**Summary of Changes**

Manuscript "Ms. No. MEDIA-D-21-01329: Reducing Variations in Multi-center Alzheimer's Disease Classification with Convolutional Adversarial Autoencoder"‬‬‬‬‬‬‬‬‬‬

Thank you very much for your attention and the reviewers' valuable comments on our manuscript. We have revised the manuscript according to the constructive reviews and suggestions. Please find enclosed the detailed response to the reviewers**.** Specifically, in this revised version, we have made the following changes.

1. As suggested by the reviewer, we have performed additional experiments on two external datasets: the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL) and the National Alzheimer's Coordinating Center (NACC).
2. As suggested by the reviewer, we have included references on Alzheimer's disease classification using the attention mechanism.
3. As suggested by the reviewer, we have provided sequence differences between MRIs captured by 1.5T and 3T scanners.
4. The whole paper has been polished for better presentation by all authors.
5. We revised the manuscript in two ways: adding and removing, where new additions are depicted in blue color, and deletions are noted in red.

**Response to Reviewers**

"Ms. No. MEDIA-D-21-01329: Reducing Variations in Multi-center Alzheimer's Disease Classification with Convolutional Adversarial Autoencoder" ‬‬‬‬‬‬‬‬

We thank the reviewers and editor for spending substantial time looking over the paper and providing valuable comments. In the following, we give a point-by-point response to your concerns.

**Detailed Response to Reviews**

**Reviewer #1:**

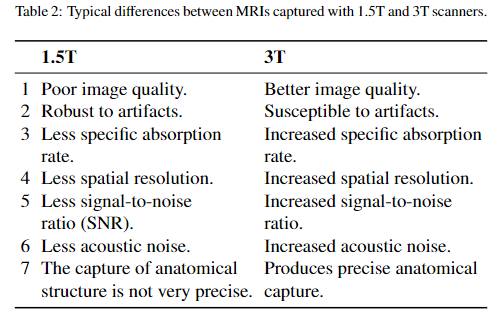
*Globally, the techniques is novel and the technical aspect sounds and the manuscript is well written and easy to follow.*

***Comment 1****: Authors state that their method enables to not remove preprocessing step to directly apply on raw data. However, it seems that what they really proposed is to replace classical preprocessing which usually includes alignment to common space, skull stripping and intensity harmonization. Current preprocessing methods are totally automatic and does not require user interaction. Can the authors explain what kind of bias classical method introduce?*

**Reply:** Thanks for your valuable comment. We want to point out that this proposed method is to augment current classical methods and not replace the entire preprocessing pipeline for AD MRI scans. Our work is geared more toward brain structure and atrophy variability in this work. It is popularly known in the literature that classical methods are time-consuming (it can take several days for preprocessing), need experienced radiologists with domain experience, and has the challenge of reproducible preprocessing of MRI data. Furthermore, the increasing complexity of different processing tools and pipelines across various institutions introduces these biases.

***Comment 2***: *I suggest the author to better emphasize sequence differences between 1.5T and 3T. I suggest to create a table. It is also unclear if author used both T1w and T2w as input of their method, or just T1w.*

**Reply**: Thanks for the valuable comment. We have provided a table to illustrate the difference between 1.5T and 3T MRI acquisition differences in Table 2 of this current version. In section 2.2, we added the statement, "*In Table 2, we present typical sequence differences between MRIs captured with 1.5T and 3T scanners."* See page 4 for details. Concerning the use of T1w and T2w sequence scans, we assert that the two structural MRI acquisition sequences were used.



