MULTIMODAL ATTENTION-BASED DEEP LEARNING FOR ALZHEIMER'S DISEASE DIAGNOSIS

Supplementary Material

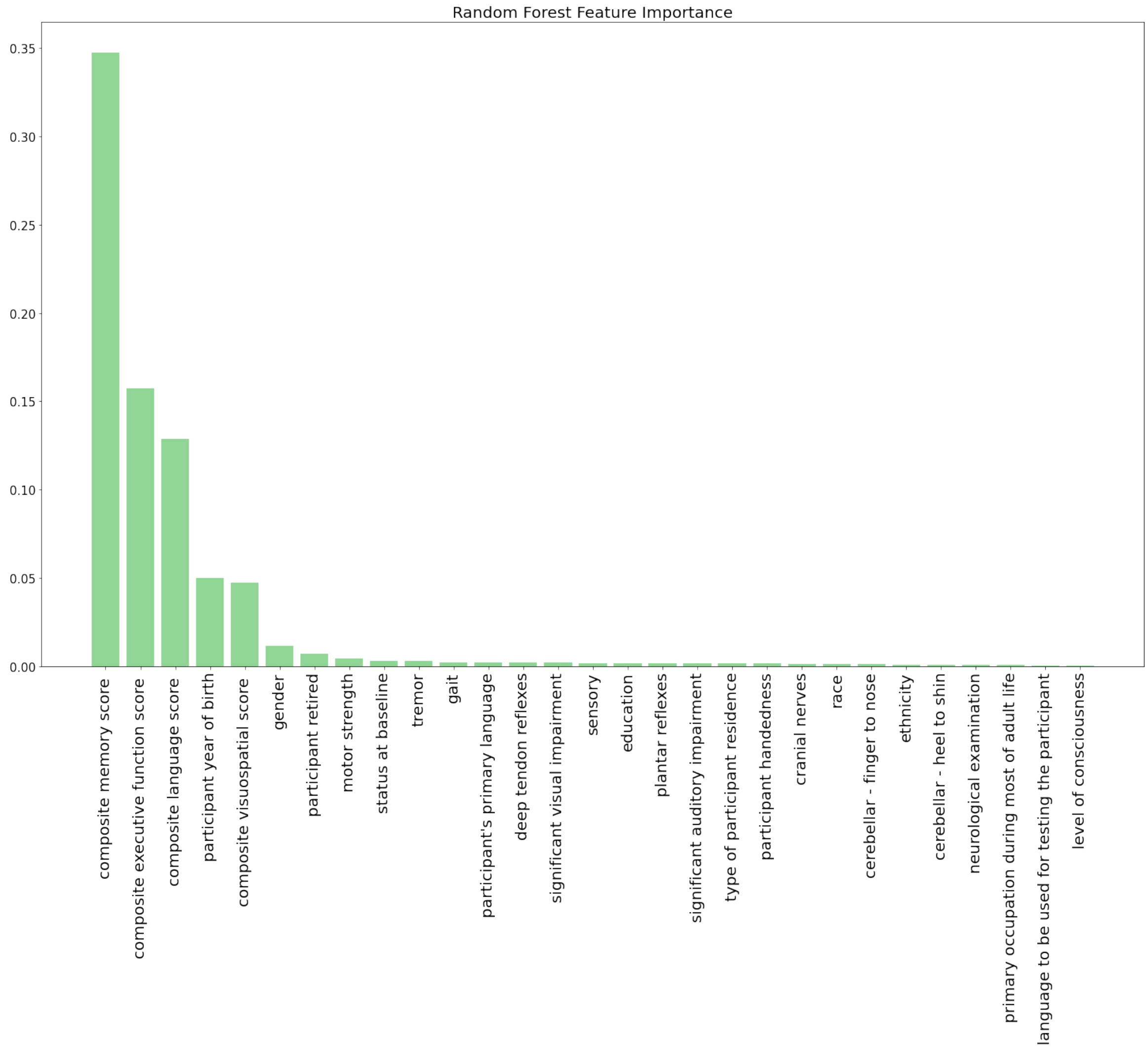
S1 Genetic Data Pre-processing

The genetic data consists of the whole genome sequencing (WGS) data from 805 ADNI participants by Illumina's non-Clinical Laboratory Improvement Amendments (non-CLIA) laboratory at roughly 30–40 × coverage in 2012 and 2013. The resulting variant call files (VCFs) have been generated by ADNI using Broad best practices (Burrows-Wheeler Aligner (BWA) and Genome Analysis Toolkit (GATK)-haplotype caller) in 2014. We first filtered the SNPs by the Hardy-Weinberg equilibrium (HWE) test for each site (p-values) by removing SNPs with HWE $p < 0.05$. We then checked the genotype quality (GQ) and removed SNPs with GQ < 20. Next, we filtered by minor allele frequency (MAF) and removed sites with MAF < 0.01. Lastly, we performed genotype value filtering where we excluded sites based on the proportion of missing data and removed sites with a missing rate > 0.05. After filtering with the above criteria, we utilized genes known to be related to Alzheimer's Disease. In this step, we first downloaded a list of all AD-related genes from the AlzGene Database (`http://www.alzgene.org/`), which contains 680 genes in total. Then we searched these genes in the UCSC genome browser (`https://genome.ucsc.edu/`) and kept the 640 genes that matched NCBI RefSeq annotation. We extracted these gene regions from RefSeq Annotation (gff file) in Bed format and use them to filter the SNPs further. We only retain the genes that are located in these regions. After selecting the 680 genes of known association with AD, we had 547,863 SNPs left. As discussed in the Genetic Data Pre-processing Section, we needed to find a way to reduce the number of features. We used a Random Forest Classifier to create a list of the most important features. Since this is a supervised method of creating features, this brought more promising results to the performance, in contrast to an approach such as principal component analysis (PCA) which is unsupervised. To find the best set of features, we tried using 50, 100, 150, and 200 forests as the parameter in the classifier. After creating the four sets of features, we used a validation set (10% from the training) to do hyperparameter tuning (as described in S4), for each set of features, we found that the set from 100 forests performed the best on the validation set resulting in the accuracy described in the Performance of Unimodal Models Section.

