Exploring Multimodal Fusion Strategies for Alzheimer's Detection

Lola Kovalski

Abstract

Early, accurate, and explainable diagnosis of Alzheimer's Disease (AD) is critical for effective treatment planning, yet both clinicians and machine learning models struggle with borderline cases of cognitive impairment. In this study, we investigate how multimodal learning models can integrate structural MRI scans with tabular clinical and demographic data to improve detection of cognitive decline. Using the OASIS-1 dataset, we compare three modeling approaches: (1) unimodal baselines, (2) early fusion, and (3) late fusion. MRI data is processed as 2D slices using convolutional neural networks, while tabular features – including MMSE, age, gender, handedness, and education – are modeled with depth-constrained random forests. We find that late fusion models achieve the highest overall accuracy and exhibit strong robustness to feature ablation. By contrast, early fusion performs inconsistently, highlighting the challenges of combining heterogeneous modalities at the feature level in limited data settings.

1 Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that impairs memory and cognition, with particularly high prevalence in older adults. While structural MRI is widely used to detect patterns of brain atrophy associated with cognitive decline, demographic and clinical features – such as age, cognitive test scores, and education – offer essential contextual information (Beheshti et al., 2017; Shen et al., 2017). Machine learning models that integrate these heterogeneous data types may enhance diagnostic accuracy, especially in borderline or ambiguous cases where single-modality models struggle.

We use the OASIS-1 dataset, which includes MRI scans and clinical metadata for 416 adult participants, 100 of whom have been diagnosed with very mild or mild AD. To assess the contributions of each modality and integration strategy, we compare three modeling approaches: (1) a unimodal baseline using either MRI or tabular data, (2) early fusion of image and tabular features, (3) and late fusion of modality-specific model outputs (Ngiam et al., 2011; Sagi and Rokach, 2018). Structured clinical features – including MMSE scores, age, gender, handedness, and education – are modeled using depth-constrained random forests (Breiman, 2001), while MRI data are processed using convolutional neural networks trained on 2D slices (Simonyan and Zisserman, 2015).

Our evaluation focuses on diagnostic accuracy, with particular attention to mild cognitive impairment cases, where multimodal modeling may offer the greatest advantage. By systematically comparing fusion strategies using the same preprocessing pipeline and model architectures, we explore how integration choices impact predictive performance and model interpretability.

2 Preliminaries

Our objective is to train a multimodal binary classifier that predicts whether a subject show signs of cognitive impairment. We use three models – random forests, convolutional neural networks (CNNs), and multilayer perceptrons (MLPs) – along with two integration strategies – early fusion and late fusion.

2.1 Random Forest Classifier

A random forest is an ensemble of decision trees trained on bootstrapped subsets of data (Breiman, 2001). In the `scikit-learn` implementation, which we make use of, each tree partitions the input space by recursively selecting splits that minimize the Gini impurity:

$$G = 1 - \sum_{k=1}^{K} p_k^2$$

where p_k is the proportion of class k in a node. Final predictions are made by majority vote across all trees. To reduce the risk of overfitting, we constrain the depth and number of features of the tree.

We chose random forests as our tabular baseline due to their strong empirical performance on structured data, resilience to overfitting in small datasets, and minimal preprocessing requirements. Their ensemble structure captures non-linear feature interactions. crucial when modeling subtle relationships among clinical variables. Moreover, they naturally handle mixed data types and provide interpretability through feature importance metrics, making them particularly attractive in medical con-

texts where understanding model rationale is vital. Compared to simpler models such as logistic regression, random forests offer a more flexible decision boundary, while avoiding the sensitivity to noise and overfitting often seen in deeper models when data is limited.

2.2 Convolutional Neural Networks (CNNs)

Convolutional neural networks are a class of neural networks particularly effectively for image classification. Given a 2D MRI slice $x_i \in R^{H \times W}$, a CNN processes the image through a series of layers designed to extract and summarize spatial features (Hopkins, 2024).

The first layers apply convolutional filters, which are small learnable matrices that slide over the input image and compute local dot products. Each filter detects a specific pattern such as an edge, texture, or anatomical boundary. The output of the convolutional layer is then passed through non-linear activation functions (typically ReLU) and pooling layers (typically max pooling). Note that the order, number of, and size of these convolutional/pooling/activation layers are implementation dependent.

