Self-Supervised Dementia Prediction From MRI Scans With Metadata Integration

Zile Huang *

October 13, 2023

Abstract

We introduce metadata integration in the training process for dementia diagnoses as weak label information using Weakly-Supervised Modified Knowledge Distillation with No Labels (WS-MDINO). Using WS-MDINO, we fine-tuned the parameters of the original vision transformer pre-trained with DINO on ImageNet. Our model achieved equivalent to the state-of-the-art epoura rformance of 92% accuracy in the OASIS1 dataset under leave-one-out cross-validation. We visualized the performance of the model by extracting average self-attention maps and average brains from the dataset, showing that the model had learned meaningful structural information about demented brains.

1 Introduction

Alzheimer's Disease (AD) is a leading cause of dementia, affecting millions worldwide. Even to date, it has no proper medical treatment and can only be controlled with continuous medication [KMS+22]. Early diagnoses and early intervention are beneficial for both the patients and caretakers, for the treatment would be most effective and less costly [RL19]. An automated model would aid the early detection of dementia immensely as it provides a fast, cheap, and accurate reference for the diagnosing process. Past works have used MRI scans of patients' brains to develop image recognition models for AD diagnoses [SN18, SMP+21, FDH+19, CGAA22, AR14, SJS+23, IZ18]. However, past models have faced challenges such as poor interpretability, which is a symptom of most deep learning and CNN architectures, and non-optimal integration of clinically free metadata [SN18]. Many previous works failed to perform cross validation because it is too computationally expensive (for each training split the model needed to be re-trained completely) [FDH+19,IZ18]. To address these limitations, we developed a model with a self-supervised method which can incorporate the metadata as weak labels [CZWM+22] with a vision transformer (ViT) backbone [DBK+20].

*Advised by: Jan Cross-Zamirski

While many previous works used the Convolutional Neural Network (CNN), we use a small ViT with 8×8 pixel patch size (ViT-S/8) introduced by Dosovitskiy *et al.* [DBK+20]. Compared to traditional CNN models, ViTs on medical datasets have been shown to capture long-range relationships in the image, provide built-in insight into the performance of the model with self-attention maps, and provide superior adaptive-learning with the self-attention mechanism [MHSS21].

Even though ViTs require a significantly larger dataset than CNNs to achieve these qualities [MHSS21], researchers can perform transfer learning from the pre-trained weights on ImageNet [DDS+09], which consists of millions of labeled images. Past work on automatic AD diagnoses using a ViT achieved an overall accuracy of 83.27%, with 85.07% specificity and 81.48% sensitivity on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [HKK23].

While the typical training methods for both CNNs and ViTs are supervised, this paper uses an unsupervised training approach, WS-MDINO, a modified version of the DINO [CTM+21] training method that integrates the metadata of the subjects into the training process . We trained a multi-perception classifier and a K-nearest-neighbor (KNN) classifier using the features extracted from the ViT to produce the final prediction.

Compared to other models, our ViT trained with the WS-MDINO method:

- Is a multi-modal model that integrates the clinically available metadata of the patients into the training process, achieving better overall performance
- Provides less noisy self-attention maps for the image data than supervised ViTs [CTM+21]
- Allows more complicated validation methods, such as K-fold and leave-one-out validation, without extra time and computing resources, because the training process of the feature extractor is not supervised, and, thus, does not require a train-test split.

2 Background

2.1 AD classification

There are many past efforts to use machine learning to diagnose AD in early stages based on MRI scans. An early work in automatic diagnoses used Structure Tensor Analysis to extract features from the MRI scan and used Support Vector Machine to classify stages of dementia [AR14]. It achieved 88.6% two-class (demented, non-demented) accuracy, 87.6% sensitivity, and 84.8% specificity. While this method required relatively less computational resources as it did not use neural network, it could not effectively integrate the clinically free metadata into the training process. It also lacked three-class classification. Later works used Convolutional Neural Network for the image recognition task. Fulton *et al.* [FDH+19] took the center 51 slices of the axial plane of the 3-D

image and trained a ResNet50 model for three-class (non-demented, very-mildly-demented, mildly-demented), achieving a 98.99% accuracy. However, this result is not convincing as it did not use a K-fold validation and it is likely that the slices of the same brain were assigned to both training and validation sets, causing data leakage. Using training-test split and 5-fold cross validation, we could not reproduce the results listed in the paper. Islam *et al.* [IZ18] trained three separate CNN models for each of the sagittal, coronal, and axial views of the brain, and combined the prediction of each model using vote. Their proposed model achieved 93% accuracy, 93% sensitivity, and 94% specificity. However, they failed to use N-fold validation as they considered it too computationally expensive, adding greater randomness to their performance. Newer studies introduced the ViT approach [ZK22], achieving 86% accuracy on ADNI dataset with convolutional voxel values as the input. Compared to CNN models, ViTs had better interpretability and could capture more long-range relations in the image.

