Journal of the American Medical Informatics Association, 29(12), 2022, 2014–2022
https://doi.org/10.1093/jamia/ocac168
Advance Access Publication Date: 23 September 2022
Research and Applications

Research and Applications

Multimodal attention-based deep learning for Alzheimer's disease diagnosis

Michal Golovanevsky[1], Carsten Eickhoff[1,2], and Ritambhara Singh[1,3]

[1]Department of Computer Science, Brown University, Providence, Rhode Island, USA, [2]Center for Biomedical Informatics, Brown University, Providence, Rhode Island, USA, and [3]Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, USA

Corresponding Author: Carsten Eickhoff, PhD, 233 Richmond Street, Providence, RI 02903, USA; carsten@brown.edu

Received 15 June 2022; Revised 10 August 2022; Editorial Decision 1 September 2022; Accepted 21 September 2022

ABSTRACT

Objective: Alzheimer's disease (AD) is the most common neurodegenerative disorder with one of the most complex pathogeneses, making effective and clinically actionable decision support difficult. The objective of this study was to develop a novel multimodal deep learning framework to aid medical professionals in AD diagnosis.

Materials and Methods: We present a Multimodal Alzheimer's Disease Diagnosis framework (MADDi) to accurately detect the presence of AD and mild cognitive impairment (MCI) from imaging, genetic, and clinical data. MADDi is novel in that we use cross-modal attention, which captures interactions between modalities—a method not previously explored in this domain. We perform multi-class classification, a challenging task considering the strong similarities between MCI and AD. We compare with previous state-of-the-art models, evaluate the importance of attention, and examine the contribution of each modality to the model's performance.

Results: MADDi classifies MCI, AD, and controls with 96.88% accuracy on a held-out test set. When examining the contribution of different attention schemes, we found that the combination of cross-modal attention with self-attention performed the best, and no attention layers in the model performed the worst, with a 7.9% difference in F1-scores.

Discussion: Our experiments underlined the importance of structured clinical data to help machine learning models contextualize and interpret the remaining modalities. Extensive ablation studies showed that any multimodal mixture of input features without access to structured clinical information suffered marked performance losses.

Conclusion: This study demonstrates the merit of combining multiple input modalities via cross-modal attention to deliver highly accurate AD diagnostic decision support.

Key words: Alzheimer's disease, clinical decision support, artificial intelligence, machine learning, deep learning, multimodal deep learning

INTRODUCTION

Background and significance

Alzheimer's disease (AD) is the most common neurodegenerative disorder affecting approximately 5.5 million people in the United States and 44 million people worldwide.[1] Despite extensive research and advances in clinical practice, less than 50% of AD patients are diagnosed accurately for pathology and disease progression based on clinical symptoms alone.[2] The pathology of the disease occurs several years before the onset of clinical symptoms, making the disease difficult to detect at an early stage.[3] Mild cognitive impairment

(MCI) is considered an AD prodromal phase, where the gradual change from MCI to AD can take years to decades.[4] As AD cannot currently be cured, but only delayed in progression, early detection of MCI before irreversible brain damage occurs is crucial for preventive care.

The urgent need for clinical advancement in AD diagnosis inspired the Alzheimer's Disease Neuroimaging Initiative (ADNI) to collect diverse data such as imaging, biological markers, and clinical assessment on MCI and AD patients.[5] Such distinct data inputs are often referred to as individual *modalities*; a research problem is characterized as *multimodal* when it considers multiple such modalities and *unimodal* when it includes just one. Thanks to data collection efforts such as the one spearheaded by ADNI, it became possible to create unimodal machine learning models capable of aiding AD diagnosis, most commonly using imaging data,[6–10] or clinical records.[11,12] Recently, deep learning (DL) has shown considerable potential for clinical decision support and outperformed traditional unimodal machine learning methods in AD detection.[7,13,14] The major strength of DL over traditional machine learning models is the ability to process large numbers of parameters and effectively learn meaningful connections between features and labels. Even with DL's advantage, single-modality input is often insufficient to support clinical decision-making.[15]

AD diagnosis is multimodal in nature—health care providers examine patient records, neurological exams, genetic history, and various imaging scans. Integrating multiple such inputs provides a more comprehensive view of the disease. Thus, several DL-based multimodal studies[16–19] have leveraged the richer information encoded in multimodal data. Despite an overall convincing performance, they all miss a crucial component of multimodal learning—cross-modal interactions. The existing methods simply concatenate features extracted from the different modalities to combine their information, limiting the model's ability to learn a shared representation.[20] In response, we propose a novel Multimodal Alzheimer's Disease Diagnosis framework (MADDi), which uses a cross-modal attention scheme[21] to integrate imaging (magnetic resonance imaging [MRI]), genetic (single nucleotide polymorphisms [SNPs]), and structured clinical data to classify patients into control (CN), MCI, and AD groups.

Many successful studies were conducted using the ADNI dataset.[5] Only a small subset of them used multimodal data, and an even smaller subset attempted 3-class classification. In this work, we propose to use attention as a vehicle for cross-modality interactions. We show state-of-the-art performance on the challenging multimodal 3-class classification task, achieving 96.88% average test accuracy across 5 random model initializations. Next, we investigated the contribution of each modality to the overall model. While for unimodal models, images produced the most robust results (92.28% accuracy), when we combined all 3 data inputs, we found that the clinical modality complements learning the best. Finally, since our method utilizes 2 different types of neural network attention, we investigated the contribution of each type and found significant performance improvements when using attention layers over no attention. Through our experiments, we were able to highlight the importance of capturing interactions between modalities.