The Rectified Linear Unit (ReLU) activation function is defined as:

$$\text{ReLU}(x) = \max(0, x)$$

ReLU introduces non-linearity by zeroing out negative values, allowing the network to model complex patterns while mitigating the vanishing gradient problem.

As the network deepens, successive layers capture increasingly abstract features, starting from local textures and progressing to complex structures. After the final convolutional layers, the spatial outputs are flattened into a 1D feature vector and passed through fully connected layers which output logits $z_i = (z_{i,0}, z_{i,1})$, in the case of binary classification, representing the model's unnormalized confidence in each class.

The logits are then converted to class probabilities using the softmax function, which is equivalent to using the sigmoid function for binary classification tasks.

$$\text{softmax}(z_{i,j}) = \frac{e^{z_{i,j}}}{\sum_k e^{z_{i,k}}}$$

For this project, we use a VGG-16 CNN (Simonyan and Zisserman, 2015) pretrained on ImageNet and fine-tuned on our MRI data. The model is trained using cross-entropy loss:

$$L = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{e^{z_i y_i}}{\sum_j e^{z_{i,j}}} \right),$$

where weights $(w_k = (\sum_j c_j - c_k)/c_k)$ are applied to compensate for class imbalance.

2.3 Multilayer Perceptrons (MLPs)

A multilayer perceptron is a fully connected neural network used to model relationships between input features and output class. Unlike CNNs, which exploit spatial structure, MLPs treat inputs as flat vectors. Given an input $x \in R^d$, the MLP passes it through a sequence of linear transformations followed by non-linear activation functions (Bhattacharya, 2025).

The MLP computes a sequence of hidden layer representations:

$$h^{(1)} = \text{ReLU}(W^{(1)}x + b^{(1)}), h^{(2)} = \text{ReLU}(W^{(2)}h^{(1)} + b^{(2)}),$$

where each $h^{(l)}$ represents the output of a hidden layer. The final output layer produces unnormalized logits, which are converted to class probabilities using the softmax/sigmoid function:

$$\hat{y} = \text{softmax}(W^{(3)}h^{(2)} + b^{(3)})$$

We use MLPs in two parts of our pipeline, early fusion and stacking.

2.4 Fusion Strategies

Early fusion combines modalities at the input level. We extract an embedding from the CNN and concatenate it with tabular features to form a single vector, which is passed through a MLP. This setup allows for joint representation learning but may generalize poorly when data is limited due to the challenge of reconciling heterogeneous inputs.(Barnum et al., 2020).

Late fusion combines predictions rather than raw features. We train separate models on each modality – a CNN for MRI and a random forest for tabular features – and integrate their predicted class probabilities using several strategies. These include simple averaging, weighted averaging, and stacking meta-learners. This approach avoids parameter sharing across modality and is robust to missing or noisy modalities at test time (Ngiam et al., 2011; Huang et al., 2020).

Stacking, treated as a form of late fusion, uses a meta-learner to combine the output probabilities from each unimodal model. We experiment with logistic regression, random forests, and MLPs as meta-classifiers. These are fine tuned to the training set to optimally reweight base model predictions, allowing the ensemble to adaptively leverage each modality depending on the prediction context (Sagi and Rokach, 2018).

3 Data

We use the OASIS-1 cross-sectional dataset, which includes MRI and demographic data for 416 adult participants aged 18 to 96. Of these, 100 participants over age 60 have been clinically diagnosed with very mild to mild Alzheimer's Disease (AD).

3.1 Sample Inclusion Criteria

Subjects were included if both MRI imaging data and corresponding metadata were available. Approximately 60 participants were excluded due to missing MRI scans or missing tabular resulting in a final sample of 355. Among these, the mean age was 53.6 (SD = 25.3); 224 (63.1%) of observations were female; and 87 (24.5%)

have AD. Across the board, the relative proportions are consistent with the full $n = 416$ sample.