A comprehensive review [WTSDM$^+$20] about machine learning models in AD classification presented the challenges faced in past classification works. It showed that many works only did a train-test set split and did not perform cross validation, making their performance less convincing. It also showed that many past works, such as the work we failed to reproduce [FDH$^+$19], suffered data leakage, knowingly or not, which caused inaccurate representation of models' performance. The review showed that many proposed performances were not reproducible and, in fact, if with proper train-set split and validation method, most proposed models would be outperformed by Support Vector Machine (SVM) with image score.

2.2 Machine Learning

2.2.1 Vision Transformer (ViT)

Inspired by transformers in Natural Language Processing (NLP), the ViT [DBK$^+$20] is a newer network architecture for computer vision. ViTs first split the inputted image into small patches (the original paper provided 8×8 pixel patches and 16×16pixel patches, but other dimensions are possible). Patches are then linearly projected to a flattened vector and, along with learnable class tokens, fed to the transformer encoder, which consists of multi-head attention layers and multi-layer-perception (MLP) layers. A normalization layer is added to each of the two main layers to improve performance and training efficiency. The multi-head attention layer consists of multiple self-attention heads, whose outputs are concatenated for the MLP layer. Each self-attention head can be visualized with a self-attention map. All the embedding are fed to a final MLP classifier for final classification. This structure is illustrated in Figure 1.

Compared to traditional CNN architectures, ViTs are more adaptive for image distortion and can capture long-range relations. However, this comes at a cost of heavy dependency on augmentations, hyper-parameter tuning, and large datasets [MHSS21]. For medical datasets, which can be relatively small,

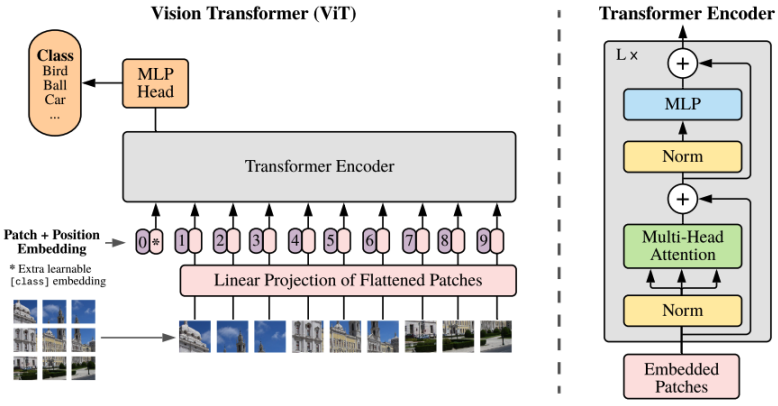

Figure 1: Vision Transformer architecture - figure from the original paper [DBK+20]

researchers primarily use or fine-tune ViTs pretrained on ImageNet [MHSS21].

2.2.2 Knowledge Distillation with No Labels (DINO)

Caron *et al.* proposed Knowledge Distillation with No Labels (DINO) as a self-supervised training scheme [CTM+21]. Similar to knowledge distillation, DINO trains two networks, the student network g_{θ_s} with parameters denoted as θ_s and the teacher network g_{θ_t} with parameters denoted as θ_t. DINO uses special data augmentation that, for each image x, generates 2 global crops covering large areas, denoted as x_1^g and x_2^g, and n local crops covering small areas, denoted as V (n is a hyper-parameter). DINO feeds the teacher network only global crops, and the student network global crops and local crops. DINO trains the student network to maximize its agreement with the teacher network by minimizing the Cross Entropy Loss:

$$Loss = \sum_{x \in \{x^{g,1}, x^{g,2}\}} \sum_{\substack{x' \in V \\ x' \neq x}} -P_t(x) log(P_s(x')) \tag{1}$$