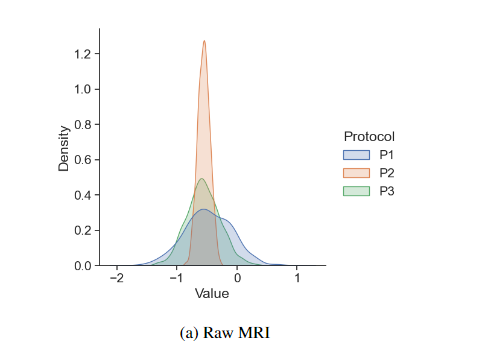
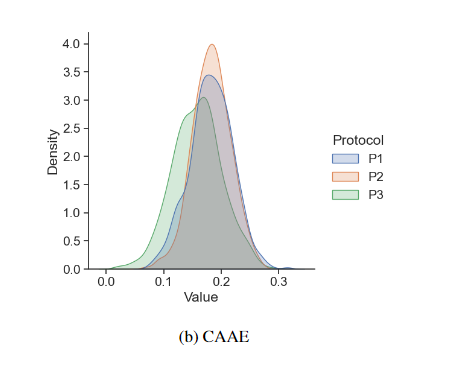
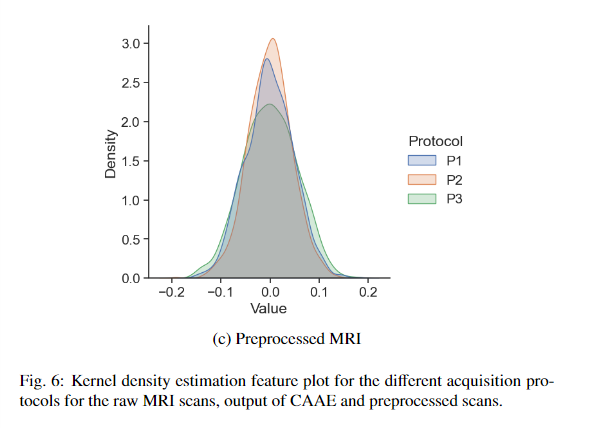
***Comment 3****: The author compare there method for MRI alignment but did not specify which method or what type of transformation has been done to align scans to MNI152 template, did the author used ANTs, FSL, etc. Did they perform rigid, affine, elastic transformation? Without this information it is impossible to assess the quality of the alignment.*

**Reply**: Thanks for the concern. The preprocessing for all other baselines was done with MALPEM and FSL. Specifically, the following preprocessing pipes were applied to the MRI scans: registration, skull-stripping, bias field correction, enhancement using histogram equalization, tissue segmentation, subcortical segmentation, and normalization. We used the FLIRT tool available within the FSL package to align the scans to the MNI152 template while performing affine transformations.

After the image registration, we normalized intensities of all the voxels [mean = 0 and standard deviation (SD) = 1. We then performed background removal where all the voxels from background regions outside the skull were set to −1 to ensure uniform background intensity. Furthermore, cortical and subcortical structures from volumetric MRI scans were segmented using FreeSurfer (Fischl, 2012). In-built functions such as 'recon-all', 'mri\_annotation2label', 'tkregister2', 'mri\_label2vol', 'mri\_convert' and 'mris\_calc' were used to obtain the segmented structures.

***Comment 4****: Following the previous point, one of the big challenges of multi-site dataset (beside MRI alignment which is not really a multi-site problem but more a inter-subject/acquisition variability problem) is the difference in terms of T1w signal distribution differences. However, the author does not discuss of it. How does their CAAE perform compare to classic intensity standardization method?*

**Reply**: Thanks for the follow-up comment. As already mentioned, the main variabilities earmarked in this work are the factors mentioned in the introduction section: "MRI scanners' differences, center-specific parameter settings, imaging protocols, different individual brain morphology, MRI scans acquisition imperfections and patient positioning across various clinics and hospitals." These variabilities exist because the MRI scans inherit one of these variability factors. These influencing factors are likely to present differences in MRI feature distribution among different centers and make AD disease prediction difficult. In contrast, the variability is not only introduced by the T1 or T2 but is probable to induce variability in MRI scans captured from different sites. Also, this current work focuses on the overall feature distribution these variabilities introduce, which makes the prediction of the AD difficult using multi-center MRI scans. Kindly see Figure 6 for the KDE distribution plot for the raw MRI scans, the CAAE output, and the preprocessed MRI scans. Even though the KDE of the preprocessed MRI is better than our proposed method, it possesses the capability of learning a near to perfect common transformation space for the raw MRI scans. Most importantly, these new transformation contains AD discriminative features for the robust prediction of the disease.



