FLIQA-AD: a Fusion Model with Large Language Model for Better Diagnose and MMSE Prediction of Alzheimer's Disease

Junhao Chen[1], Zhiyuan Ding[2], Xiangzhu Zeng[3], Yan Liu[4], Ling Wang[1*]**
[1] University of Electronic Science and Technology of China, Chengdu, China
[2] Johns Hopkins University, Baltimore, USA
[4] Peking University Third Hospital, Beijing, China
[3] University of Chinese Academy of Sciences, Beijing, China
Correspondence: eewangling@uestc.edu.cn, yanliu@ucas.ac.cn

Abstract

Tracking a patient's cognitive status early in the onset of the disease provides an opportunity to diagnose and intervene in Alzheimer's disease (AD). However, relying solely on magnetic resonance imaging (MRI) images with traditional classification and regression models may not fully extract finer-grained information. This study proposes a multi-task Fusion Language Image Question Answering model (FLIQA-AD) to perform AD identification and Mini Mental State Examination (MMSE) prediction. Specifically, a 3D Adapter is introduced in Vision Transformer (ViT) model for image feature extraction. The patient electronic health records (EHR) information and questions related to the disease work as text prompts to be encoded. Then, an AD-Former model, which combines self-attention and cross-attention mechanisms, is used to capture the correlation between EHR information and structure features. After that, the extracted brain structural information and textual content are combined as input sequences for the large language model (LLM) to identify AD and predict the corresponding MMSE score. Experimental results demonstrate the strong discrimination and MMSE prediction performance of the model, as well as question-answer capabilities. [1]

1 Introduction

Alzheimer's disease (AD) is one of the most common forms of dementia. It takes several years from the onset of normal cognition (NC) to AD, so it provides an opportunity for early diagnosis and intervention. The Mini-Mental State Examination (MMSE) is a widely used cognitive assessment tool for evaluating the progression of cognitive and behavioral states. Alternatively, magnetic resonance images (MRI) can obtain more detailed structural changes, such as the presence of senile plaques (SP) and atrophy of the cerebral cortex (Duc et al., 2020). AD identification and MMSE score are interrelated, which underscores the necessity of combining MRI and other non-imaging data for dementia analysis (Qiu et al., 2018).

Therefore, some researchers have introduced multi-task learning to predict MMSE and detect AD jointly. For instance, in (Liu et al., 2021) an interaction module is designed to connect the shared features to the tasks. To include the demographic text information, a deep multi-task multi-channel learning (DM^{2}L) framework is proposed for classification and regression (Liu et al., 2018). To solve the task relevance issue, feature relevance is exploited by adding three multi-task interaction layers between two task backbones (Tian et al., 2022). However, such work tends to perform better on AD identification or MMSE score prediction tasks exclusively, and a decline in performance is observed on multi-target tasks. Using an additional interaction module for interacting still requires extracting features for different tasks. Simply designing multiple interaction layers without incorporating any electronic health records (EHR) prompts information will not assess early-stage AD effectively due to ignoring demographic characteristics.

In recent years, the vision language pre-trained (VLP) model has provided a better reference for solving the above challenges. For example, CLIP (Radford et al., 2021) learns representations from natural language supervision and performs well for zero-shot transfer to various downstream tasks. BLIP-2 (Li et al., 2023) uses an efficient pretraining strategy that freezes the visual encoder and large language model. The modal gap is bridged by training the Q-former. LLaVA (Liu et al., 2024) trains a projection layer to connect the frozen visual encoder and large language model (LLM), with better zero-shot capabilities.