Where $P(x)$ represents the probability distributions for the output, the Temperature Softmax:

$$P(x)^{(i)} = \frac{\exp\left(g(x)^{(i)}/\tau\right)}{\sum_{k=1}^{K} \exp\left(g(x)^{(k)}/\tau\right)} \tag{2}$$

Where K is the dimensionality of the output and τ is the temperature, different for student and teacher, denoted as τ_s and τ_y ($\tau > 0$). The teacher parameters are updated with an exponential moving average (ema) based on the student parameters:

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s \tag{3}$$

Where λ is the momentum hyper-parameter. While DINO also works with other architectures such as ResNet, it performs best with a ViT backbone. DINO with a ViT backbone presents clearer semantic segmentation information than supervised ViTs and works excellently with k-NN classifiers using extracted embeddings.

2.2.3 Weak Supervised form of DINO (WS-DINO)

Cross-Zamirski *et al.* proposed weak supervision during DINO (WS-DINO) training using weak labels on medical datasets which have clinically free metadata [CZWM+22]. WS-DINO first creates a pseudo class for each image using the metadata without using the real label. Different from DINO, WS-DINO then sources local views V from images of the same pseudo class. Minimizing the same loss function as Eq. 1, WS-DINO not only maximizes the agreement between the teacher and student networks, but also maximizes the agreement between images of the same pseudo class, therefore achieving weak supervision.

WS-DINO provides an elegant solution of integrating metadata into the training process. Therefore, it is especially powerful for datasets with relevant guiding metadata.

3 Methods

The implementation of our methods is available in a GitHub repository[1]. We summarize our training and evaluation in Figure 2.

3.1 Weakly-Supervised Modified DINO (WS-MDINO)

We propose Weakly-Supervised Modified DINO (WS-MDINO) training method for brain images. WS-MDINO is an adaptation of the WS-DINO method on brain images which allows the model to cluster subjects directly using weak labels while preserving important brain features such as symmetry and alignment.

Given the great structural similarity between aligned brain images, we consider random resized crop, the primary cropping method used in WS-DINO and DINO, unsuitable for brain image, as the model would interpret the augmented structural variation as a more significant factor than the actual information the images carry.

Therefore, WS-MDINO feeds the teacher network one global view of image x, denoted as v_t, and the student network n global view of n different images of the same pseudo class, denoted as V_s. It is trained to maximize the agreement between images of the same pseudo class by minimizing the Cross Entropy Loss:

[1] https://github.com/powerLEO101/WS-MDINO_OASIS1

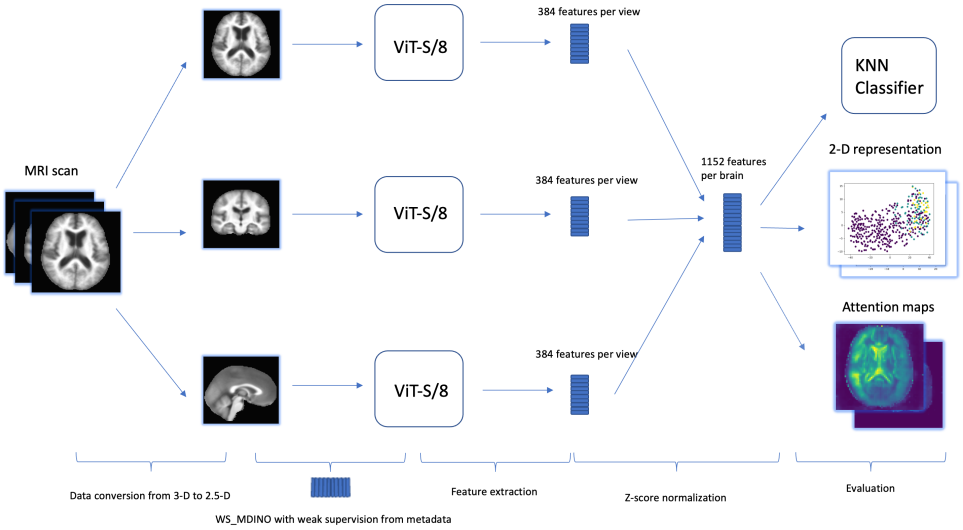

Figure 2: Summary of data preprocessing, training, feature extraction, and evaluation pipeline

Table 1: Caption

$$Loss = \sum_{x \in V_s} -P_t(v_t) log(P_s(x)) \tag{4}$$

3.2 Dataset analysis

We selected the OASIS1 dataset for our study [MWP+07]. The OASIS1 dataset provides a cross-sectional collection of 436 subjects aged from 18 to 96. For each subject, the dataset provides a $176 \times 208 \times 176$ pixel 3-D image of the MRI scan of each subject and a table of subjects' metadata. We present the details and completeness of the metadata in Table 2.