***Comment 5****: It seems odd to compare classification results between method using different modality without describing this difference. For example, author compare their work with* *Maggipinto et al. 2017, that used DTI data to classify AD. Do the author run the method proposed by Maggipinto et al. on their on dataset using structural MRI. More emphasis has to be made on this point.*

**Reply**: Thanks very much. Actually, for the method proposed by Maggipinto et al. 2017, we did follow their machine learning classification method with random forest robustly. But concerning the preprocessing pipeline for their DTI dataset, we did not follow holistically because of the difficulty accessing their preprocessing code and the subtle difference between these modalities. We confirm that we followed their fundamental preprocessing pipeline except for the diffusion tensor fitting and extraction of gradient directions and b-values proposed in their data preprocessing methodology.

***Comment 6****: Following point 1. I would be interested to see how the CAAE enables to increase classification performances compare to different classical preprocessing pipeline (see point 4, spatial transformation and intensity standardization)*

**Reply**: Thanks. From table 3, it could be seen that although the proposed CAAE approach scans achieved significant results on the raw MRI, it achieved more performance gain in the AD prediction when applied together with the classical preprocessing pipeline. From the experimental results, our proposed method on the raw multi-center MRI scans achieved the classification accuracy of 91.90%, 90.05%, and 88.10%, respectively, for AD vs. HC, AD vs. MCI, and MCI vs. HC, respectively, but for the preprocessed scans we achieved 93.10%, 91.4%1 and 89.26%. The proposed CAAE is to simultaneously learn the discriminative features of the raw multicenter MRI scans and map them to a common space (i.e., the same Gaussian distribution) as described in detail in subsection 3.2 in section 3. Specifically, the proposed CAAE robustly learns the most important features for AD prediction and discards unimportant features like the skull. Therefore, our proposed method carries the central idea of the classical preprocessing pipeline.

**Reviewer #2:**

***Comment 1***: *Since the cross-validation was only performed on the CRAT module (i.e., the reconstructed images from CAAE), and the optimization of the CRAT and CAAE were done separately, it is important to perform the same CV in the CAAE first, i.e., the data split should be done at the really beginning. Although CAAE is unsupervised learning, data leakage could happen if all image were used at this stage. See Wen et al. 2020, for details. Moreover, is it possible to jointly optimize the two losses from the two modules?*

**Reply**: Thanks for the valuable comment. We performed the same cross-validation strategy for the training of the CAAE model. Thus, the data splitting for the model was performed at the very beginning before training the two models. We have included this additional information in this current version to clarify any ambiguity in the experimental section, "*To prevent data leakage, we maintain the same data splitting strategy for the CAAE and the CRAT at the very beginning of the experiments.".* Also, concerning the optimization of the two losses from the two modules, we initially attempted this approach, but because of the huge dimensionality of the 3D MRI scans and several 3D convolutional layers in the architecture of both models, we ran into Out of Memory (OOM) exception making us approach this current task as a two-step optimization problem.

***Comment 2***: *Is it mandatory to have the pair of T1 and T2 images to train or apply the model? This is not clear to me. What the added value of T2 images on top of T1? All individuals have both T1 and T2 scans? If not, how did the model handle with the missing modality?*

**Reply**: Thanks. It is not mandatory to have a pair of T1 and T2 images for the model training. Specifically, all the acquired images are structural MRI (sMRI) scans, with subjects having either T1 or T2 images. Since both T1 and T2 are sMRI scans but only with different voxel intensities (sequence acquisition), handling the missing modality is not important in this task.