Inspired by these works, in this study, we pro-

[1] The code is following: https://github.com/junhao667/FLIQA-AD.git

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 587–594
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

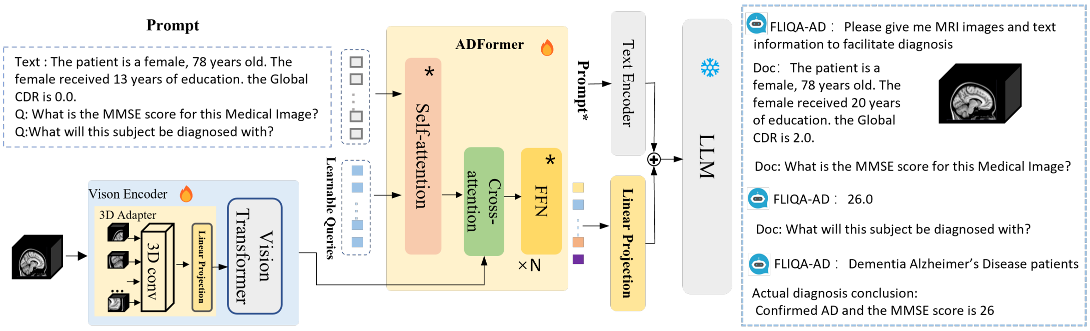

Prompt
Text : The patient is a female, 78 years old. The female received 13 years of education. the Global CDR is 0.0.
Q: What is the MMSE score for this Medical Image?
Q:What will this subject be diagnosed with?
Vison Encoder
3D Adapter
3D conv
Linear Projection
Vision Transformer
Learnable Queries
ADFormer
*
Self-attention
Cross-attention
FFN
×N
Prompt*
Text Encoder
Linear Projection
LLM
FLIQA-AD : Please give me MRI images and text information to facilitate diagnosis
Doc: The patient is a female, 78 years old. The female received 20 years of education. the Global CDR is 2.0.
Doc: What is the MMSE score for this Medical Image?
FLIQA-AD : 26.0
Doc: What will this subject be diagnosed with?
FLIQA-AD : Dementia Alzheimer's Disease patients
Actual diagnosis conclusion:
Confirmed AD and the MMSE score is 26

Figure 1: The framework of our proposed method. The text prompts and images are encoded. After that, we can obtain the query output from ADFormer. The text we input into the model is denoted as prompt*. The demonstration is shown on the right.

pose an AD MRI diagnostor, FLIQA-AD, for better diagnosis and prediction of MMSE. Specifically, the diagnostor is constructed by the vision encoder module, ADFormer fusion module, and LLM module as shown in Fig.1. In the vision encoder module, a 3D Adapter is used to convert 3D images into processable tokens, preserving the spatial structure information of the images. Then, we utilize the bio-ClinicalBERT model, which has been pre-trained on specialized diagnostic question-answering texts (Alsentzer et al., 2019), as the text encoder. The patient's EHR information and questions related to the disease will be used as text prompts. To extract the most diagnostically beneficial visual features from different types of patients, ADFormer is proposed to fuse the EHR information and vision features through a cross-attention manner. Finally, LLM is used as a decoder that outputs AD detection and MMSE scores from the text and visual features input.

2 Method

2.1 3D Adapter

Since patients have different global and localized presentations, both global and local structural information is important for classification and regression tasks. So, a 3D adapter is used to project the image patch into the embedding space while also capturing the local structural information inside the patch before inputting. Let **I** of size (H, W, D) be the input MRI image, the patch size of each MRI volume image is (P, P, P), then the total number of patches is $N_p = HWD/P^3$. These patches serve as the effective input sequence length for the Vision Transformer (ViT) (Dosovitskiy et al., 2021). Since the embedding dimension of all transformer layers is uniformly D, we use a learnable linear projection layer that projects each sequence into the D-dimensional space. Then the input embedding is (B, N_p, D), where B is the batch size.

2.2 ADFormer

Personal information from EHRs (gender, age, education level, etc.) is related to brain states, and taking this non-MRI structural information into account can influence AD diagnosis and MMSE prediction results (Koga et al., 2002; Liu et al., 2017; Ding et al., 2009). Therefore, we propose the ADFormer, which fuses this textual information with MRI structural information through the cross-attention layer. To encode the EHR information, (Alsentzer et al., 2019) is used, which was trained on a large corpus of medical texts, including PubMed and MIMIC-III, and the EHRs of patients in the intensive care unit (ICU). We introduced this text encoder into our ADFormer and fine-tuned it so that it could relearn relying on already existing basic medical knowledge without a mass of data.