The CDR in the metadata is used as the ground-truth label, separating the image data into 4 classes: 0 being healthy, 0.5 being very mildly demented, 1 being mildly demented, and 2 being moderately demented.

Column Name	Data Completeness
Identification (ID)	Complete
Gender (M/F)	Complete
Dominant Hand (Hand)	Complete
Education (Educ)	Missing 201 rows
Socioeconomic Level (SES)	Missing 220 rows
Mini Mental State Examination Score (MMSE)	Missing 201 rows
Clinical Dementia Rating (CDR)	Complete
Estimated Total Intracranial Volume (eTIV)	Complete
Normalize Whole Brain Volume (nWBV)	Complete
Atlas Scaling Factor (ASF)	Complete
Delay	Missing 416 rows

Table 2: Summary of OASIS Data Completeness

3.3 Data preprocessing

In the dataset, there are 336 healthy subjects, 70 very mildly demented subjects, 28 mildly demented subjects, and 2 moderately demented subjects. We merged the mildly demented and moderately demented subjects into one class in-line with other studies [FDH+19, SMP+21].

Our model was mostly unaffected by the missing data except for the MMSE score. We noticed that all subjects without an MMSE score are non-demented. Therefore, we automatically gave them a full score of 30 for their MMSE score, signifying they are cognitively healthy [DPC17].

We created two kinds of pseudo class for each subject: 1) pseudo class only using the MMSE score and 2) pseudo class using a combination of the MMSE score and Age, the compound class. The detailed pseudo class can be found in the csv file in our Github Repository.

The raw images for subjects are atlas-registered gain field-corrected, brain masked, and re-sampled to 1mm isotropic voxels [MWP+07]. The dataset provides the processed file for this part of the preprocessing, which can be found in the "T88_111" folder in the dataset. We took one middle slice of the sagittal, coronal, and axial planes of the 3-D MRI image, converted them into arrays, and stored them in separate files for each subject.

By taking the center three slices of each plane of the brain, we converted the original 3-D images into 2.5-D images, which:

- Preserve the important features for diagnosing dementia (the ventricles and hippocampus area).
- Allow greater compatibility with the existing computer vision architectures and weights, such as ResNet and ViTs
- Have a much smaller size compared to the original dataset, demanding less computational resources and decreasing the training time, compared to 3-D models such as [SMP+21].

3.4 Data Augmentation

We used limited data augmentations to preserve important features of the brains. After several trials with various data augmentations such as resizing and translating, we concluded that, because each brain image is so structurally similar to another, the model would interpret the noise caused by the data augmentation a more significant information than the actual information the images carry. For this reason, we found that models would generally perform better on brain datasets like OASIS1 with little data augmentation.

Therefore, unlike the original DINO implementation[2], we avoided rotation and translations to preserve the symmetrical structure of the brain; We avoided color jitter, solarization, and Gaussian noise for the model to understand that the input is single-channel, even though the gray-scale channel is copied into RGB channels to fit the ViT structure and has a black background. For each global crop, we resized the image to 256×256 pixels and centered cropped the image to 224×224 pixels.

3.5 Network Details and Training

We trained separate models for each of the sagittal, coronal, and axial planes. For each model we used the same hyper-parameters as follows: a ViT-S/8 backbone; each augmented image data gives one 224×224 pixel global crop and three other 224×224 pixel global crops of images of the same pseudo class; teacher momentum is 0.99; gradient norm for gradient clipping is 3.0; teacher temperature is 0.04 without warm-up; student temperature is 0.07; center momentum is 0.8; batch size is 4; weight decay is none; optimizer is adamW; warm-up epoch is 10 epochs; learning-rate is $3e^{-6}$ to $2e^{-6}$ with a cosine scheduler; number of total epochs is 40; any other parameters are the same as the original DINO implementation. We initialized each model with the weights from DINO trained on ImageNet.