MATERIALS AND METHODS

Data description

The data used in this study were obtained from the ADNI database (https://adni.loni.usc.edu/), which provides imaging, clinical, and genetic data for over 2220 patients spanning 4 studies (ADNI1, ADNI2, ADNI GO, and ADNI3). The primary goal of ADNI has been to test whether combining such data can help measure the progression of MCI and AD. Our study follows the common practice of using patient information from only ADNI1, 2, and GO, since ADNI3 is still an ongoing study expected to conclude in 2022. To capture a diverse range of modalities, we focused on patients with imaging, genetic, and clinical data available. We trained unimodal models on the full number of participants per modality. For our multimodal architecture, we only used those patients who had all 3 modalities recorded (referred to as the *overlap dataset*). The number of participants in each group is detailed in Table 1.

Ground truth labels

Since ADNI's launch in 2003, more participants have been added to each phase of the study, and the disease progression of initial participants is continuously followed. Over time, some patients who were initially labeled as CN and MCI had a change in diagnosis as their disease progressed. While some patients had as many as 16 MRI scans since the start of the study, clinical evaluations were collected much less frequently, and genetic testing was only performed once per patient. Thus, combining 3 modalities per patient was a unique challenge as, at times, there were contradictory diagnoses, making the ground truth diagnosis unclear. For our overlap dataset, we used the latest MRI and clinical evaluation for each patient and the most recent diagnosis. Several studies focused on using time-series data to track the progression of the disease.[16,17,22–24] However, we aimed to accurately classify patients into groups at the most recent state of evaluation so our method can be generalized to patients who are not part of long-term observational studies.

Clinical data preprocessing

For clinical data, we used 2384 patients' data from the neurological exams (eg, balance test), cognitive assessments (eg, memory tests), and patient demographics (eg, age). The clinical data are quantitative, categorical, or binary, totaling 29 features. We removed any feature that could encode direct indication of AD (eg, medication that a patient takes to manage extant AD). A full list of features can be found in Supplementary Material S2. Categorical data were converted to features using one-hot encoding, and continuous-valued features were normalized.

Genetic data preprocessing

The genetic data consist of the whole genome sequencing data from 805 ADNI participants by Illumina's non-Clinical Laboratory Improvement Amendments. The resulting variant call files (VCFs) were generated by ADNI using Burrows–Wheeler Aligner and Genome Analysis Toolkit in 2014. Each subject has about 3 million SNPs in the raw VCF file generated. However, not all detected SNPs are informative in predicting AD. We followed the established preprocessing steps detailed in Venugopalan et al[19] to reduce the number of SNPs and keep only the relevant genetic components. After such filtering (detailed further in Supplementary Material S1), we had 547 863 SNPs per patient. As we only have 805 patients with genetic data, we were left with a highly sparse matrix. We used a Random Forest classifier as a supervised feature selection method to determine which are the most important features, reducing our feature space to roughly 15 000 SNPs. Note that data points used for model testing were not seen by the classifier. While the result was still sparse, we found that this level was a reasonable stopping point

Table 1. Number of participants in each modality and their diagnosis

	Total participants	Control	Mild cognitive impairment	Alzheimer's disease
Clinical	2384	942	796	646
Imaging	551	278	123	150
Genetic	805	241	318	246
Overlap	239	165	39	35

Note: This table shows the number of participants in each modality and further separates the participants into their diagnoses. The overlap section refers to patients who had all 3 modalities recorded.

as determined by the experiment detailed in Supplementary Material S1. The final data were grouped into 3 categories: no alleles, any 1 allele, or both alleles present.

Imaging data preprocessing

The imaging data in this study consist of cross-sectional MRI data corresponding to the first baseline screenings from ADNI1 (551 patients). The data publisher has standardized the images to eliminate the non-linearities caused by the scanners from different vendors. From each brain volume, we used 3 slices corresponding to the center of the brain in each dimension. An example input is shown in Figure 1. Further details on the ADNI preprocessing steps and experiments justifying the use of 3 images per patient can be found in Supplementary Material S3.

Finalizing the training dataset

To train our multimodal architecture, we used 239 patients who had data available from all 3 modalities. The overlap dataset was chosen out of the original data mentioned above—imaging (551 patients), SNP (805 patients), and clinical (2284 patients) to predict AD stages. While the SNP data were unique per patient, the clinical and imaging data appeared multiple times. To ensure a proper merger, we used the timestamps in the clinical data and matched it to the closest MRI scan date. Next, we used the most recent evaluation on a patient to avoid repeating patients. The patients' demographic information is shown in Table 2.

Multimodal framework

The proposed framework, MADDi, is shown in Figure 2. The model receives a patient's preprocessed clinical, genetic, and imaging data and outputs the corresponding diagnosis (CN, AD, or MCI). Following the input, there are modality-specific neural network architecture backbones developed in the single modality setting (further detailed in the Performance of Unimodal Models Section). For clinical and genetic data, this is a 3-layer fully connected network, and for imaging data, we have a 3-layer convolutional neural network. The output of those layers then enters a multi-headed self-attention layer, which allows the inputs to interact with each other and find what features should be paid most attention to within each modality. This layer is followed by a cross-modal bi-directional attention layer,[21] which performs a similar calculation to self-attention but across different pairs of modalities. The purpose of cross-modal attention is to model an interaction between modalities; for example, clinical features may help reinforce what the imaging features tell the model and thus lead to more robust decision making. Both attention types are rigorously defined in the Neural Network Attention Section. The last step concatenates the output of the parallel attention computations and feeds it into a final dense layer that makes the prediction.