3.2 MRI Imaging Data

We use the processed (i.e. atlas-aligned, gain-field and bias-field corrected, 1mm^3 isotropic resolution) T1 weighted scans stored in Analyze 7.5 format. Each 3D MRI is loaded using the `nibabel` library and converted into a tensor. We then extract 9 representative 2D slices – three each from the axial, sagittal, and coronal planes at 40%, 50%, 60% depth. See fig. 1 for an example of a randomly selected patient. Each slice is then normalized to $[0, 1]$ and resized to 224×224 pixels. The slices are then stacked into a final tensor of shape $[1 \times 9 \times 224 \times 224]$, where the first dimension represents a single-channel grayscale image. As described in later sections, we ultimately reduce to a single representative slice $[1 \times 224 \times 224]$ to accommodate our CNN.

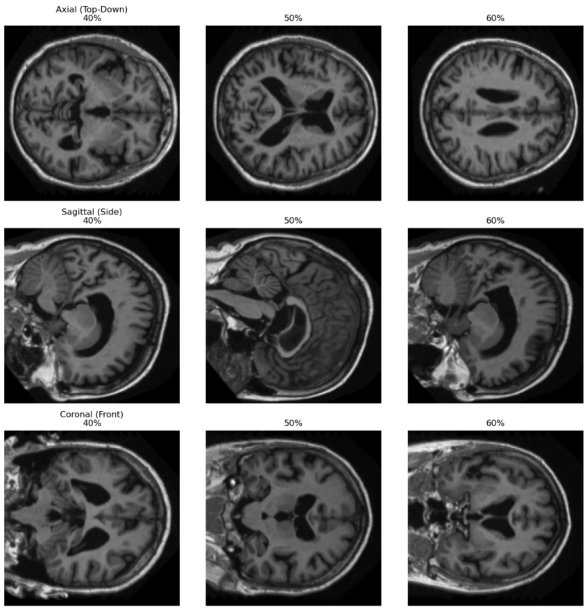

Figure 1: Extracted MRI slices from randomly selected patient

3.3 Tabular Features

Each subject's demographic, cognitive, and anatomical information is parsed from a text metadata file. We use the following features: age, sex (1 = female, 0 = male), education (years), MMSE (mini mental state examination score), handedness (right or left), estimated total intracranial volume (eTIV), atlas scaling factor (ASF), normalized whole brain volume (nWBV), and clinical dementia rating (CDR).

As of now, we do not include eTIV, ASF, and nWBV as features for the model. To construct labels, we use CDR > 0 to classify an observation as having AD (label = 1) and CDR = 0 as not (label = 0). In reality, the CDR is on a scale of 0–3: 0 = no dementia, 0.5 = questionable dementia, 1 = mild cognitive impairment (MCI), 2 = moderate, 3 = severe. For binary classification, we use CDR = 0 as the negative class and CDR > 0 as the positive class.

If a subject is missing a value of a feature (i.e. there is not education reported), we replace with the mean value of that feature across the dataset. Subjects with missing values for all features are excluded and part of the aforementioned ~ 60 omitted subjects.

3.4 Final Dataset and Loading

Each subject's MRI tensor and tabular features are paired with their binary outcome label. The dataset is randomly split into a 70/15/15 training/val/test split.[1] Data is loaded using PyTorch's `DataLoader` enabling batched processing and multiprocessing for efficient model training. [2]

4 Unimodal Models

Before exploring multimodal fusion strategies, we establish baseline performance using single-modality models. These unimodal classifiers allow us to assess the independent predictive power of MRI imaging and tabular clinical data. In addition to serving as performance anchors, these models are foundational to the multimodal models.

4.1 Tabular Baseline – Random Forest

est	depth	feat	Train F1	Val F1
200	5	None	0.957	0.667
100	5	sqrt	0.957	0.667
200	10	sqrt	1.000	0.667
200	5	sqrt	0.948	0.667
50	15	sqrt	1.000	0.609

Table 1: Top performing random forest hyperparameter configurations ranked by F1 score (validation set)

We use random forests to model Alzheimer's risk based solely on tabular clinical features. To optimize performance, we perform a grid search over key hyperparameters: number of estimators (n_estimates), tree depth (max_depth), and the number of features considered at each split (max_features).

Table 1 summarizes the top-performing random forest hyperparameter combinations. Based on training set F1 scores, we selected the top 10 performing hyperparameter combinations. Among these high capacity models,

[1] Given the skewed distribution of CDR scores, a future refinement might be to stratify observations across splits so we maintain a proportional representation of the different scores. Many of the random val and test splits fail to include an observation from each possible CDR (typically CDR = 2, which is rare in this dataset).