Let the input image-text feature pair be $\{v_n, t_n\}_{n=1}^{N_s}$, where N_s is the number of samples, v_n is the visual features extracted by ViT, t_n is text. Textual information t_n is fed into the model via self-attention blocks, which are parameterized and trained in medical text based on bio-Clinicalbert (Alsentzer et al., 2019). Queries interact with the visual features through a cross-attention module to extract the most effective visual features by combining the existing knowledge. The cross-attention module is subsequently followed by the feed-forward neural (FFN) network, which is also trained in the medical literature. To maintain the abundant detailed information inherent in high-resolution 3D medical images, we avoid downsampling and cropping operations. Our visual input features are greatly reduced in the order of magnitude of the features from the visual features obtained from the original ViT $(344, 1408)$ to the final $(32, 768)$.

2.3 Question Answering Decoder

To detect AD and MMSE prediction, we use the fine-tuning-based FLANT5 (Chung et al., 2024) as a language model. Each task we consider (including regression prediction, classification, Q&A, etc.) can be treated as text models and trained together to reach the final target. For the classification task, the model can only predict a single word corresponding to the target. The prediction remains the basic paradigm of language modeling, i.e., the new token is related to both the input and the previous prediction tokens (Raffel et al., 2020).

Let $t = \{t_1, t_2, ..., t_i\}$ be the input text sequence, where i is the length of the text token, $q = \{q_1, q_2, ..., q_n\}$ denotes the sequence of ADFormer output, n denotes the number of learnable Queries, and $a = \{a_1, ..., a_j\}$ is the previous prediction, where j is the tokens of the previous output. we compute language generation loss L_{LG}:

$$L_{LG} = -\sum_{j=1}^{T} \log \mathrm{P}(a_j \mid t_1, \ldots, t_i, q_1, \ldots, q_n, \\ a_1, \ldots, a_{j-1}). \quad (1)$$

Assuming that E_{vit} denotes the vision encoder, Q denotes the learnable queries from ADFormer, the feature extracted by ADFormer is formulated as:

$$q_D = Q(E_{vit}(\mathbf{I})), \quad (2)$$

To match the dimensions of query and LLM. We first project the original features of ADFormer query output q_D to the embedding space of LLM by a learnable projection f:

$$q = f(q_D), \quad (3)$$

Finally, the input of the LLM model is formulated as the concatenation of t and q.

2.4 Training Objective

To align image and text representations, it is necessary to maximize their mutual information. We also feed questions with text into ADFormer to perform image-text contrast learning. Specifically, question tokens as one of the inputs interact with the query through the self-attention layer, which directs the ADFormer's cross-attention layer to focus on the more informative image regions. Therefore, the contrastive learning loss is formulated as:

$$L_{I \leftrightarrow T} = CrossEntropy(I_f, T_f), \quad (4)$$

where I_f denotes visual features. The text and question feature is T_f. $L_{I \leftrightarrow T}$ denotes the contrast loss between the image I and text-question T.

Furthermore, for the supervised task, we also introduce the image and result comparison loss as:

$$L_{I \leftrightarrow \hat{P}} = CrossEntropy(I_f, \hat{P}), \quad (5)$$

where $\hat{P}$ denotes the target of the prediction. And the final loss function is formulated as:

$$L_{total} = L_{I \leftrightarrow T} + L_{I \leftrightarrow \hat{P}} + L_{LG}. \quad (6)$$

3 Experiments

3.1 Data and preprocessing

We use the ADNI (Petersen et al., 2010) and OASIS (Marcus et al., 2007) datasets to validate our approach. The volume images of MRI T1 were collected as samples, the statistics of the data information are shown in Table 1. All the images of ADNI are officially pre-processed: Gradwrap Correction, B1 Non-Uniformity Correction and N3 Non-Uniformity Correction. The FMRIB Software Library (FSL) software (Jenkinson et al., 2012) was used to register the original images to the MNI152 standard template. Textual information, including age, MMSE, education level, CDR score, etc. was extracted to construct input text.