Experimental design

Neural network attention

As a part of our experimental design, we evaluate the importance of attention in our model. Previous methods[16,17,19] explored the value that DL brings to automating AD diagnosis. We build on top of previous multimodal DL frameworks and examine the need for intermodal interactions through attention. Thus, we used the same framework but toggled the presence of attention based on 4 criteria: self-attention and cross-modal attention, just self-attention, just cross-modal attention, and no attention. The different types of attention are introduced in the following.

Generalized attention. Attention is a mechanism that captures the relationship between 2 states and highlights the features that contribute most to decision-making. An attention layer takes as input queries and keys of dimension d_k, and values of dimension d_v. A key is the label of a feature used to distinguish it from other features, and a query is what checks all available keys and selects the one that matches best. We compute the dot products of the query with all keys, divide each by the square root of d_k, and apply a Softmax function, which converts a vector of numbers into a vector of probabilities, to obtain the weights on the values:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Following the success of the Transformer,[25] we use the multi-head attention module, which allows the model to jointly attend to information from different representation subspaces at different positions.

Self-attention. For self-attention mechanisms, queries, keys, and values are equal. The self-attention mechanism allows us to learn the interactions among the different parts of an input ("self") and determine which parts of the input are relevant for making predictions ("attention"). In our case, the prior neural network layers produce in parallel 3 latent feature matrices for each modality that act as the keys, queries, and values: imaging matrix I, genetic matrix G, and clinical matrix C. Self-attention, in our terms, refers to attention computation done within the same modality. Thus the self-attention module performs the following operations:

$$\text{self} - \text{attention}(I \rightarrow I) \tag{2}$$

$$\text{self} - \text{attention}(G \rightarrow G) \tag{3}$$

$$\text{self} - \text{attention}(C \rightarrow C) \tag{4}$$

Cross-modal attention. In each bi-directional cross-modal attention layer,[21] there are 2 unidirectional cross-modal attention sublayers: one from modality 1 to modality 2 and one from modality 2 to modality 1.

In our case, the cross-modal attention layer takes the output of each self-attention computation: image self-attention output I_s, genetic self-attention output G_s, and clinical self-attention output

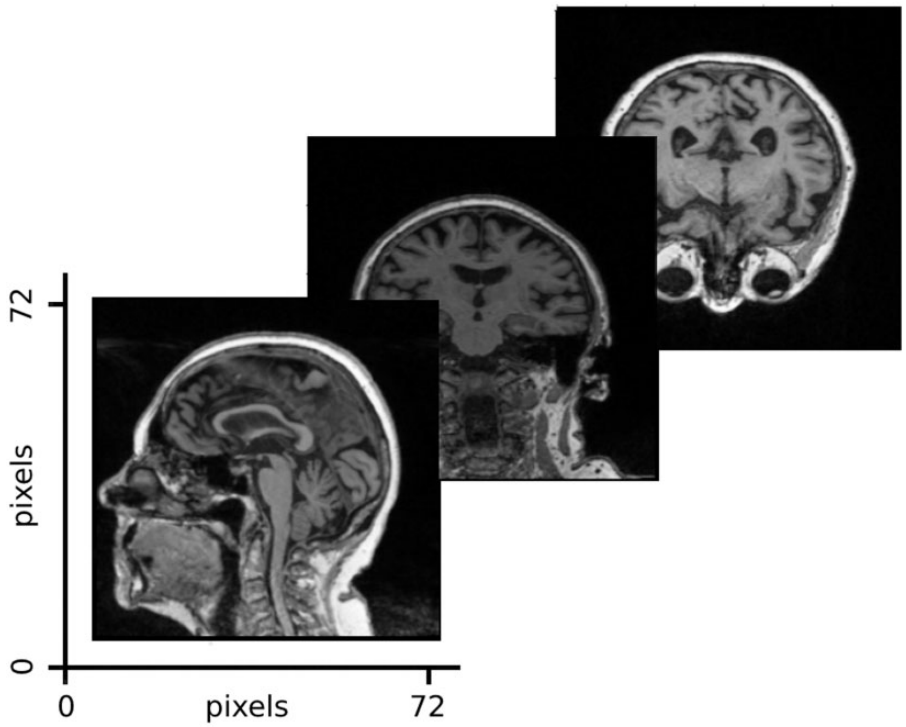

Figure 1. Imaging input example. The imaging model took as input 3 slices from the center of the MRI brain volume, which were uniformly shaped to 72 × 72 pixels.

Table 2. Participants' demographic information

Group	Participants (n)	Female sex (%)	Mean age (years)
Control	165	53.9	77.8
Mild cognitive impairment	39	34.2	76.6
Alzheimer's disease	35	31.4	78.1

Note: This table shows the number of participants in each diagnosis group along with the percent of females and the average age of each group.

C_s. Thus the cross-modal attention module performs the following operations:

$$\text{concatenation}(\text{cross}-\text{modal attention}(I_s \rightarrow C_s), \text{cross}-\text{modal attention}(C_s \rightarrow I_s)) \quad (5)$$

$$\text{concatenation}(\text{cross}-\text{modal attention}(C_s \rightarrow G_s), \text{cross}-\text{modal attention}(G_s \rightarrow C_s)) \quad (6)$$

$$\text{concatenation}(\text{cross}-\text{modal attention}(G_s \rightarrow I_s), \text{cross}-\text{modal attention}(I_s \rightarrow G_s)) \quad (7)$$

Finally, we created a model with no attention module at all, where, following the neural network layers, we directly proceed to concatenate and produce output through the final dense layer. This setting represents the previous state-of-the-art methods used for integrating multimodal datasets for our task.

