MCI (EMCI) and late MCI (LMCI). We discard this data for our classification results unless otherwise stated.

Class	total (train/test)	
	Subjects	Sessions
CN	**711** (568/143)	**711** (568/143)
EMCI	**326** (260/66)	**1814** (1414/401)
MCI	**388** (310/78)	**928** (760/168)
LMCI	**177** (141/36)	**933** (743/190)
DAT	**276** (220/56)	**756** (612/144)

Table 1: Data distribution for ADNI dataset.

Since medical datasets tend to be smaller for positive classes, we first augment our dataset by treating each session as a "psuedo-subject" for nonhealthy patients. We ensure no contamination between the training set and testing set by keeping each subject's scans in either one or the other. This leads to a very balanced distribution for the 2-class and 3-class classification setups. Then, we convert the raw DICOM files to the established BIDS format [14] for all subjects. Finally, we utilize fmriprep [15] to perform the following preprocessing tasks standard in the medical literature:

- N4 bias field correction (brightness uniformity)
- Volume merging (multiple scans for one session)
- Skull stripping (isolating the brain)
- Spatial normalization (image registration)
  - MNI152 linear template with 1mm resolution

We utilize the TorchIO [16] Python library for data loading and augmentation. Our general pytorch pipeline includes the following operations:

- Fix data to the canonical orientation
- Resample to 2mm spatial resolution
- Pad the volumes to (96,108,96) for patch-friendly data
- Normalize intensity values to (-1,1)
- We augment by uniformly picking one of:
  - Random affine transformation
  - Random patient motion artifact
  - Random noise

4. Methods

4.1. β-VAE

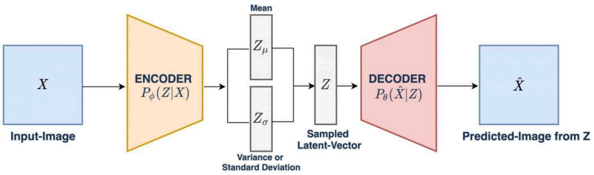

Figure 1: Figure adapted from [17].

In machine learning, self-supervised learning marks a significant shift from supervised approaches. β-VAE emphasizes learning disentangled latent representations of the data, where the hyperparameter β encourages independent latent variables. The data itself is used to supervise the learning process without labels. This is similar to a kind of "nonlinear PCA" approach with learnable parameters to compress the data to a lower dimensional space.

Supervised methods often require extensive labeled datasets and may struggle to generalize beyond their training scenarios due to their reliance on direct supervision. In contrast, β-VAE aims to uncover the underlying causal factors of the data it observes, promoting a form of generalization that is not tied to specific supervised tasks but is rather aimed at a broader understanding of the data's structure. Because of this, we utilize this model to compare against supervised approaches in our possibly size constrained classification problem.

4.2. ConvNeXt

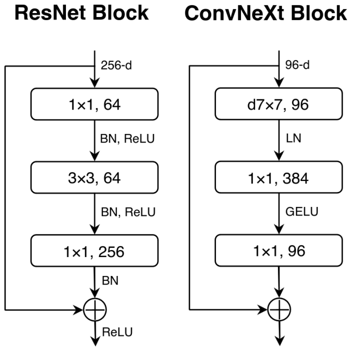

Figure 2: ConvNeXt block design figure taken from [18].

ConvNeXt [18] is a recent convolutional network that updates the ResNet architecture [19] by reintroducing modern ideas first seen in the ViT architecture [20]. A few key ideas for the micro block design are: 1) replacing batch normalization with layer normalization [21] which is great to reliably sample statistics, especially when batch size is small, 2) replacing Rectified Linear Unit (RELU) activation function with Gaussian Error Linear Unit (GELU) function [22] which adds smoothness around zero, making it easier for backpropogation, 3) changing convolution parameters and where they take place in the block design, and 4) introducing an inverted bottleneck MLP after the convolution operation in a similar fashion to ViT. By doing this, and changing some macro level design, they are able to achieve competitive performance compared to SOTA models.

We utilize this architecture to establish a baseline and observe whether the CNN inductive bias is helpful for our smaller dataset. Since this is designed for 2D images, we generalize the model by *inflating* the layers. Meaning, we learn 7x7x7 kernels instead of 7x7 kernels. This is justifiable since we have an extra spatial dimension in contrast to video, where temporal data is more redundant.