S2 Clinical Features

In the Model Robustness Section we discussed the value of clinical features to the model's performance. To ensure that there are no variables in the data that could potentially give an unfair advantage to the model (e.g. medication that a patient takes, only when they already have AD), we carefully examined all available variables. We fit a Random Forest classifier (from the scikit-learn package [30]) which outputs the features along with their importance score. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature [30]. Figure S1 shows the full list of features and their importance.

Figure S1: **Clinical feature importance.** The graph shows all the clinical features used in our model in order of most important to least important



S3 Imaging Data Pre-processing

The images used for our study are pre-processed by ADNI with specific image correction steps:

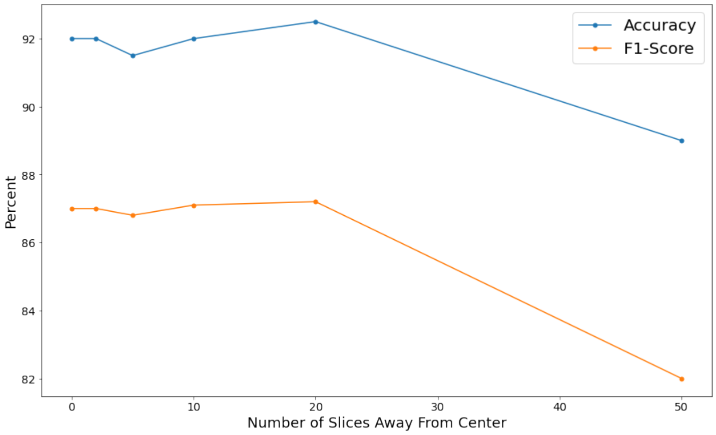
1. Gradwarp: gradwarp is a system-specific correction of image geometry distortion due to gradient non-linearity. The degree to which images are distorted due to gradient non-linearity varies with each specific gradient model. We anticipate that most users will prefer to use images which have been corrected for gradient non-linearity distortion in analyses.

2. B1 non-uniformity: this correction procedure employs the B1 calibration scans noted in the protocol above to correct the image intensity non-uniformity that results when RF transmission is performed with a more uniform body coil while reception is performed with a less uniform head coil.

3. N3: N3 is a histogram peak sharpening algorithm that is applied to all images. It is applied after grad warp and after B1 correction for systems on which these two correction steps are performed. N3 will reduce intensity non-uniformity due to the wave or the dielectric effect at 3T. 1.5T scans also undergo N3 processing to reduce residual intensity non-uniformity.