Data	Image	Group (AD/MCI/NC)	Gender (M/F)	Age	MMSE (Mean)
ADNI	8315	2613/3667/2035	4024/2808	55–93	5–30 (26.3)
OASIS	373	146/-/227	160/213	60–98	4–30 (27.3)

Table 1: Data details, AD, MCI, and NC within the "group" category represent Alzheimer's Disease, Mild Cognitive Impairment, and Normal Control, respectively.

3.2 Experimental Setting Detail

We randomly sampled 300 of each category (AD, MCI, NC) by patient level from ADNI to form a testing set, and the remaining 7380 samples from ADNI were used as training and validation sets. Multiple visits of the same subject are treated as separate images. The validation set consists of 300 randomly selected image samples from each category. All 373 samples from OASIS-2 were used for zero-shot tests.

Method		Multi-class Disease Identification						MMSE Prediction		
		ACC_{AD}	ACC_{MCI}	ACC_{NC}	ACC	AUC	$Kappa$	$RMSE$	R^2	CC
Single-task	**MedBLIP (T5)**	0.71	0.94	0.91	0.85	0.89	0.77	2.44	0.62	0.80
	ViT+MLP	0.83	0.94	0.96	0.91	0.93	0.88	2.41	0.63	0.80
	ViT+Qformer+MLP	0.83	0.97	0.95	0.92	0.94	0.87	2.21	0.69	0.87
Multi-task	**LLaVA-Med(7b)**	0.87	0.55	0.96	0.72	0.79	0.57	3.01	0.42	0.72
	BLIP-2 (T5)	0.83	**0.99**	0.95	0.92	0.94	0.88	5.05	-0.63	0.26
	Ours	**0.94**	0.98	**0.99**	**0.97**	**0.98**	**0.96**	**1.25**	**0.90**	**0.95**

Table 2: Performance comparison of AD/MCI/NC classification and MMSE prediction on single-task and multi-task.

In the vision encoding process, the input registered images are uniformly resized to 126 × 126 × 126, the patch size is set to 18, and each volume is eventually divided into 343 patches. Finally, the size of 344 × 1408 (with class token preserved) is passed through the ViT. The visual encoder uses the EVA_CLIP (Fang et al., 2023) model that can be efficiently fine-tuned. The language model FLANT5 (T5) is used for text encoding.

The fusion model ADFormer, with 32 learnable queries and the last hidden layer is used as the final output features. AdamW is used as the optimizer, and the learning rate is dynamically adjusted using WarmupCosine. The initial learning rate is set to be 2e-5, the batch size is 8, and all experiments were performed on a single A100 × 40G GPU.

3.3 Performance of Our Proposed Method

In this study, the same data and computational resources were used to train the model ViT (Dosovitskiy et al., 2021). We also fine-tuned the multimodal model as comparison. The BLIP2 (Li et al., 2023) model was fine-tuned with T5, the 3D Adapter structure was added to the ViT and fine-tuned for 3D image processing. We also train and fine-tune the MedBLIP (Chen and Hong, 2024) model with T5 on our dataset, and fine-tune the LLaVA model following the LLaVA-Med(Li et al., 2024).

For the classification, we use accuracy (ACC), Area Under the ROC curve (AUC), and the $Kappa$ coefficient which can assess the concordance between model predictions and the truth. For the MMSE prediction, we utilize the Square root of the mean ($RMSE$), the coefficient of determination (R^2) as a statistical measure of explained variability, and Pearson's correlation coefficient (CC) to reflect the alignment trend and linearity of the predictor for evaluating the performance of the proposed method (Liu et al., 2021). For further details regarding the parameters can be found in the Appendix A

The results are shown in Table 2, which shows that our model outperforms all the other approaches except MCI accuracy. The identification accuracy and Pearson correlation coefficient are reached to 97% and 95%.

To evaluate the generalization ability of the model, we test the zero-shot performance on OASIS. The results are shown in Table 3. We can find that the performances of AD identification and MMSE prediction of the models that use image-text fusion techniques are much better than those of using only image information (ViT+MLP) or simple contrast learning method (MedBLIP).