Unified hyperparameter tuning scheme

The modality-specific neural network part of MADDi was predetermined based on the hyperparameter tuning done on each unimodal model (Supplementary Material S4). We did not use the overlapping test set during hyperparameter tuning was done. To fairly evaluate the need for attention, we tuned using the same hyperparameter grid for each of the other experimental models. Meaning, that each model (self-attention only, the cross-modal attention only, and the model with no attention) gets its own set of best hyperparameters. We first randomly split our 239 patients into a training set (90%) and a held-out testing set (10%). We chose a 90–10 split for consistency with all the papers we compared against (shown in Table 3). We designed a 3-fold cross-validation scheme and took the parameters that gave the best average cross-validation accuracy. Next, we used 5 random seeds to give the model multiple attempts at initialization. The best initialization was determined based on the best training performance on the full train and validations set (ie, validation was added into training). This pipeline was repeated to find until we found the best parameters for each baseline.

Evaluation metrics

The following metrics were calculated for each model: accuracy, precision (positive predictive value), recall (sensitivity), and F1-score (harmonic mean of recall and precision). F1-score was the primary performance metric for evaluating our baselines. Accuracy was used to evaluate our best model against previous papers, as that was the metric most commonly reported on this task. The metric calculations are detailed in Supplementary Material S5.

RESULTS

Performance of unimodal models

To demonstrate the success of our multimodal framework, MADDi, we first examined the performance of a single-modality model. Our evaluation pipeline was consistent across all modalities in that we used a neural network and then tuned hyperparameters to find the best model. We split each modality into training (90%) and testing (10%) data, where the testing set was not seen until after the best

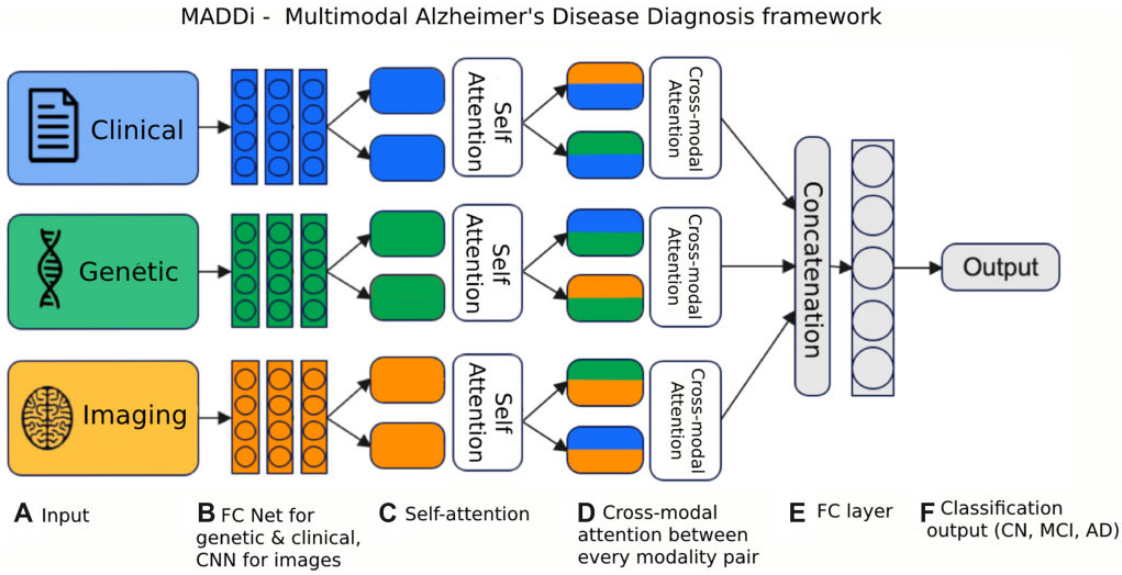

Figure 2. Model architecture. (A) Data inputs—clinical data (demographics, memory tests, balance score, etc.), genetic (SNPs), and imaging (MRI scans). (B) The input sources are combined and fed into a fully connected (FC) neural network architecture for genetic and clinical modalities and a convolutional neural network (CNN) for imaging data. (C) Using the obtained features from the neural networks, a self-attention layer reinforces any inner-modal connections. (D) Then, each modality pair is fed to a bi-directional cross-modal attention layer which captures the interactions between modalities. (E) Finally, the outputs are concatenated and passed into a decision layer for classification into the (F) output Alzheimer's stages (CN, MCI, and AD).

Table 3. Comparison with related studies

Study	Modality	Accuracy	F1-score	Method
Bucholc et al, 2019[18]	MRI, PET, clinical	82.90%	Not reported	SVM
Fang et al 2020[26]	MRI, PET	66.29%	Not reported	GDCA
Abuhmed et al, 2021[17]	MRI, PET, clinical	86.08%	87.67%	Multivariate BiLSTM
Venugopalan et al, 2021[19]	MRI, SNP, clinical	78%	78%	DL + RF
MADDi	**MRI, SNP, clinical**	**96.88%**	**91.41%**	**DL + attention**

Note: This table shows the comparison between our study and 5 other previous studies that attempted to solve a similar problem to ours. MADDi performed with 96.88% average accuracy and 91.41% average F1-score across 5 random initializations on a held-out test set, achieving state-of-the-art performance on the multimodal 3-class classification task.