[2] If you are trying to replicate results *be careful with dataloaders* as the instantiation of one dataloader object might be different than another (i.e. training set is not always the same, so don't use dataloaderX to train the tabular baseline and dataloaderY to train the MRI model as it will not work and/or lead to data leaks). As you will see in my code, this is why I had several disgusting jupyter notebooks.

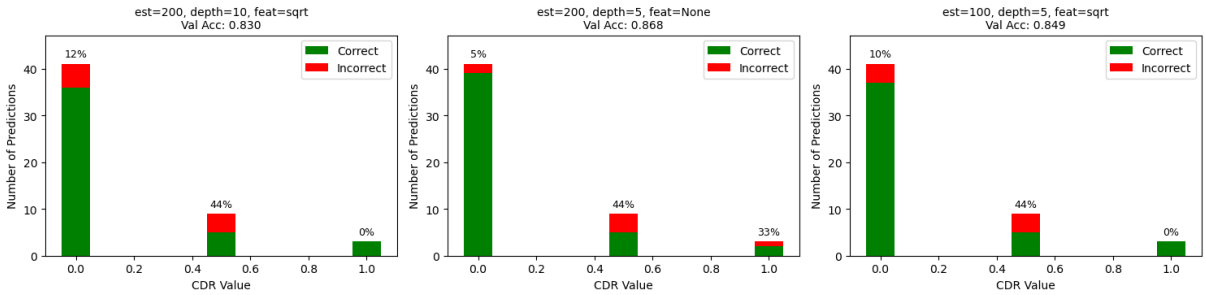

Figure 2: Prediction distribution across CDR levels for top random forest configurations (validation set)

we choose our final configuration based on validation set performance – these results are what you see in the table. This two-step process ensures that we focus on models that demonstrate distinct and strong representational capacity on the training set, while ultimately favoring models that avoid overfitting and perform reliably on unseen data.

Although our classifier is binary, fig. 2 visualizes validation set prediction performance for the top-performing trees across levels of cognitive impairment as measured by the Clinical Dementia Rating (CDR). Error rates are lowest for both CDR = 0 (no impairment) and CDR greater than or equal to 1 (clear dementia). However, performance consistently drops for borderline cases (CDR = 0.5), where all models struggle to accurately identify individuals with very mild cognitive decline.

Interestingly, while the configuration with 200 estimators and a maximum depth of 10 achieves perfect training performance, it exhibits the highest false positive rate for the majority class (CDR = 0), incorrectly flagging healthy individuals as impaired 12% of the time. Reducing the maximum depth to 5 (with the same number of estimators) lowers false positives for CDR = 0 to 5%, but at the cost of a 33% misclassification rate for CDR = 1—indicating a failure to reliably detect clearly impaired individuals. In contrast, the configuration with 100 estimators and a maximum depth of 5 strikes a more effective balance, with a false positive rate of 10% for CDR = 0 and perfect accuracy for CDR = 1. This trade-off suggests that limiting tree depth constrains model complexity, while reducing the number of features considered at each split (via square root selection) encourages simpler, more robust decision boundaries. Together, these constraints reduce the risk of overfitting to noise in the clinical data and improve generalization to unseen cases.

Given this balance between predictive accuracy and clinical relevance, we select the model with 100 estimators, a maximum depth of 5, and square root feature selection for all subsequent analyses.[3]

[3]Additional performance metrics for these configurations across training, validation, and test sets are provided in the Appendix, further supporting this model selection decision.

4.2 MRI Baseline – Finetuned VGG16

We train our CNN-based MRI classifier using a modified VGG-16 architecture (Simonyan and Zisserman, 2015) pretrained on ImageNet and fine-tuned to our dataset. The first convolutional layer is modified to accept grayscale input, a batch normalization layer is added after the first convolution, and the final fully connected layer is modified to behave like a binary classifier. This approach follows established practices in medical fine-tuning, where pretrained weights provide a strong initialization and improve generalization on small datasets (Lin et al., 2018).