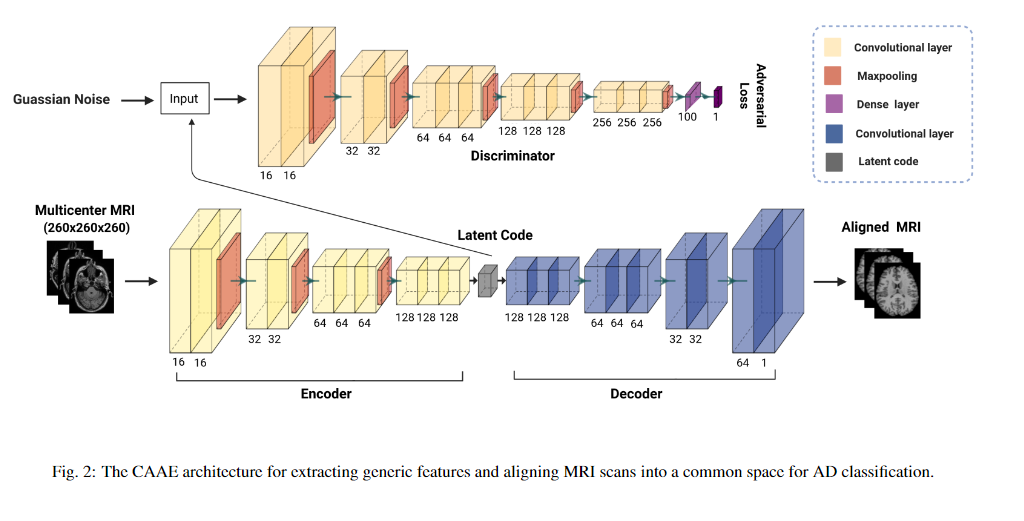
***Comment 3***: *Although the current study adopted a nested cross-validation, it is not clear how authors tuned the hyperparamters of the DL models. Please be specific for these parameters, how many layers, number of neurons, learning rates, etc*

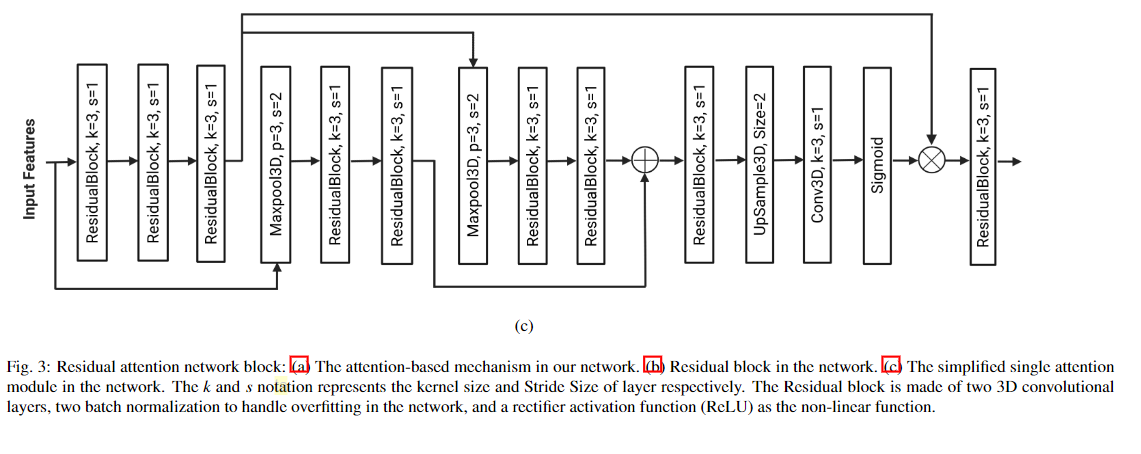
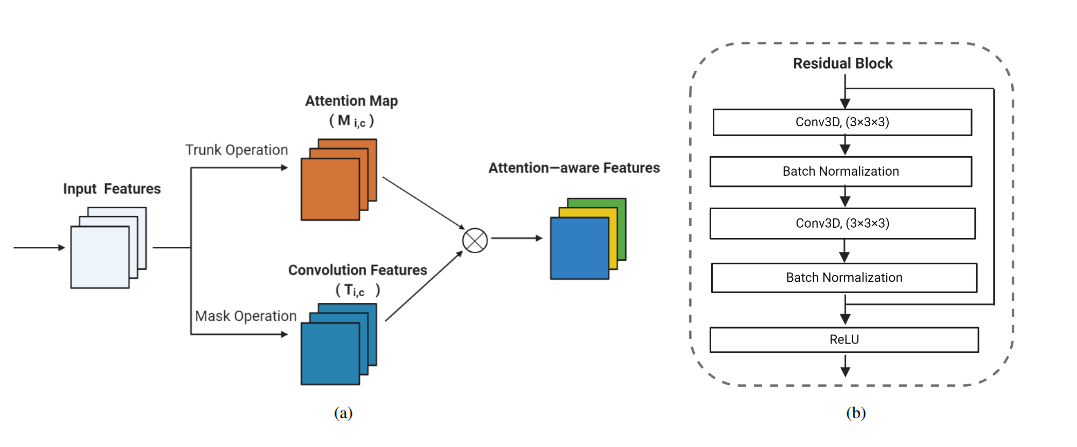
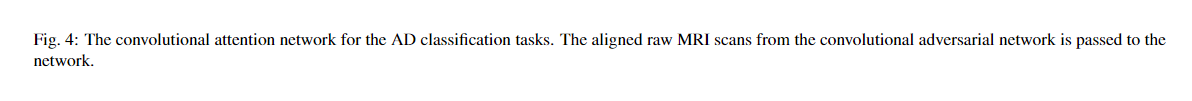
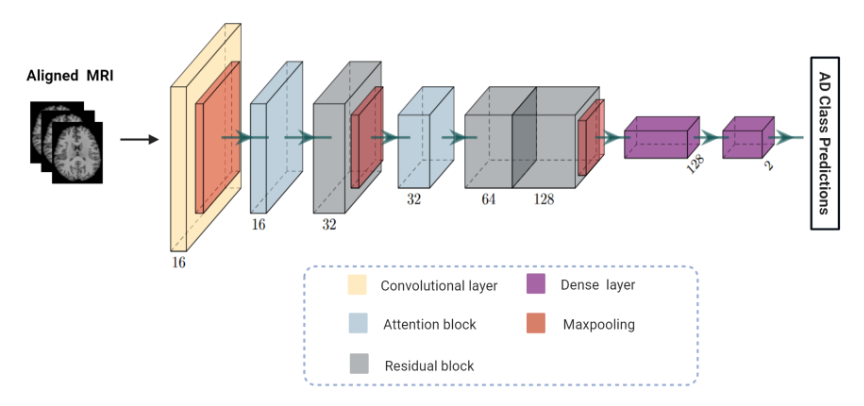
**Reply**: Thanks. We followed the literature on AD classification tasks and performed several ablation studies in selecting the best hyperparameter for the training of the two models. For the unsupervised CAAE network, we used a learning rate (lr) of 0.0001. Also, we used a lr =0.00001 for the supervised classification task with CRAT. Most AD classification tasks used a very low initial lr=0.0001, so we fine-tuned to a lesser lr (<0.0001) for this current work. Because Adam optimizer is an adaptive gradient learning algorithm, we chose not to apply any other learning scheduling techniques like step scheduling, exponential scheduling, etc. but keep its step-size upper bound with our initial learning rates. In arriving at our chosen lr, we manually fined-tuned the model with 0.00001≤lr≤ 0.0001 and settled on 0.00001 based on ablation experiments for the classification tasks. We kept the learning rate constant without learning rate decay throughout the training for the CRAT network. The AD classification task is very complex and needs small optimization steps for the network to reach its global optimum. The batch size for all training is five (5) because a very big batch size will cause an OOM exception due to the huge dimension of the AD dataset. The experiment was run for 150 epochs.