parameters were chosen using the average accuracy across 3-fold cross-validation. The reported test accuracies are averaged across 5 random initializations, which remained consistent across all modalities. The results are summarized in Figure 3 (and in Supplementary Material S6, Table S3). For the clinical unimodal model, we created a neural network with 3 fully connected layers (other hyperparameters can be found in Supplementary Material S4). The best model yielded 80.5% average accuracy across 5 random seeds. The model was trained on 2384 patients. For imaging results, we created a convolutional neural network with 3 convolution layers. The best model yielded 92.28% average accuracy across 5 random seeds. The model was trained on 551 patients, but we allowed for patient repetitions as we found that only using 551 images was not enough to train a successful model. Thus, we had 3674 MRI scans from 551 patients (some patients repeated up to 16 times). We selected our testing set such that it has 367 unique MRIs (10% of training), and we do not allow for any repeating patients in the testing set. We only allowed repetition during training, and no patient was included in both training and testing sets. For genetic data, we created a neural network with 3 fully connected layers. The best model yielded 77.78% average accuracy across 5 random seeds. The model was trained on 805 unique patients.

Performance of multimodal models

Table 3 contrasts the performance and architecture of MADDi with state-of-the-art multimodal approaches from the literature. Note that due to the differences in dataset characteristics and multitask settings, it was not possible to directly compare performance among methods that only report binary classification or use a single modality. Thus, we only report studies that used 2 or more modalities and did 3-class classification. For our proposed method, we report the average accuracy across 5 random initializations on a held-out test set. Therefore, we also use the test (as opposed to cross-validation) accuracy from other studies. Bucholc et al[18] used support vector machines to classify patients into normal, questionable cognitive impairment, and mild/moderate AD, comparable to our definitions of CN, MCI, and AD. They reported 82.9% test accuracy but did not rely on DL. Fang et al[26] used Gaussian discriminative component analysis as a method of multi-class classification using 2 different imaging modalities, achieving 66.29% accuracy on the test set. Abuhmed et al and El-Sappagh et al[16,17] both used MRI, PET, and various health record features. The key difference between the 2 is that El-Sappagh et al considered a 4-class classification of CN, AD, stable MCI (patients who do not progress to AD), and progressive MCI. Since they did not report 3-class classification, we could not

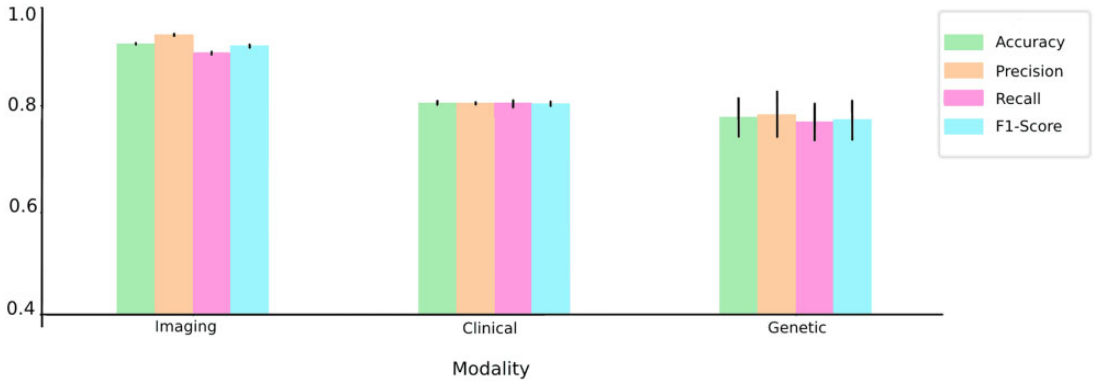

Figure 3. Metric evaluation of unimodal models. The graph shows all 4 evaluation metrics for the best neural network model for each modality—imaging, clinical, and genetic. The imaging model gives the best performance overall, whereas the genetic modality gives the lowest performance with highest variation.

directly compare to their work, but note that they achieved 92.62% accuracy on the 4-class task. Both methods utilized DL, but they focused more on disease progression diagnosis with time-series data rather than static disease diagnosis. Venugopalan et al[19] were most similar to our study in structure, modality choice, and preprocessing. They, too, did not utilize the recent advancement of attention-based multimodal learning, which is where our architecture stands out. At 96.88%±3.33% average accuracy, MADDi defined state-of-the-art performance on the multimodal 3-class classification task.

Model robustness

To definitively conclude that both self-attention and cross-modal attention are necessary additions to the model, we ablated the attention schemes in 3 conditions (self-attention only, cross-modal attention only, and the model with no attention) on the held-out test set using the best parameters for each respective model. To demonstrate that our success was not dependent on initialization, we used 100 different random seeds and recorded the distribution of F1-scores on the testing set. For more information on our test sample selection, refer to Supplementary Material. Figure 4 (and Supplementary Table S4 in Supplementary Material S7) shows that self-attention and cross-modal attention layers together have the narrowest distribution, with the highest median F1-score. The next best distribution is the cross-modal attention layer alone, which has a slightly wider distribution but still the second-highest median F1-score. The success of the 2 methods involving cross-modal attention becomes apparent and provides strong evidence that capturing interactions between modalities positively influences the model's decision-making. All 3 models that utilize attention achieved 100% F1-score for at least one initialization, while the model with no attention layers only reached at most 92.8% F1-score. Furthermore, the performance of our final model was 7.9% average F1-score higher than a model with no attention, and was significant ($P < .0001$, 2-sample Z-test)—providing further evidence that attention was beneficial for multimodal data integration.