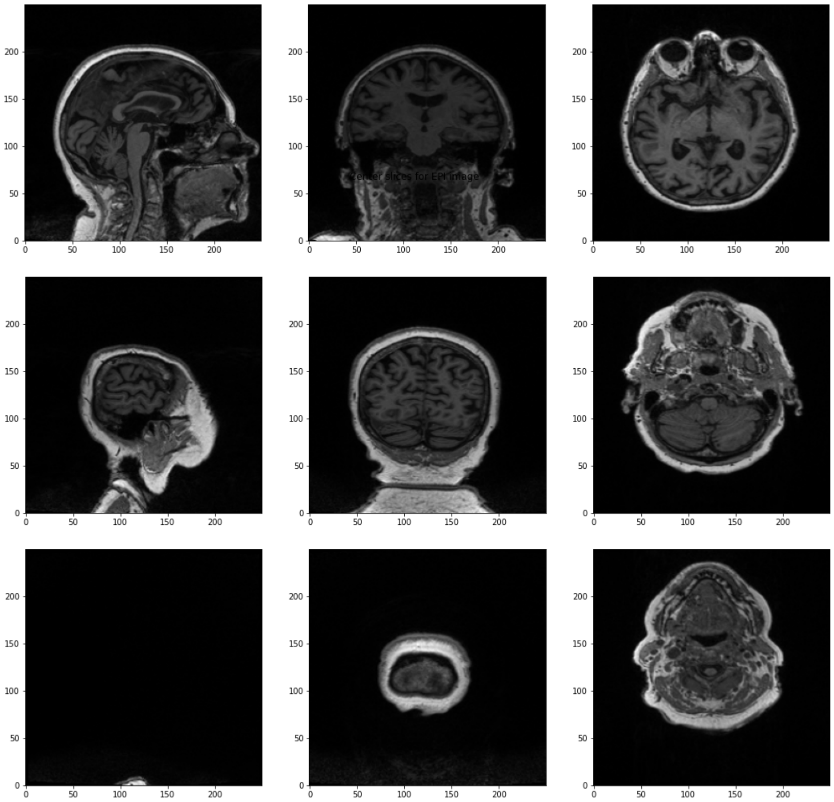
We followed the same pre-processing steps as Bucholc et al. [19], El-Sappagh et al. [17], and Abuhmad et al. [18], which relied on ADNI's correction steps without further modification.
To demonstrate that the performance of the unimodal imaging model is not significantly impacted by the addition of more brain slices, we report the optimized performance of the model with just the middle three slices, 2 more images per angle (6 more in total), 5 more images per angle, 10 more images per angle, 20 more images per angle, and 50 more images per angle. We report both F1-Scores and accuracy (averaged across 3-fold validation set), which follow a similar trend shown in S2. The difference in performance between no additional slices (used in the paper) and 20 additional slices are all within 1 percent. When adding 50 slices to each angle, we observe a significant decline in performance. Thus, we proceeded with the original choice of just using the middle of the brain.

Figure S2: **Validation F1-Score and Accuracy Trend as Number of Images Increases.** The graph shows that the unimodal imaging model does not significantly benefit from the addition of more images.



The decline in performance can be attributed to the fact that the slices further away from the center do not contain meaningful information and add noise to the model. The example below shows the middle three slices, followed by the outer 10 slices and outer most slices.

Figure S3: **Example of MRI slices as distance increases from the center.**



S4 Hyperparameter Tuning Methods

To perform hyperparameter tuning for both unimodal and multimodal results, we randomly split the data into a training set (90%) and a testing set (10%). The testing set was set aside and withheld from tuning. We picked the best hyperparameters on the average validation accuracy of the 3-fold cross-validation scheme. For either a fully connected neural network or a convolutional neural network, the architecture, batch size, number of epochs, and learning rate were chosen via tuning. Table S1 describes all hyperparameters considered.

Table S1: **Hyperparameter Grid**

Hyperparameters	Values
Learning Rate	[0.00001, 0.1]
Dropout Values	{0.1, 0.2, 0.3, 0.4, 0.5}
Number of Layers	[1, 6]
Batch Size	{16, 32, 64, 128}
Number of Epochs	{10, 20, 50, 80, 100, 150, 200}

The hyperparameters that gave the highest accuracy for each type of model are shown in S2. For the multimodal framework, the best unimodal neural network values were added into the architecture.