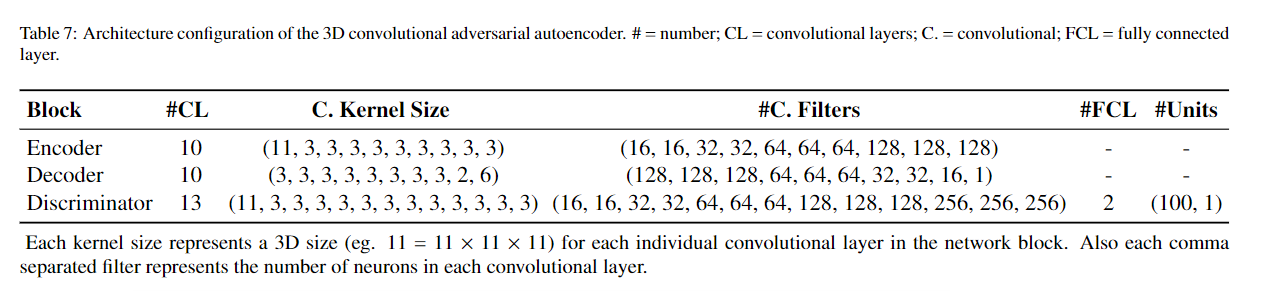
For each epoch iteration during model training, we compared the model's validation loss at the current epoch with the previous known lowest loss at . If , then we save the current model's instance (with its parameters, specifically the weights and bias) as the best checkpoints. These checkpoints are then restored during testing. We applied the early stopping technique for each fold's training with a patience of 5 when there was no decrease in the validation loss.