Using the self-attention and cross-attention model (MADDi), we investigated the performance of the model with respect to the individual classes as seen in Table 4. We report metrics averaged across 5 random initializations. We find that, regardless of the initialization, the model is extremely accurate at identifying AD patients. However, for some cases, it tends to mistake MCI patients for CNs. We hypothesize that, since our data does not include different stages of MCI, it may have MCI patients with mild symptoms that could be mistaken for CNs by the model. These observations can be seen in detail through 5 confusion matrices from the 5 initializations in the Supplementary Material S8, Figure S4.

Modality importance

Finally, we investigated the importance of each modality to bring more transparency into the model's decision-making and motivate future data collection efforts. Knowing how valuable each modality is to disease classification and what happens when it is excluded from the experiment can shed light on where to focus scientific resources. While every study participant had at least some clinical data available, only a few hundred patients had MRI scans. To fairly compare each modality's importance to the model, we performed our analyses on the same exact patients. Thus, we evaluated the contribution of the modalities on the overlap patient set (detailed in Figure 5). For single modalities, we only used self-attention. For a pair of modalities, we used both self-attention and cross-modal attention. We performed hyperparameter tuning for each model to ensure fair evaluation, with all the parameters provided in Supplementary Material S4, Table S1. As seen in Figure 5, combining the 3 modalities performs the best across all evaluation metrics. A full table with the numeric results can be found in Supplementary Material S10, Table S6. The interesting discovery here was the clinical modality contribution to this performance. While the use of 2 modalities was better than one in most cases, when clinical data were withheld, we saw a significant drop in performance; clinical data alone achieved 82.29% accuracy and 78.30% F1-score, whereas genetic and imaging together achieved 78.33% accuracy and dropped to 50.07% F1-score. These results suggest that the clinical dataset is an important catalyst modality for AD prediction. We hypothesize that this empirical merit stems from the fact that clinical features offer the necessary patient context that grounds the additional modalities such as vision or omics information and helps the model in their effective representation and interpretation.

To further investigate the clinical data, we used a Random Forest classifier (which fits the clinical training data to the diagnosis labels in a supervised manner) to find the most dominant features from the clinical modality: memory score, executive function score, and language score. A full list of features in order of importance can be found in Supplementary Material S2, Figure S1.

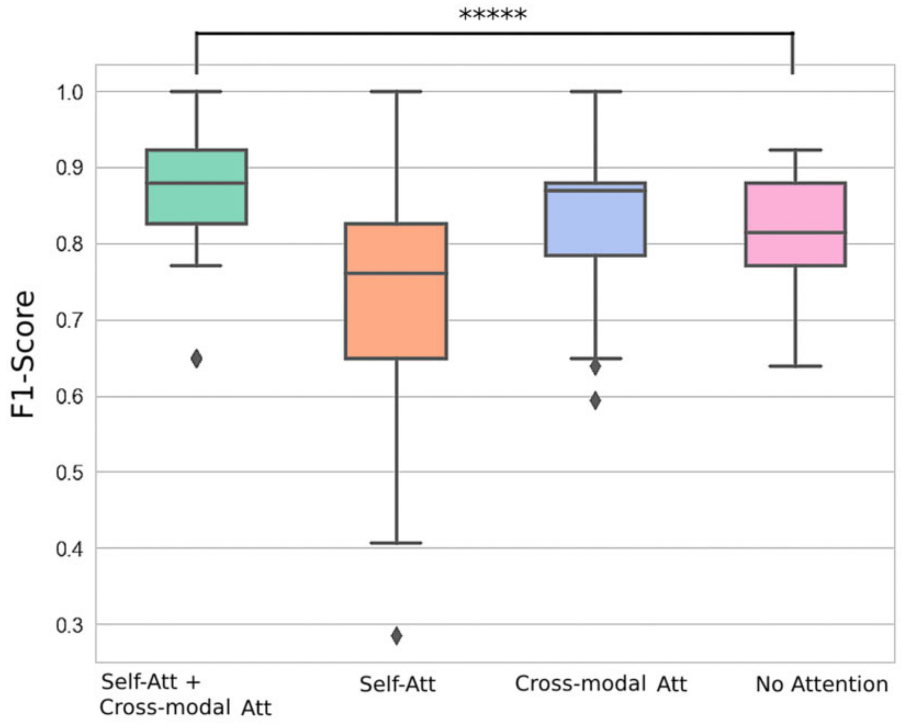

Figure 4. F1-score distribution for different attention-based and attention-free baselines. Box plots showing the F1-score distribution across 100 random seeds to demonstrate the value of attention in a deep learning model. The F1-scores were calculated from a held-out test set. The horizontal line represents the median F1-score, and the boxes represent the first and third quartiles. The whiskers represent quartile 1 – (1.5 × interquartile range) and quartile 3 + (1.5 × interquartile range). The dots represent the individual F1-scores for each model. ****$P \leq .0001$. The figure demonstrates that the combination of self-attention with cross-modal attention performs the best with the most narrow variation.