Table S2: **Best Hyperparameters**

	Learning Rate	Batch Size	Number of Layers	Dropout Value	Number of Epochs
Unimodal Clinical	0.0001	32	3	{0.2, 0.3, 0.5}	100
Unimodal Genetic	0.001	32	3	{0.3,0.5}	50
Unimodal Imaging	0.001	32	3	{0.3,0.5}	50
Multimodal	0.001	32	{3, 3, 3}	{0.2, 0.3, 0.5}	50

S5 Evaluation Metrics

For our multi-class setting, we used the formulas below for each class. For example, for class 0 (control), we calculated the number of true positives, true negatives, false positives, and false negatives just for class 0. Then, we use “macro” averaged F1-score using the arithmetic mean of all the per-class F1-scores.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

S6 Performance of Unimodal Models

Table S3 shows the numeric information presented in Figure 3. We report all four evaluation metrics for the best neural network model for each modality - imaging, clinical, and genetic. The imaging model gives the best performance overall, whereas the genetic modality gives the lowest performance.

Table S3: **Results of Unimodal Models.**

	Accuracy (%)	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Clinical	80.59	80.56	80.48	80.47
Imaging	92.23	94.02	90.4	91.83
Genetic	77.78	78.37	76.92	77.24

S7 Model Robustness

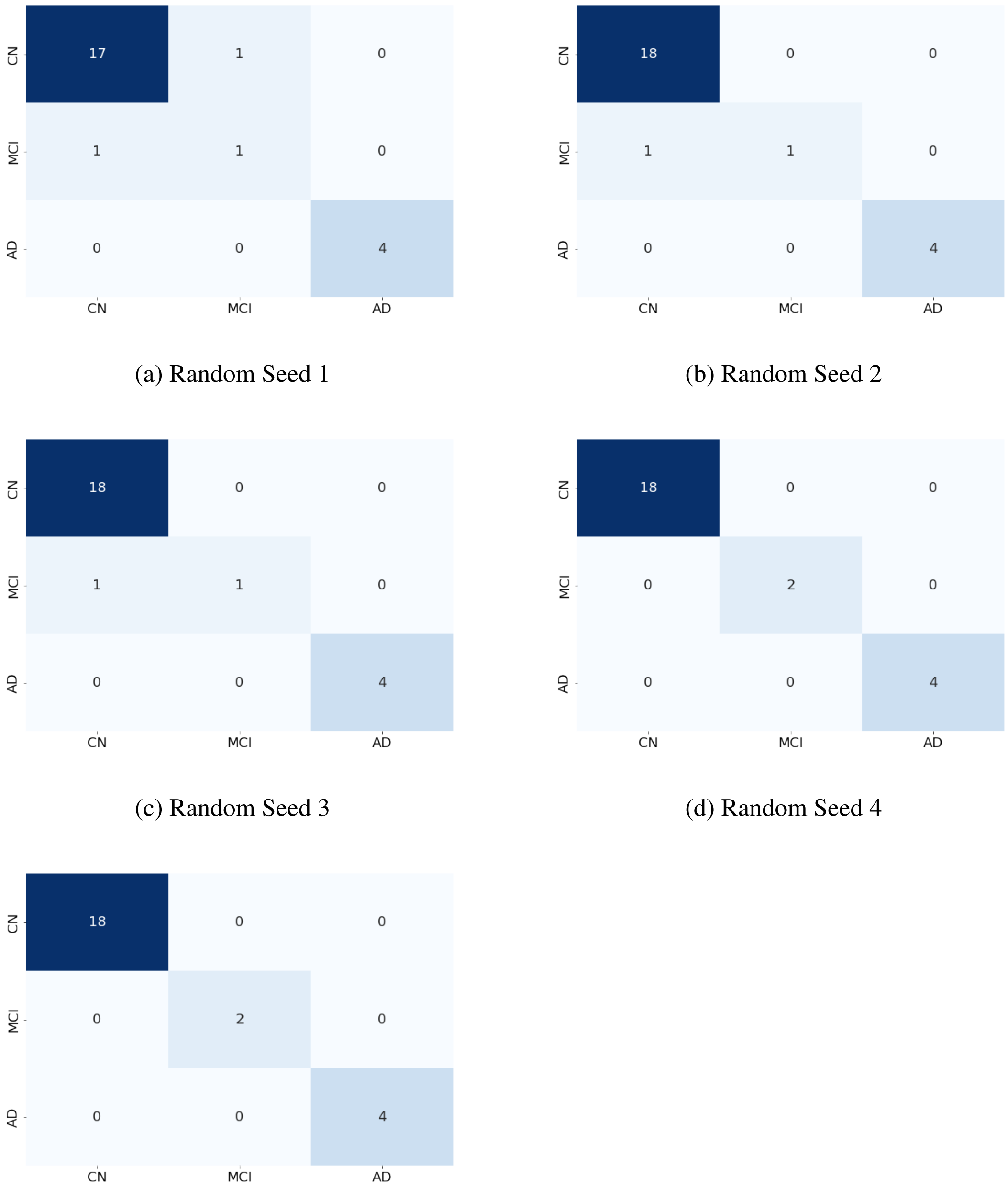
Table S4 shows the numeric information presented in Figure 4.