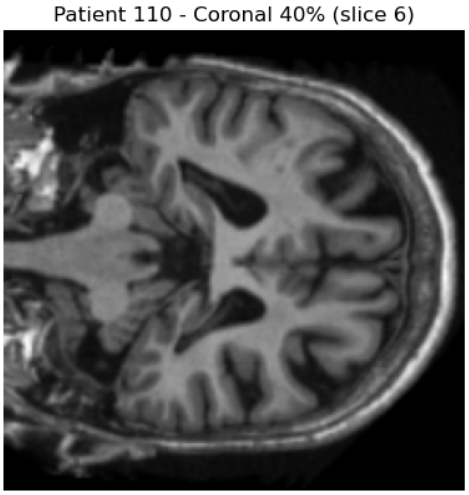

Figure 4: Example of slice 6 view

As mentioned in the data section, we extract 9 representative 2D slices from each subject's 3D T1-weighted MRI scan – three each from the axial, sagittal, and coronal planes at 40%, 50%, and 60% depth. We train nine separate CNNs, one per slice, to assess which anatomical region is most predictive of cognitive impairment – and to reduce downstream computational costs by focusing on a single 2D slice. Of these, the index six slice, corresponding to the 40% depth coronal plane view of the brain achieves the highest F1 score and accuracy. Interestingly, the coronal plane is most commonly used in clinical settings for detecting Alzheimer's.It is best for visualizing the hippocampus and medial temporal

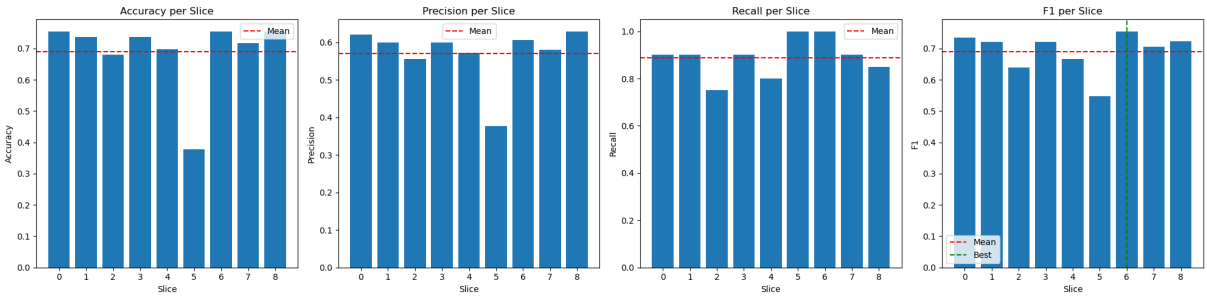

Figure 3: Validation set performance metrics across models fine tuned on a single slice (0-8)

lobe, which are the earliest regions to show atrophy in Alzheimer's (Zach et al., 2020). As a result of its performance, we select the slice 6 fine-tuned model for use in downstream fusion models. See the appendix A for a more detailed illustration of the performance differences among the 9 models.

Training is performed using stochastic gradient descent (SGD) with a learning rate of 0.01, momentum of 0.9, and a weight decay of 1×10^{-4}. To account for class imbalance in the data, we use weighted cross-entropy loss to account for class imbalance. Training proceeds for up to 20 epochs with early stopping, halting after 5 consecutive epochs without validation $F1$ improvement – the reason for this being I wanted to avoid overfitting and did not want my computer to blow up. In addition, we use a learning rate scheduler, `ReduceLROnPlateau` to reduce the learning rate by a factor of 0.1 if performance plateaus.

The model achieves a final validation F1 score of 0.65, with high recall (0.833) but slightly lower precision (0.5), suggesting a tendency to produce false positives.

To further interpret model behavior, we once again visualize classification performance across Clinical Dementia Rating (CDR) levels, see fig. 5. The model performs well across all level of cognitive impairment, achieving perfect accuracy on moderate (CDR = 1) and severe (CDR = 2) cases and low error (12.2%) on healthy individuals (CDR = 0). Notably, it also correctly classifies the majority of very mild cases (CDR = 0.5), with a misclassification rate of just 14.3%. This stands in stark contrast to the Random Forest baseline, which, while achieving 100% accuracy on healthy individuals misclassified nearly 90% of mild (CDR = 0.5) cases. These results highlight the CNN's strength in detecting early signs of cognitive impairment, a particularly challenging and clinically meaningful target.