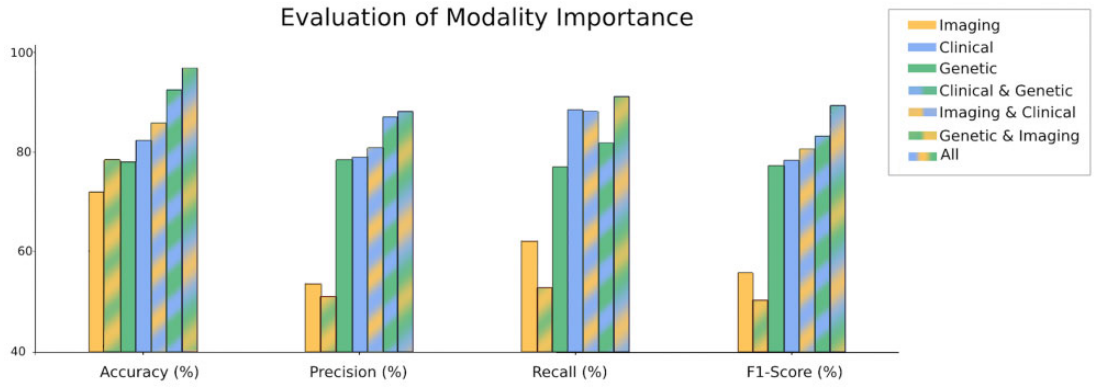

Figure 5. Evaluation of modality importance. This figure evaluates the possible combinations of modalities. The metrics were calculated as an average of 5 random initializations on a held-out test set. The combination of the 3 modalities performs the best across all evaluation metrics. Excluding the clinical modality causes a drop in performance, demonstrating the value of clinical information.

DISCUSSION

Clinical importance and implications

Detecting AD accurately is clinically important as it is the sixth leading cause of death in the United States and is the most common cause of dementia.[1] Without treatment or other changes to the trajectory, aggregate care costs for people with AD are projected to increase from $203 billion in 2013 to $1.2 trillion in 2050.[2] Despite studies such as ADNI collecting various imaging, genetic, and clinical data to improve our understanding of the underlying disease processes, most computational techniques still focus on using just a single modality to aid disease diagnosis. Our state-of-the-art model allows for an effective integration scheme of 3 modalities and can be expanded as more data sources become available.

Future extensions

The proposed model architecture can be used in other multimodal clinical applications, such as cancer detection.[27,28] As the efforts to make health care data more broadly available continue to increase, we believe that our model will help aid the diagnostic process. The framework we propose does not rely on modality-specific processing within the model itself. Thus, our future work will include other data (PET scans, clinical notes, biomarkers, etc.). While it is

Table 4. Investigating performance for each individual class

	Accuracy	Precision	Recall	F1-Score
Control	96.66%	96.78%	98.88%	97.81%
Moderate cognitive impairment	96.66%	90.00%	70.00%	76.66%
Alzheimer's disease	100%	100%	100%	100%

Note: This table shows the performance metrics averaged across 5 random initializations of MADDi on each class (control, Moderate Cognitive Impairment, and Alzheimer's disease). We observe that Alzheimer's disease is predicted correctly regardless of initialization and the only mistake the model makes is misclassifying MCI patients as control patients.

straightforward to simply interchange the current modalities with new ones and only use 3 modalities at a time, we plan on expanding our work beyond this current level as there is no rigid constraint on the number of modalities used with the given codebase. Furthermore, similar to the task El-Sappagh et al[16] explored, we will extend our task to more than 3-class classification and use our work to detect different types of MCI (stable MCI and progressive MCI). This will help better understand AD progression and delay the change from MCI to AD.

Limitations

When creating our unimodal performance baselines, we often struggled with finding the ground truth labels for the genetic data. While every patient had a diagnosis attached to an MRI scan, and most of the clinical exams also had a diagnosis listed, genetic data did not. Out of the 808 patients with genetic data, we used 805 patients where diagnosis on their most recent MRI scan agreed with their clinical diagnosis. Thus, we proceeded with 805 patients to eliminate any error in the ground truth labeling. This gap is natural, as a patient may have had a more recent MRI that changed the diagnosis, leaving the recent clinical evaluation outdated (and vice versa).

During preprocessing of the MRI images, we chose to use the middle slice of the brain rather than the full brain volume. This could mean that our model did not see certain areas of the brain. When running unimodal experiments on the MRI data, the performance remained the same (within 1%) when using just the middle slice of the brain compared to the full brain volume, shown in the Supplementary Material S3. Since processing thousands of slices per patient is much more computationally expensive, we proceeded with this simplification. While on our task, there was no significant difference in performance; in other applications, integrating the full brain volumes into the model could further increase performance.

CONCLUSIONS

In this work, we presented a MADDi, which uses attention-based DL to detect AD. The performance of MADDi was superior to that of existing multimodal machine learning methods and was shown to be consistently high regardless of chance initialization. We offer 3 distinct contributions: integrating multimodal inputs, multi-task classification, and cross-modal attention for capturing interactions. Many existing multimodal DL models simply concatenate each modality's features despite substantial differences in feature spaces. We employed attention modules to address this problem; self-attention reinforces the most important features extracted from the neural network backbones, and cross-modality attention[21] reinforces relationships between modalities. Combining the 2 attention modules resulted in a 96.88% accuracy and defined state-of-the-art on this task. Overall, we believe that our approach demonstrates the potential for an automated and accurate DL method for disease diagnosis. We hope that in the future, our work can be used to integrate multiple modalities in clinical settings and introduce the highly effective attention-based models to the medical community.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS

All authors contributed to the design of the methodology and the experiments. MG implemented the data preprocessing, modeling, and data analysis. All authors discussed the results and contributed to the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We thank Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu/) for providing data for this study. For a complete list of ADNI investigators, refer to: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. We acknowledge Pinar Demetci for her help in the discussion of genetic data preprocessing.

CONFLICT OF INTEREST STATEMENT

None declared.

CODE AVAILABILITY

Available at: https://github.com/rsinghlab/MADDi.

DATA AVAILABILITY

The data underlying this article were provided by Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu/) by permission. Data will be shared on request to ADNI.