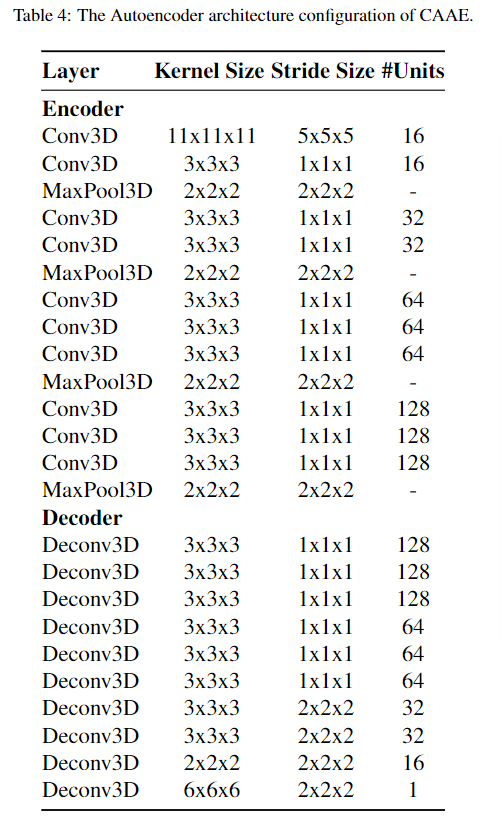
The architecture configuration of CAAE is as follows: There are forty-three (43) layers present in this model. The encoder module has ten convolutional layers and three (3) max-pooling layers. The decoder module has ten deconvolutional layers. For the discriminator, it has twenty network layers. Of these, thirteen are convolutional layers, five max-pooling layers, and two dense layers. Please see Figure 2 in the main manuscript and also Table A.4, Table A.5, Table A.7 in Appendix C: Network Configuration for the detailed architecture of the Adversarial autoencoder, CAAE.

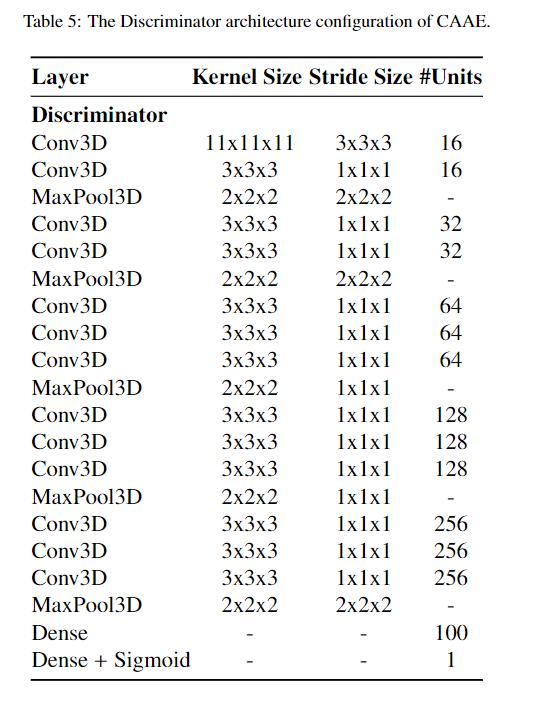
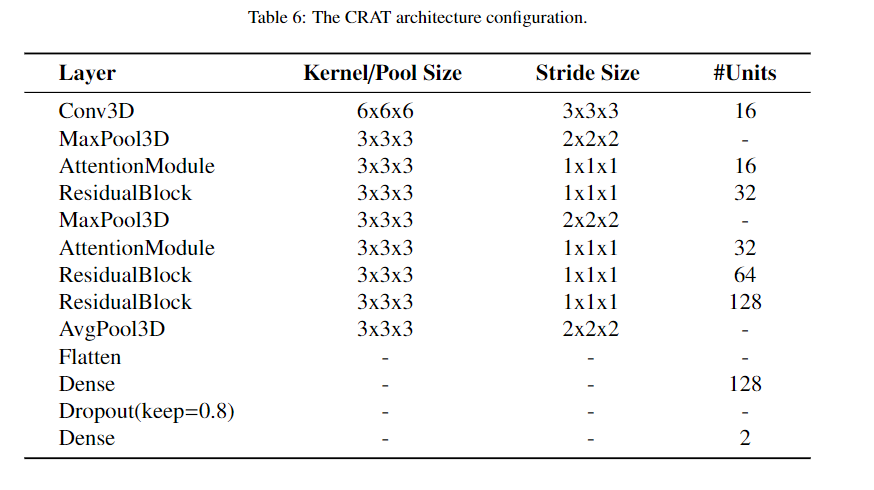
The CRAT model architecture has fifty-six (56) network layers in total. It comprises two attention blocks, one convolutional layer, three residual blocks, three max-pooling layers, and two dense layers. Each attention block comprises nine residual blocks, two max-pooling layers, and a convolutional layer. Also, each residual block is made of two convolutional layers. There are forty-five convolutional layers, seven max-pooling layers, two deconvolutional layers, and two dense layers. See Figure 3 and Figure 4 for an overview of the CRAT architecture and the detailed architecture configuration in Table A.6 in Appendix C: Network Configuration.