We extracted 384 features from the ViT head for each plane of view and combined them into a vector with 1152 elements. Finally we performed a Z-score normalization on the combined feature vector.

The model was trained with a INTEL I7-12700K GPU and NVIDIA RTX3080 GPU. The total training took approximately two hours.

3.6 Evaluation, visualization, and interpretation

To evaluate the features extracted by our model, we trained a KNN classifier (k=2) with leave-one-out cross validation. Leave-one-out cross validation is the logical extreme of cross validation and is the most unbiased. We evaluate the KNN classifier using 3-class accuracy, 2-class accuracy, sensitivity, and specificity.

To visualize the performance of our model, we performed a Principle Components Analysis (PCA) and reduced the dimensionality to 2. We then plotted

[2]https://github.com/facebookresearch/dino

Method	3-class Acc.	2-class Acc.	Sensitivity/Specificity
ResNet50	76.9%	79.6%	46%/87%
WS-MDINO with MMSE labels	84%	89%	64%/96%
WS-MDINO withCompound labels	**85%**	**92%**	**71%/98%**
WS-MDINO with Real labels (CDR)	100%	100%	100%/100%
CNN with vote [IZ18] [3]	N/A	93%	93%/94%
2-D CNN [SMP+21]	N/A	84%	N/A
3-D CNN [SMP+21]	N/A	84%	N/A
Forward Neural Network [JKK17]	N/A	90%	92%/87%

Table 3: Comparison of our work to existing works

each subject in a 2D space with 3 different colors representing non-demented, very mildly demented, and mildly demented and moderately demented in Figure 3, 4, and 5

To interpret our model, we extracted the self-attention maps for each subject. We then took the average pixel value of self-attention maps of subjects in each class and produced an average self-attention map for each class. Taking the average pixel value, we also evaluated an average brain for subjects in each class in Figure 6.

4 Results and discussion

4.1 Performance

We trained a ResNet50 on the same images set using 5-fold cross-validation as our baseline model, a WS-MDINO model using the real labels (CDR) as a proof of concept, a WS-MDINO model with the MMSE score as the pseudo class, and a WS-MDINO model with compound classes using both MMSE and age. We present our models' performances with other existing best performing models in Table 3.

We show that our model achieved the equivalent of state-of-the-art performance using compound labels under leave-one-out cross validation, a stricter metrics compared those of other works. It is worth noticing that, even though the WS-MDINO trained with real labels achieved 100% accuracy, it is not a valid network and only serves to be a proof of concept, because CDR is not a weak label and would cause data leakage.

4.2 Class representation

We used PCA to reduce the dimensionality to 2 and scattered subjects with three colors representing three classes (Purple = Non-demented, Mint = Very Mildly Demented, Yellow = Mildly to Moderately Demented). We present the class representation plots for three stages of the fine-tuning (no fine-tuning, 20 epochs of fine-tuning, and 40 epochs of fine-tuning) to show the clustering

process under weak supervision in Figure 3, 4, and 5. We show that the weak supervision with weak labels is effective as subjects cluster over time in the 2-D representation. It is worth noticing that, even though we used 7 weak labels in total, our model clustered subjects into roughly 4 clusters. This shows that, while our model learns from the weak labels, it also effectively learns from the similarities between images of the same class.

4.3 Average self-attention maps and brains

The self-attention map is a powerful feature of ViTs which visualizes the weights of the self-attention heads. Such visualization is very useful as it shows the features that the model has learned, which simplify the fine-tuning process and allow the researchers to interpret the model. We present average self-attention maps from the fine-tuned weights and average brain for each of the sagittal, coronal, and axial views in Figure. 6.

Using average self-attention maps and brains, we show that our model has learned meaningful information from brains with different stages of dementia. In general, demented subjects have shrunken hippocampus and cerebral cortex, and enlarged ventricles [IZ18]. In Figure 6, we show that our model successfully captured the aforementioned features for demented subjects. From the sagittal and coronal views, we show that the area around the hippocampus is the most highlighted by the attention map. From the axial view, we show that the ventricle is the most highlighted and the area around it, the cerebral cortex, is also more highlighted than other structures. Thus, we show that our model has learned the significance of the hippocampus, cerebral cortex, and ventricles in the diagnoses of dementia. It is also notable that the highlighted areas for demented subjects are generally dimmer than those for non-demented subjects, signifying our model has learned the structural difference between demented and non-demented brains.