4.3. MLP-Mixer

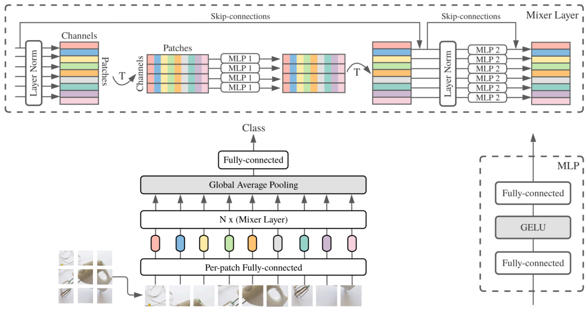

Figure 3: Figure taken from [23].

MLP-Mixer [23] is a novel architecture for vision that diverges from traditional models by exclusively using multi-layer perceptrons (MLPs) instead of convolution or self-attention mechanisms. Key design choices include: 1) employing two types of MLP layers that mix features across tokens (image patches) and channels, which allows the model to interact with different aspects of the input data without spatial convolutions or attention, 2) using simple matrix multiplication, which simplifies the computational requirements and makes the architecture scalable, and 3) leveraging layer normalization and Gaussian Error Linear Unit (GELU) activations to maintain training stability and smoothness around zero.

For our 3D data, we had to redesign the patch embedding process to include the extra dimension, transforming it into a cube extraction process. This change allows the architecture to capture spatial relationships not only in height and width but also across depth while maintaining its original premise of using MLPs for both channel and spatial mixing.

4.4. Vision Transformer (ViT)

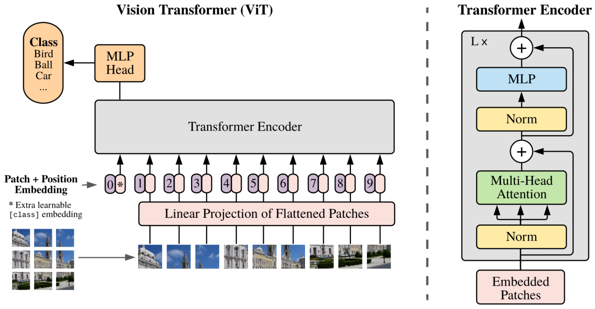

Figure 4: Figure taken from [20].

Vision Transformer (ViT) [20] is an image classification model designed to replicate the success of transformer models in natural language processing (NLP) tasks. The ViT model first divides the input image into patches. Tokens are formed by unrolling each image patch into a vector. Positional encoding is then added to each token describing the location from which the patch originates in the image. After that, the tokens are then sent into transformer layers, as depicted in Figure 4. The transformer layer contains a multi-head self-attention mechanism and an MLP block. Classification is performed by utilizing the class token that is appended to the token matrix.

Just like with MLP-Mixer, we created a 3D patch embedding block to create tokens from patches. Additionally, we generalized the 1D positional encoding formula (made for language) from [24] to the 3D setting instead of the learnable positional encoding used in [20].

5. Experiments

5.1. Setup

We use the Adam optimizer [25] for all of our models without weight decay. We did not experience overfitting since we only used small complexity models (~30-40M), as larger models did not improve classification accuracy. For the cost function, we use weighted cross entropy loss for the supervised models. The weights are set by the inverse frequency of occurance for each class in order to combat class imbalance, although our data is mostly balanced for the vast majority of experiments. For the VAE model, the loss function is a combination of the mean square error between input/output volumes and the Kullback–Leibler divergence term that promotes a standard distribution for the latent variables. We use a learning rate of 10^{-5} for all models, as this results in ideal performance for each model. Generally, each model was trained using a different number of epochs. Some models required more epochs than others to achieve convergence using classification accuracy as the validation metric. For the VAE model, PSNR was used as the validation metric during training.

5.2. Results

In Table 2, we list results on the binary classification problem (CN/DAT). Note that due to the quadratic complexity of the transformer model, ViT is fixed to a patch size of 8x8x8 which is part of the reason why ViT performs worse than other models. Due to time limitations, we did not explore linear computation methods such as ProbSparse Self-Attention from Informer [26]. Moreover, since ViT has less inductive bias, this also leads to worse results given the limited dataset size. MLP-Mixer (p=3) generally performed well, even slightly better than ConvNeXt. This result is surprising since we expected CNNs to perform better under this limited data regime. We believe this is due to ConvNext kernel sizes being tweaked for natural images in the original paper (from ImageNet, for example) instead of our medical application. Possibly, with further tweaking, ConvNeXt would perform equally well. In Table 3, we extend our results to the multi-class setting (CN/MCI/DAT).