***Comment 4:*** *The classification results here are high, especially for AD vs MCI and CN vs MCI. AD vs CN has been an relatively easy task in the literature. Recent works published on Nature journal has identified that MCI/AD are very heterogeneous, and multiple subtypes exist in these disease labels. Please see:*

*a) https://doi.org/10.1038/s41467-018-05892-0*

*b) https://doi.org/10.1038/s41467-021-26703-z*

*So, it is important to apply the trained model to external independent datasets. Is it possible to apply the trained model to AIBL and OASIS?*

**Reply**: Thanks for the comment. Though AD vs. MCI and CN vs. MCI have been difficult tasks in some AD classification literature, other works using attention networks have proven to perform these tasks with good performance, as reported in the references below [1-5]. Most importantly, our CAAE and CRAT possess the discriminative capability of identifying the subtle discriminative features of AD atrophy, making them classify AD vs. NC well as AD vs. MCI and CN vs. MCI very well.

We have performed additional experiments and applied the trained model on two popular external datasets: the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL) and the National Alzheimer's Coordinating Center (NACC). Kindly refer to section 4.6. Ablation study on external datasets, at page 15 for detail. It reads, "*To further verify the proposed method's effectiveness, we tested the trained models on two popular external AD datasets. The datasets are the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL) Ellis et al. (2010) and the National Alzheimer's Coordinating Center (NACC) Beekly et al. (2004). For each external dataset, we used 195 subjects where 65 belong to each of the classifying groups. The performance is illustrated in Table 6. It can be seen that although the prediction on the preprocessed MRI scans achieved better results than the raw scans aligned to a common space, the performance still is comparable. The experimental results prove that the convolutional adversarial model and the attention mechanism can predict AD without a robust preprocessing pipeline*. *Also, the proposed model is likely to have improved performance when trained from scratch on the external dataset and a larger dataset.* "

*[1] Zhang, J., Zheng, B., Gao, A., Feng, X., Liang, D., Long, X., 2021a. A 3d densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification. Magnetic Resonance Imaging 78, 119–126.*