Method	AD vs NC	MMSE Prediction		
	ACC	$RMSE$	R^2	CC
MedBLIP (T5)	0.20	4.56	-0.53	0.22
ViT+MLP	0.25	4.31	-0.38	0.05
ViT+Qformer+MLP	0.69	**3.20**	0.23	0.55
LLaVA-Med(7b)	0.50	3.45	**0.57**	0.26
BLIP-2 (T5)	0.64	3.51	0.09	**0.58**
Ours	**0.81**	3.60	0.25	0.56

Table 3: Zero-shot identification performance on OASIS

3.4 Ablation Study

In this experiment, we explore the effectiveness of the proposed method. To be fair, we follow the previous experimental setup of the data division strategy and move out the ADFormer module, LLM module, and T5 respectively. Each module is replaced by a simple multi-layer perceptron (MLP). We also examined the impact of prompts on LLM and Adformer. The ablation results are shown in Table 4. It illustrates that there is progressive 6% improvement in accuracy, and 15% improvement in Pearson correlation coefficient with ADFormer and LLM. When ADFormer and LLM respectively discarded the EHR information as text prompt input, all indicators dropped significantly, for example, ACC dropped by 18%.

Component	Multi-class Disease Identification				MMSE Prediction		
	ACC$_{AD}$	ACC$_{MCI}$	ACC$_{NC}$	ACC	RMSE	R^2	CC
w/o T5&ADFormer	0.83	0.94	0.96	0.91	2.41	0.63	0.80
w/o ADFormer	0.87	0.96	0.97	0.93	1.89	0.77	0.88
w/o T5	0.88	0.97	0.97	0.94	1.99	0.74	0.90
ADFormer w/o prompt	0.62	0.91	0.83	0.79	4.82	-0.48	0.31
LLM w/o prompt	0.86	0.93	0.95	0.91	2.72	0.53	0.75
Ours	**0.94**	**0.98**	**0.99**	**0.97**	**1.25**	**0.90**	**0.95**

Table 4: Comparison of different components of our models

3.5 Interpretability Analysis

In the medical diagnostic, the MMSE score of AD, MCI and NC usually be clinically categorized into a range of values. In this experiment, we also plot the predicted MMSE scores with true values on ADNI test data, the results are shown in Appendix. B Fig.(2a)- (2c), where the ranges are also marked out. The overlap between predicted and true values of AD, MCI and NC are 78.3%, 89.3% and 86% respectively. We also demonstrate the efficiency and interpretability of our feature fusion module ADFormer by t-SNE feature downscaling on the ADNI dataset. As shown in Appendix. B Fig. (2d), the extracted features have reliable category separation, with clusters of data points in each category more clearly separated from the others.

4 Conclusion

In this work, to better identify AD and predict MMSE, we propose a fusion model ADFormer to interact with the patient's EHR information and MRI images. 3D Adapter extracted local features from 3D MRI images, which are divided into blocks and projected into ViT embedding space to extract visual representations. Subsequently, the patient EHR information and questions, along with visual features are fused through the self-attention and cross-attention blocks in the ADFormer module. LLM is used to help reasoning. The model responds with the corresponding category or MMSE score according to the specific question. The model also illustrates outstanding performance on zero-shot identification tasks, and the experiment results show state-of-the-art performance on large datasets. In the follow-up work, we will work on improving the model's ability to respond to open medical questions and its zero-shot capability.

5 Limitation

This paper proposes a FLIQA-AD model based on EHR information and MRI images to diagnose AD. However, in the medical domain, especially Alzheimer's disease, text and image information is extremely scarce due to privacy protection and other issues, and the amount compiled in this paper is limited, which greatly limits the open question-answering ability of this model.

Secondly, the model is pre-trained on the ADNI dataset. When it is transferred to the OASIS dataset, although we have performed a series of preprocessing to keep the basic features of the image consistent, the performance on OASIS has declined due to differences in information such as image resolution. In the experiments in this paper, we found that increasing the trainable dataset can improve the model's ability on image datasets that are significantly different from the training set. That may be the thing worth trying in the future.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Qiuhui Chen and Yi Hong. 2024. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. In *Proceedings of the Asian Conference on Computer Vision*, pages 2404–2420.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Bei Ding, Ke-Min Chen, Hua-Wei Ling, Fei Sun, Xia Li, Tao Wan, Wei-Min Chai, Huan Zhang, Ying Zhan, and Yong-Jing Guan. 2009. Correlation of iron in the hippocampus with mmse in patients with alzheimer's disease. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 29(4):793–798.