Model	Precision		Recall		Accuracy
	CN	DAT	CN	DAT	
ConvNeXt	0.79	**0.85**	**0.83**	0.81	0.82
ViT (p=8)	0.79	0.76	0.69	0.84	0.77
MLP-Mixer (p=3)	**0.87**	0.82	0.76	**0.90**	**0.84**
VAE	0.67	0.80	0.80	0.66	0.73

Table 2: Binary classification (CN/DAT) results.

Model	Precision			Recall			Acc.
	CN	MCI	DAT	CN	MCI	DAT	
Conv NeXt	0.52	**0.53**	**0.67**	**0.69**	0.44	0.61	0.57
ViT (p=8)	0.59	0.42	0.55	0.49	0.37	**0.69**	0.52
MLP Mixer (p=3)	**0.73**	0.45	0.64	0.68	**0.53**	0.57	**0.59**
VAE	0.47	0.45	0.57	0.64	0.40	0.46	0.49

Table 3: Three-class classification results.

Patch Size	(CN/AD) Accuracy
2x2x2	**0.84**
3x3x3	**0.84**
4x4x4	0.82
6x6x6	0.80
12x12x12	0.73

Table 4: Effects of patch size on MLP-Mixer model.

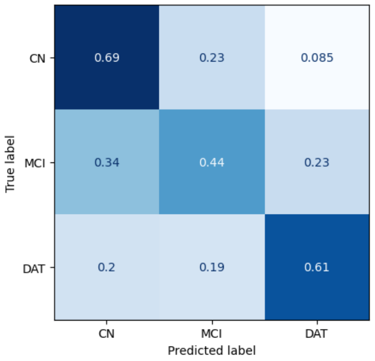

Figure 5: Confusion matrix for the 3-class classification problem using ConvNeXt model.

In Table 4, we explore the effects of patch size on classification performance and observe that small patch sizes are ideal to learn features relevant to diagnosis.

Not surprisingly, as seen in Figure 5, there is a large overlap between CN and MCI, indicating similar features, which makes sense since structural atrophy-based changes occur later in AD subjects. Additionally, some MCI subjects might be impaired due to confounders such as chronic stress and severe depression.

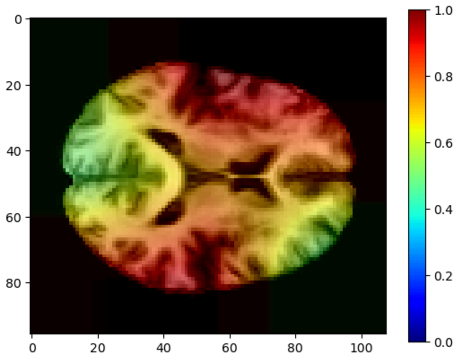

Figure 6: Saliency map generation of DAT subject using Grad-CAM [27] method.

In Figure 6, we generate a saliency map for a DAT subject with the ConvNeXt model to see what features the model learned are relevant for diagnosis. As illustrated, the model pays close attention to the temporal lobe (top/bottom sides in red), a region known for long-term memory recall and language. Moreover, the model also looks at the ventricles (the center of the brain in orange), an area that is known to be enlarged in DAT subjects.

We performed some additional experiments and summarize our results below.

Ablation/Data Studies (ConvNeXt only):

- Increasing model complexity from ~30M to ~80M resulted in no accuracy change.
- Increasing spatial resolution from 2mm (~100^3) to 1mm (~200^3) resulted in no accuracy change.
- Incorporating transfer learning using 11-class 3D CT organ classification (MedMNIST [28]) resulted in faster convergence but no increase in accuracy.
- Applying label smoothing to treat EMCI/LMCI classes with labels $[0.5, 0.5, 0]$ and $[0, 0.5, 0.5]$ for (CN,MCI,DAT) classes resulted in a 3% accuracy decrease for the 3-class validation set.

6. Conclusion

In conclusion, AD classification using sMRI data is a challenging problem due to the confounders and reliability of "ground truth" labels. Besides scaling up our dataset with more samples, given these issues, it would be interesting to explore the addition of other potential biomarkers (such as functional MRI [29]) into the classification system to further improve accuracy. In addition, exploring the "learning with noisy labels" research field [30] would be highly applicable to our setting. Some have worked on modifying the CE loss to be more robust against this effect [31], and not using clean labels at all during training [32], to provide a few examples.