Table S4: **F1-Score Distribution for different attention-based and attention-free baselines.** The table shows the F1-score distribution across 100 random seeds to show the value of attention in a deep learning model. The table demonstrates that the combination of self-attention with cross-modal attention performs the best with the most narrow variation.

	Lower Whisker	**Lower Quartile**	**Median**	**Upper Quartile**	**Upper Whisker**
Self-Att + Cross-Modal Att	0.7767	0.8268	0.8799	0.9238	1
Cross-Modal Att	0.4068	0.6491	0.7657	0.8268	1
Self-Att	0.7175	0.8148	0.8682	0.8799	0.9238
No Attention	0.6396	0.7714	0.8148	0.8799	0.9238

S8 Investigating Individual Class Performance

We include confusion matrices for each of the 5 random initializations to supplement Table 4 in the Model Robustness Section. Each confusion matrix represents the results of our best multimodal model with respect to a random seed.



(a) Random Seed 1

(b) Random Seed 2

(c) Random Seed 3

(d) Random Seed 4

(e) Random Seed 5

Figure S4: **Confsion matricies for 5 random initializations of MADDi.**

S9 Sample Selection

To demonstrate that our sample selection process was thorough, we show in Table S5 the results of the models described in Model Robustness Section on the 3-fold cross validation scheme. The metrics in the table are averaged across 5 random initializations. Since these results are similar to the ones reported on the test set in Table 3, we consider the test set a fair sample of our data.

Table S5: **Cross-Validation Results**

	F1-Score Val Set 1	**F1-Score Val Set 2**	**F1-Score Val Set 3**
Cross-Modal Att + Self-Att	89.74%	97.44%	92.30%
Cross-Modal	87.18%	91.02%	88.46%
Self-Att	76.92%	85.89%	83.34%
No Attention	79.48%	89.74%	87.17%

S10 Evaluation of Modality Importance

Table S6 shows the contribution and performance of each modality on the overlap patient set. The metrics were calculated as an average of five random initializations on a held-out test set. It captures the same information as Figure 5 in the Modality Importance Section but provides all the numeric results.

Table S6: **Evaluation of Modality Importance**

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Clinical	82.29 ± 9.49	78.92 ± 3.68	88.43 ± 5.29	78.30 ± 6.70
Genetic	77.78 ± 3.91	78.37 ± 4.64	76.92 ± 3.78	77.24 ± 4.03
Imaging	71.66 ± 4.68	53.38 ± 9.55	62.03 ± 9.77	55.46 ± 8.86
Clinical and Genetic	92.50 ± 3.18	87.05 ± 9.36	81.85 ± 1.36	83.21 ± 4.21
Genetic and Imaging	78.33 ± 1.86	50.88 ± 8.19	52.59 ± 6.88	50.07 ± 4.28
Imaging and Clinical	85.83 ± 10.73	80.81 ± 8.56	88.15 ± 15.52	80.52 ± 13.34
Clinical, Genetic, Imaging	**96.88 ± 3.33**	**88.15 ± 14.22**	**91.23 ± 13.37**	**89.32 ± 15.59**