5 Conclusion

WS-MDINO is a powerful method of integrating metadata as weak supervision for DINO training, allowing models to learn effectively from both images and metadata. Capable of generating 3-class and 2-class predictions with high accuracy, our dementia diagnosis model trained using WS-MDINO with compound weak labels successfully captures important features of demented and non-demented brain. Our model also provides insight into weakly supervised training methods for datasets that are sensitive to data augmentations, such as brain MRI scan datasets.

While the OASIS1 dataset used in this study is a relatively small dataset, there are larger datasets for dementia diagnoses such as ADNI[4]which consists of thousands of subjects. Future work should encompass testing our method on such larger datasets. However, this is beyond the scope of this study. It is also

[4]https://adni.loni.usc.edu/

possible that a stronger pseudo class could further improve the model's performance. However, it is important that the pseudo classes do not have dataset-specific information, which decreases the model's generalization ability. Thus, we suggest building pseudo classes as simple as possible and following existing studies on metadata's influence on the subject, such as [RSH+13]. For future work, WS-MDINO has the potential to seamlessly combine machine learning approaches with classical approaches, which are reflected in the creation of one or multiple pseudo classes.

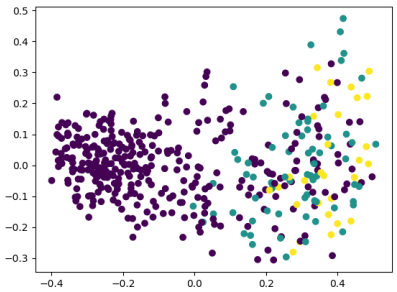

Figure 3: Class representation in 2-D space using ImageNet weights with no fine-tuning

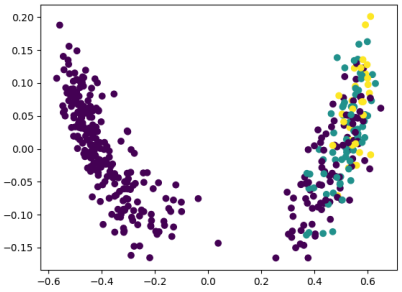

Figure 4: Class representation in 2-D space using ImageNet weights fine-tuned with WS-MDINO with compound label after 20 epochs

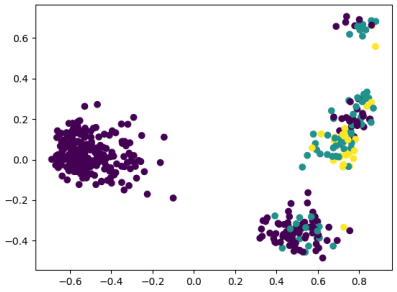

Figure 5: Class representation in 2-D space using ImageNet weights fine-tuned with WS-MDINO with compound label after 40 epochs

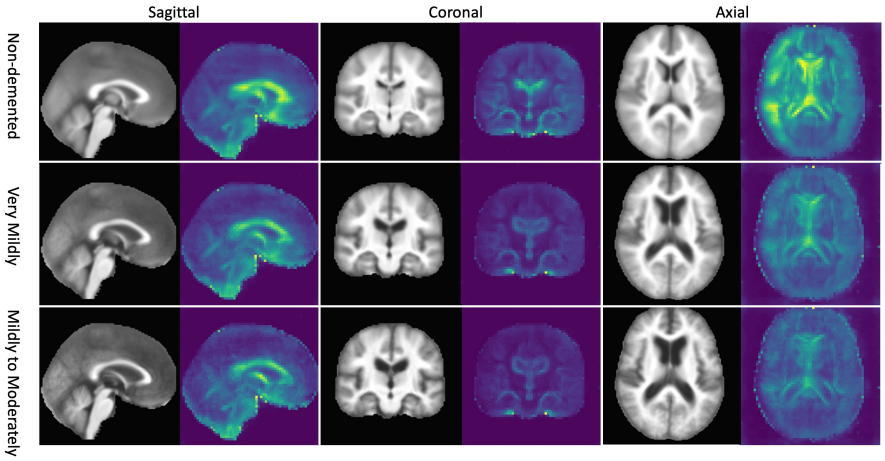

Figure 6: Average attention maps and brains from sagittal, coronal, and axial view, produced from WS-MDINO using compound labels (left: average brain; right: average self-attention map)

References

[AR14] M Archana and S Ramakrishnan. Detection of alzheimer disease in mr images using structure tensor. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1043–1046, 2014.