References

[1] M. A. DeTure and D. W. Dickson, “The neuropathological diagnosis of Alzheimer’s disease,” *Molecular neurodegeneration*, vol. 14, no. 1, p. 32, 2019.

[2] M. S. Albert *et al.*, “The diagnosis of mild cognitive impairment due to Alzheimer’s disease: recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease,” *Focus*, vol. 11, no. 1, pp. 96–106, 2013.

[3] Z. Breijyeh and R. Karaman, “Comprehensive review on Alzheimer’s disease: causes and treatment,” *Molecules*, vol. 25, no. 24, p. 5789, 2020.

[4] T. Jiang *et al.*, “Multimodal magnetic resonance imaging for brain disorders: advances and perspectives,” *Brain Imaging and Behavior*, vol. 2, pp. 249–257, 2008.

[5] T. G. Beach, S. E. Monsell, L. E. Phillips, and W. Kukull, “Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005–2010,” *Journal of neuropathology and experimental neurology*, vol. 71, no. 4, pp. 266–273, 2012.

[6] M. Ávila-Villanueva, A. Marcos Dolado, J. Gómez-Ram\irez, and M. Fernández-Blázquez, “Brain structural and functional changes in cognitive impairment due to Alzheimer’s disease,” *Frontiers in psychology*, vol. 13, p. 886619, 2022.

[7] B. Magnin *et al.*, “Support vector machine-based classification of Alzheimer’s disease from whole-brain anatomical MRI,” *Neuroradiology*, vol. 51, pp. 73–83, 2009.

[8] A. Savio, M. Garc\ia-Sebastián, M. Graña, and J. Villanúa, “Results of an adaboost approach on Alzheimer’s disease detection on MRI,” in *International Work-Conference on the Interplay Between Natural and Artificial Computation*, 2009, pp. 114–123.

[9] L. V. Fulton, D. Dolezel, J. Harrop, Y. Yan, and C. P. Fulton, “Classification of Alzheimer’s disease with and without imagery using gradient boosted machines and ResNet-50,” *Brain sciences*, vol. 9, no. 9, p. 212, 2019.

[10] T. Akan, S. Alp, and M. A. N. Bhuiyan, “Vision Transformers and Bi-LSTM for Alzheimer's Disease Diagnosis from 3D MRI,” in *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, 2023, pp. 530–535.

[11] R. Hedayati, M. Khedmati, and M. Taghipour-Gorjikolaie, “Deep feature extraction method based on ensemble of convolutional auto encoders: Application to Alzheimer’s disease diagnosis,” *Biomedical Signal Processing and Control*, vol. 66, p. 102397, 2021.

[12] Q. Zuo, Y. Zhu, L. Lu, Z. Yang, Y. Li, and N. Zhang, “Fusing Structural and Functional Connectivities using Disentangled VAE for Detecting MCI,” in *International Conference on Brain Informatics*, 2023, pp. 3–13.

[13] R. C. Petersen *et al.*, “Alzheimer's disease Neuroimaging Initiative (ADNI) clinical characterization,” *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.

[14] K. J. Gorgolewski *et al.*, “The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments,” *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[15] O. Esteban *et al.*, “fMRIPrep: a robust preprocessing pipeline for functional MRI,” *Nature methods*, vol. 16, no. 1, pp. 111–116, 2019.

[16] F. Pérez-García, R. Sparks, and S. Ourselin, “TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning,” *Computer Methods and Programs in Biomedicine*, p. 106236, 2021, doi: https://doi.org/10.1016/j.cmpb.2021.106236.

[17] “https://learnopencv.com/variational-autoencoder-in-tensorflow/,” vol. 0, no. , p. , 2024.

[18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A convnet for the 2020s,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11976–11986.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[21] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.

[22] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016.

[23] I. O. Tolstikhin *et al.*, "Mlp-mixer: An all-mlp architecture for vision," *Advances in neural information processing systems*, vol. 34, pp. 24261–24272, 2021.

[24] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] H. Zhou *et al.*, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, pp. 11106–11115.

[27] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*, 2018, pp. 839–847.

[28] J. Yang *et al.*, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[29] R. Sperling, "The potential of functional MRI as a biomarker in early Alzheimer's disease," *Neurobiology of aging*, vol. 32, pp. S37–S43, 2011.

[30] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE transactions on neural networks and learning systems*, 2022.

[31] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 322–330.

[32] L. Jiang, D. Huang, M. Liu, and W. Yang, "Beyond synthetic noise: Deep learning on controlled noisy labels," in *International conference on machine learning*, 2020, pp. 4804–4815.