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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Nguyen Thanh Duc, Seungjun Ryu, Muhammad Naveed Iqbal Qureshi, Min Choi, Kun Ho Lee, and Boreom Lee. 2020. 3d-deep learning based automatic diagnosis of alzheimer's disease with joint mmse prediction using resting-state fmri. *Neuroinformatics*, 18:71–86.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. 2012. Fsl. *Neuroimage*, 62(2):782–790.

H Koga, T Yuzuriha, H Yao, K Endo, S Hiejima, Y Takashima, F Sadanaga, T Matsumoto, A Uchino, K Ogomori, et al. 2002. Quantitative mri findings and cognitive impairment among community dwelling elderly subjects. *Journal of Neurology, Neurosurgery & Psychiatry*, 72(6):737–741.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jin Liu, Xu Tian, Jianxin Wang, Rui Guo, and Hulin Kuang. 2021. Mtfil-net: automated alzheimer's disease detection and mmse score prediction based on feature interactive learning. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1002–1007. IEEE.

Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. 2017. Deep multi-task multi-channel learning for joint classification and regression of brain status. In *International conference on medical image computing and computer-assisted intervention*, pages 3–11. Springer.

Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. 2018. Joint classification and regression via deep multi-task multi-channel learning for alzheimer's disease diagnosis. *IEEE Transactions on Biomedical Engineering*, 66(5):1195–1206.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. 2007. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, non-demented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507.

Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, CR Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. 2010. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3):201–209.

Shangran Qiu, Gary H Chang, Marcello Panagia, Deepa M Gopal, Rhoda Au, and Vijaya B Kolachalama. 2018. Fusion of deep learning models of mri scans, mini–mental state examination, and logical memory test enhances diagnosis of mild cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:737–749.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Xu Tian, Jin Liu, Hulin Kuang, Yu Sheng, Jianxin Wang, and The Alzheimer's Disease Neuroimaging Initiative. 2022. Mri-based multi-task decoupling learning for alzheimer's disease detection and mmse score prediction: A multi-site validation. *arXiv preprint arXiv:2204.01708*.

A Evaluation parameters

A.1 Classification Metrics

A.1.1 Accuracy (ACC)

The Accuracy (ACC) quantifies the direct classification performance by measuring the ratio of correctly classified instances to the total instances, as defined in Equation 7:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{7}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

A.1.2 Area Under the ROC Curve (AUC)

The AUC (Area Under the ROC Curve) measures a model's ability to distinguish between positive and negative classes by plotting TPR (True Positive Rate) against FPR (False Positive Rate) at various thresholds. A higher AUC signifies better classification performance

- **True Positive Rate (TPR)**: Defined as the proportion of true positive instances among all actual positives (Equation 8), it reflects the model's sensitivity:

$$TPR = \frac{TP}{TP + FN}. \tag{8}$$

- **False Positive Rate (FPR)**: Represents the proportion of false positives among all actual negatives (Equation 9):

$$FPR = \frac{FP}{FP + TN}. \tag{9}$$

A.1.3 Cohen's Kappa Coefficient

Cohen's Kappa (κ) assesses the agreement between model predictions and ground-truth labels while accounting for chance agreement, providing a robust alternative to accuracy in imbalanced datasets. It is computed as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{10}$$

where P_o is the observed agreement ratio, and P_e denotes the probability of random agreement.

A.2 Regression Metrics

A.2.1 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) quantifies the average deviation between predicted and true values, emphasizing larger errors due to its quadratic nature (Equation 11):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \tag{11}$$

where y_i and $\hat{y}_i$ represent the true and predicted values of the i-th sample, and n is the total sample size.