REFERENCES

1. Naqvi E. *Alzheimer's Disease Statistics*. 2017. https://alzheimersnewstoday.com/alzheimers-disease-statistics/. Accessed June 20, 2022.
2. Thies W, Bleiler L. 2013 *Alzheimer's Disease Facts and Figures*. Wiley Online Library; 2013. https://alz-journals.onlinelibrary.wiley.com/doi/10.1016/j.jalz.2013.02.003. Accessed June 20, 2022.
3. Iddi S, Li D, Aisen PS, Rafii MS, Thompson WK, Donohue MC; Alzheimer's Disease Neuroimaging Initiative. Predicting the course of Alzheimer's progression. *Brain Inform* 2019; 6 (1): 6.
4. Petersen RC. Mild cognitive impairment. *Continuum (Minneap Minn)* 2016; 22 (2 Dementia): 404–18.

5. Mueller SG, Weiner MW, Thal LJ, *et al.* Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). *Alzheimer's Dement* 2005; 1 (1): 55–66.
6. Forouzannezhad P, Abbaspour A, Li C, *et al.* A Gaussian-based model for early detection of mild cognitive impairment using multimodal neuroimaging. *J Neurosci Methods* 2020; 333: 108544.
7. Uysal G, Ozturk M. Hippocampal atrophy based Alzheimer's disease diagnosis via machine learning methods. *J Neurosci Methods* 2020; 337: 108669.
8. Dimitriadis SI, Liparas D, Tsolaki MN; Alzheimer's Disease Neuroimaging Initiative. Random forest feature selection, fusion and ensemble strategy: combining multiple morphological MRI measures to discriminate among healhy elderly, MCI, cMCI and Alzheimer's disease patients: from the Alzheimer's disease neuroimaging initiative (ADNI) database. *J Neurosci Methods* 2018; 302: 14–23.
9. Beheshti I, Demirel H, Matsuda H; Alzheimer's Disease Neuroimaging Initiative. Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. *Comput Biol Med* 2017; 83: 109–19.
10. Dyrba M, Barkhof F, Fellgiebel A, *et al.*; EDSD Study Group. Predicting prodromal Alzheimer's disease in subjects with mild cognitive impairment using machine learning classification of multimodal multicenter diffusion-tensor and magnetic resonance imaging data. *J Neuroimaging* 2015; 25 (5): 738–47.
11. El-Sappagh S, Saleh H, Sahal R, *et al.* Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data. *Future Gener Comput Syst* 2021; 115: 680–99.
12. Zhou J, Liu J, Narayan VA, Ye J; Alzheimer's Disease Neuroimaging Initiative. Modeling disease progression via multi-task learning. *Neuroimage* 2013; 78: 233–48.
13. Wang H, Shen Y, Wang S, *et al.* Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease. *Neurocomputing* 2019; 333: 145–56.
14. Kruthika KR, Maheshappa HD. Multistage classifier-based approach for Alzheimer's disease prediction and retrieval. *Inform Med Unlocked* 2019; 14: 34–42.
15. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. *Brief Bioinform* 2022; 23 (2): bbab569.
16. El-Sappagh S, Abuhmed T, Riazul Islam S, Kwak KS. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. *Neurocomputing* 2020; 412: 197–215.
17. Abuhmed T, El-Sappagh S, Alonso JM. Robust hybrid deep learning models for Alzheimer's progression detection. *Knowl Based Syst* 2021; 213: 106688.
18. Bucholc M, Ding X, Wang H, *et al.* A practical computerized decision support system for predicting the severity of alzheimer's disease of an individual. *Expert Syst Appl* 2019; 130: 157–71.
19. Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for early detection of alzheimer's disease stage. *Sci Rep* 2021; 11 (1): 1–13.
20. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011; June 28–July 2, 2011: 689–96; Bellevue, WA.
21. Tan HH, Bansal M. Lxmert: learning cross-modality encoder representations from transformers. *ArXiv*, vol. abs/1908.07490; 2019. doi:10.48550/ARXIV.1908.07490.
22. Guerrero R, Schmidt-Richberg A, Ledig C, Tong T, Wolz R, Rueckert D; Alzheimer's Disease Neuroimaging Initiative (ADNI). Instantiated mixed effects modeling of Alzheimer's disease markers. *Neuroimage* 2016; 142: 113–25.
23. Jedynak BM, Lang A, Liu B, *et al.*; Alzheimer's Disease Neuroimaging Initiative. A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease neuroimaging initiative cohort. *Neuroimage* 2012; 63 (3): 1478–86.
24. Yau W-YW, Tudorascu DL, McDade EM, *et al.* Longitudinal assessment of neuroimaging and clinical markers in autosomal dominant Alzheimer's disease: a prospective cohort study. *Lancet Neurol* 2015; 14 (8): 804–13.
25. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *ArXiv*, vol. abs/1706.03762; 2017. doi:10.48550/ARXIV.1706.03762.
26. Fang C, Li C, Forouzannezhad P, *et al.*; Alzheimer's Disease Neuroimaging Initiative. Gaussian discriminative component analysis for early detection of Alzheimer's disease: a supervised dimensionality reduction algorithm. *J Neurosci Methods* 2020; 344: 108856.
27. Esteva A, Kuprel B, Novoa RA, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542 (7639): 115–8.
28. Weng S, Xu X, Li J, Wong ST. Combining deep learning and coherent anti-Stokes Raman scattering imaging for automated differential diagnosis of lung cancer. *J Biomed Opt* 2017; 22 (10): 1–10.