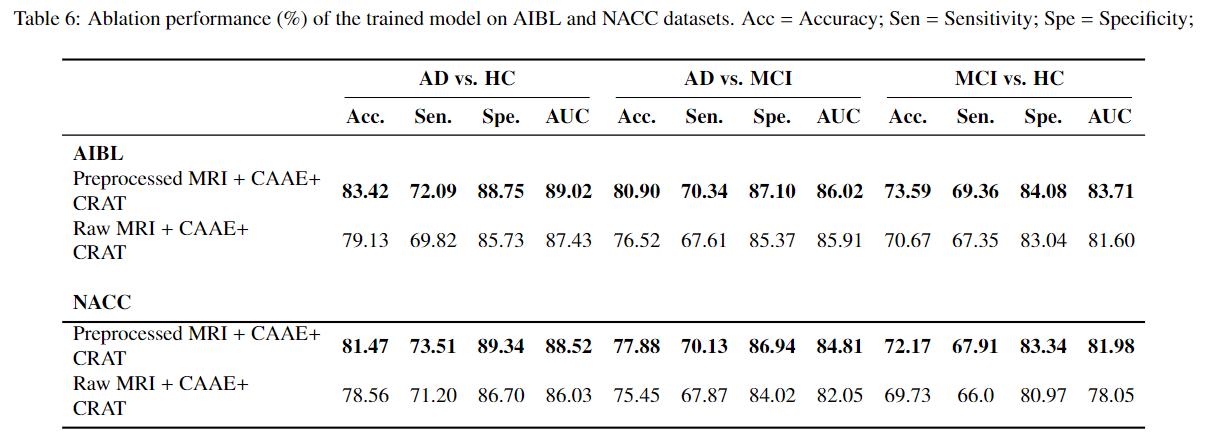
*[2] Liu, Z., Lu, H., Pan, X., Xu, M., Lan, R., Luo, X., 2022. Diagnosis of alzheimer's disease via an attention-based multi-scale convolutional neural network. Knowledge-Based Systems 238, 107942.*

*[3] Zhang, L., Wang, L., Zhu, D., 2020. Jointly analyzing alzheimer's disease related structure-function using deep cross-model attention network, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE.*

*pp. 563–567.*

*[4] Wang, S.H., Zhou, Q., Yang, M., Zhang, Y.D., 2021b. Advian: Alzheimer's disease vgg-inspired attention network based on convolutional block attention module and multiple way data augmentation. Frontiers in Aging Neuroscience 13, 313.*

*[5] Jin, D., Xu, J., Zhao, K., Hu, F., Yang, Z., Liu, B., Jiang, T., Liu, Y., 2019. Attention-based 3d convolutional network for alzheimer's disease diagnosis and biomarkers exploration, in: 2019 IEEE 16Th international symposium on biomedical imaging (ISBI 2019), IEEE. pp. 1047–1051.*



***Comment 5***: *Previous studies that adopted attention mechanism in AD classification should be discussed, or at least cited.*

**Reply**: Thanks for the suggestion. We have cited five works that leveraged attention mechanism in classifying Alzheimer's disease. A new paragraph has been added in the introduction section, "*Among the most current deep learning techniques in Alzheimer's disease prediction, attention mechanism has proven to be a potent tool, and an important component in predicting the disease using MRI scans Zhang et al. (2021a,2020); Liu et al. (2022). The attention mechanism technique has shown high accuracy for the prediction of AD, as reported in these works Liu et al. (2022); Wang et al. (2021b); Jin et al. (2019). Attention mechanism has proved efficient in capturing global dependencies because it learns the dependency relationship between each AD feature and assigns corresponding weights to map out the most discriminative features."*

[1] Zhang, J., Zheng, B., Gao, A., Feng, X., Liang, D., Long, X., 2021a. A 3d densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification. Magnetic Resonance Imaging 78, 119–126.

[2] Liu, Z., Lu, H., Pan, X., Xu, M., Lan, R., Luo, X., 2022. Diagnosis of alzheimer's disease via an attention-based multi-scale convolutional neural network. Knowledge-Based Systems 238, 107942.

[3] Zhang, L., Wang, L., Zhu, D., 2020. Jointly analyzing alzheimer's disease related structure-function using deep cross-model attention network, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE.pp. 563–567.

[4] Wang, S.H., Zhou, Q., Yang, M., Zhang, Y.D., 2021b. Advian: Alzheimer's disease vgg-inspired attention network based on convolutional block attention module and multiple way data augmentation. Frontiers in Aging Neuroscience 13, 313.

[5] Jin, D., Xu, J., Zhao, K., Hu, F., Yang, Z., Liu, B., Jiang, T., Liu, Y., 2019. Attention-based 3d convolutional network for alzheimer's disease diagnosis and biomarkers exploration, in: 2019 IEEE 16Th international symposium on biomedical imaging (ISBI 2019), IEEE. pp. 1047–1051.