A.2.2 Coefficient of Determination (R^2)

The Coefficient of Determination (R^2) measures the proportion of variance in the dependent variable explained by the model, serving as a critical indicator of goodness-of-fit (Equation 12):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{12}$$

where $\bar{y}$ denotes the mean of the true values.

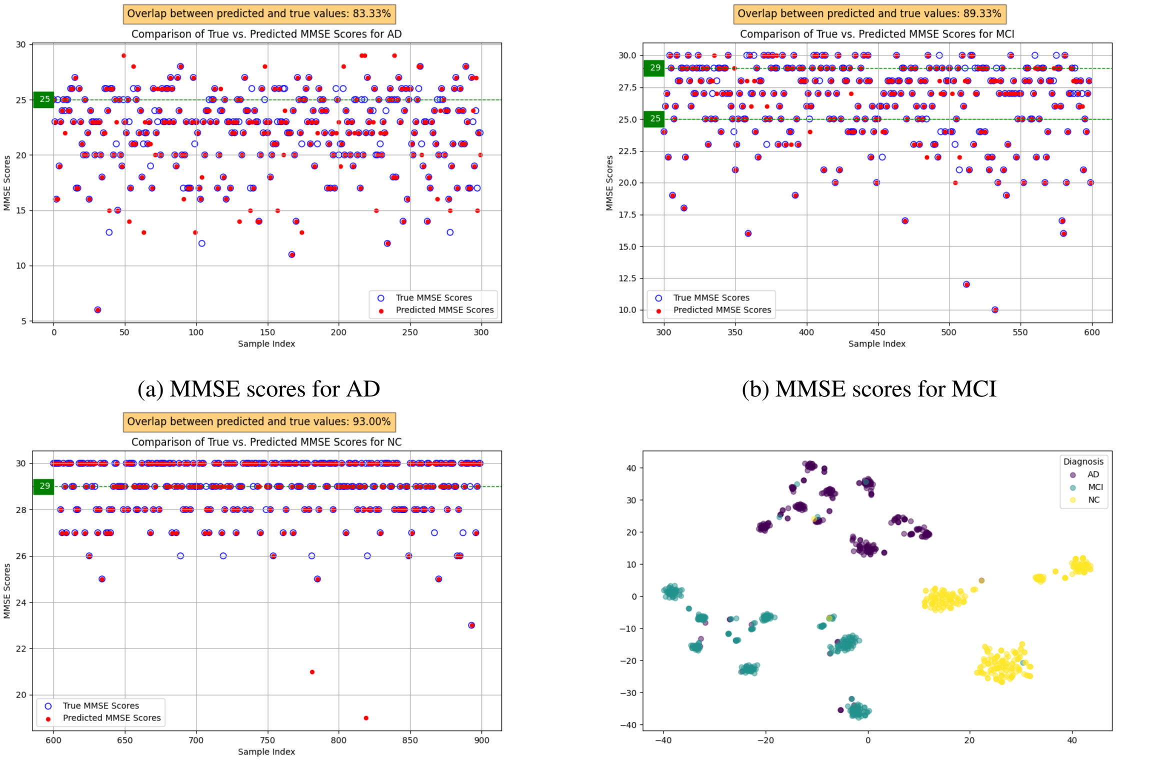
A.2.3 Pearson's Correlation Coefficient (CC)

Pearson's Correlation Coefficient (CC) evaluates the linear relationship between predicted and true values. For MMSE score prediction, it is utilized to investigate the alignment trend between model outputs and clinical observations (Equation 13):

$$CC = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) \left(\hat{y}_i - \bar{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} \left(\hat{y}_i - \bar{\hat{y}}\right)^2}}, \tag{13}$$

where $\bar{\hat{y}}$ represents the mean of predicted values.

B Interpretability Analysis



(a) MMSE scores for AD

(b) MMSE scores for MCI

(c) MMSE scores for NC

(d) Embeddings visualization of ADFormer output features

Figure 2: (a)- (c) are the MMSE score against the predicted and the true value for AD, MCI and NC. The green dashed lines in the plots represent the approximate range of MMSE scores for each category (MMSE<25 for AD, MMSE>29 for NC, and in between for MCI). (d) is a visualization of the output features of ADFormer.