[CGAA22] Kwok Tai Chui, Brij B. Gupta, Wadee Alhalabi, and Fatma Salih Alzahrani. An MRI scans-based alzheimer's disease detection via convolutional neural network and transfer learning. *Diagnostics*, 12(7):1531, June 2022.

[CTM+21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[CZWM+22] Jan Oscar Cross-Zamirski, Guy Williams, Elizabeth Mouchet, Carola-Bibiane Schönlieb, Riku Turkki, and Yinhai Wang. Self-supervised learning of phenotypic representations from cell images with weak labels, 2022.

[DBK+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2009.

[DPC17] Silvia Duong, Tejal Patel, and Feng Chang. Dementia. *Canadian Pharmacists Journal / Revue des Pharmaciens du Canada*, 150(2):118–129, February 2017.

[FDH+19] Lawrence Fulton, Diane Dolezel, Jordan Harrop, Yan Yan, and Christopher Fulton. Classification of alzheimer's disease with and without imagery using gradient boosted machines and ResNet-50. *Brain Sciences*, 9(9):212, August 2019.

[Gre93] George D. Greenwade. The Comprehensive Tex Archive Network (CTAN). *TUGBoat*, 14(3):342–351, 1993.

[HKK23] Gia Minh Hoang, Ue-Hwan Kim, and Jae Gwan Kim. Vision transformers for the prediction of mild cognitive impairment to alzheimer's disease progression using mid-sagittal sMRI. *Frontiers in Aging Neuroscience*, 15, April 2023.

[IZ18] Jyoti Islam and Yanqing Zhang. Brain MRI analysis for alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. *Brain Informatics*, 5(2), May 2018.

[JKK17] Debesh Jha, Ji-In Kim, and Goo-Rak Kwon. Diagnosis of alzheimer's disease using dual-tree complex wavelet transform, PCA, and feed-forward neural network. *Journal of Healthcare Engineering*, 2017:1–13, 2017.

[KMS+22] C. Kavitha, Vinodhini Mani, S. R. Srividhya, Osamah Ibrahim Khalaf, and Carlos Andrés Tavera Romero. Early-stage alzheimer's disease prediction using machine learning models. *Frontiers in Public Health*, 10, March 2022.

[MHSS21] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images?, 2021.

[MWP+07] Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, September 2007.

[RL19] Jill Rasmussen and Haya Langerman. Alzheimer's disease why we need early diagnosis. *Degenerative Neurological and Neuromuscular Disease*, Volume 9:123–130, December 2019.

[RSH+13] Tom C. Russ, Emmanuel Stamatakis, Mark Hamer, John M. Starr, Mika Kivimäki, and G. David Batty. Socioeconomic status as a risk factor for dementia death: individual participant meta-analysis of 86 508 men and women from the UK. *British Journal of Psychiatry*, 203(1):10–17, July 2013.

[SJS+23] Hyunji Shin, Soomin Jeon, Youngsoo Seol, Sangjin Kim, and Doyoung Kang. Vision transformer approach for classification of alzheimer's disease using 18f-florbetaben brain images. *Applied Sciences*, 13(6):3453, March 2023.

[SMP+21] Cristina L. Saratxaga, Iratxe Moya, Artzai Picón, Marina Acosta, Aitor Moreno-Fernandez de Leceta, Estibaliz Garrote, and Arantza Bereciartua-Perez. MRI deep learning-based solution for alzheimer's disease prediction. *Journal of Personalized Medicine*, 11(9):902, September 2021.

[SN18] Lauge Sørensen and Mads Nielsen. Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination. *Journal of Neuroscience Methods*, 302:66–74, May 2018.

[WTSDM+20] Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, and Olivier Colliot. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, 63:101694, July 2020.

[ZK22] Zilun Zhang and Farzad Khalvati. Introducing vision transformer for alzheimer's disease classification task with 3d input, 2022.