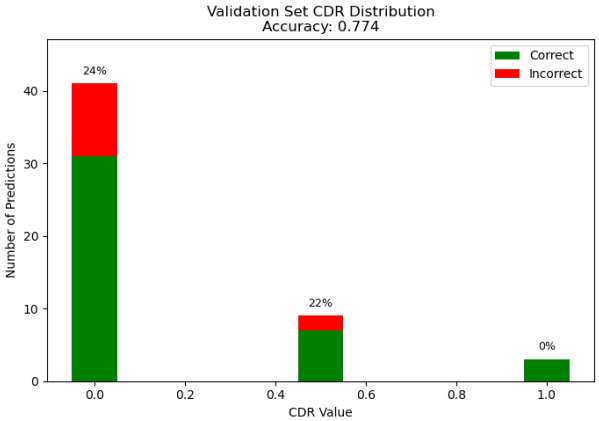

Figure 5: Prediction distribution across CDR levels for VGG-16 trained on slice 6 (validation set).

4.3 Final Performance Baseline

Using a **Random Forest with 100 estimators, a maximum depth of 5, and sqrt as the feature selection strategy**, along with a **VGG-16 model trained on slice 6** with batch normalization and weighted cross-entropy loss, we establish our final unimodal performance baselines on our test set.

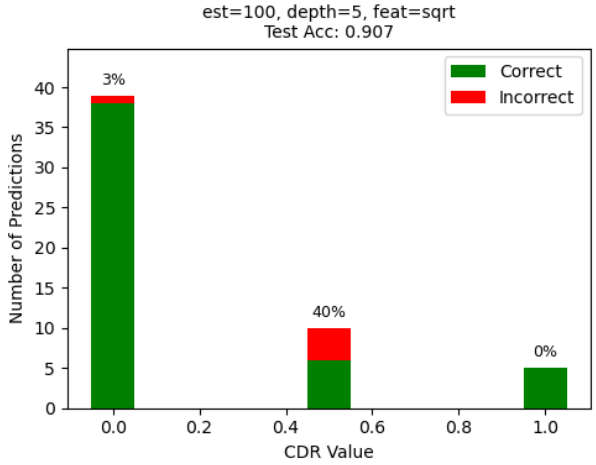

Figure 6: Prediction distribution across CDR levels for top random forest configuration (test set)

Metric	Random Forest	VGG-16 (Slice 6)
Accuracy	0.91	0.82
Precision	0.92	0.62
Recall	0.73	0.87
F1 Score	0.82	0.72

Table 2: Unimodal performance baselines for tabular and MRI models.

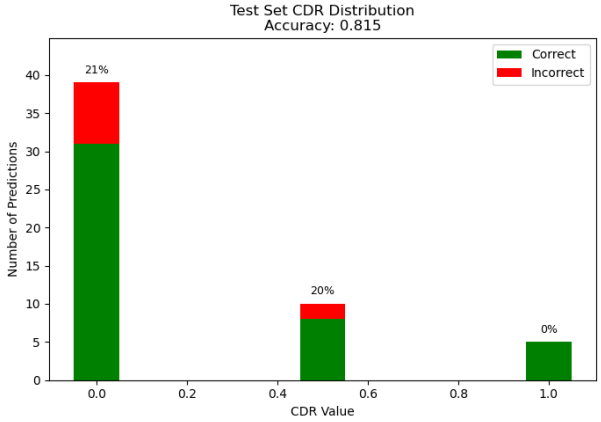

Figure 7: Prediction distribution across CDR levels for VGG-16 trained on slice 6 (test set)

5 Fusion Models

5.1 Early Fusion

In our early fusion approach, we directly combine MRI-derived features with tabular clinical data before classification. Specifically, we extract feature vectors from the CNN by taking activations from the penultimate fully connected layer and concatenate these with class probability outputs from the random forest trained on tabular data. This combined one-dimensional feature vector is then passed to a multilayer perceptron (MLP) for final classification.

We conduct a hyperparameter search over MLP architectures, dropout rates, and learning rates, using the validation set to identify the best-performing configuration. The final selected model uses two hidden layers of sizes (64,32), a dropout rate of 0.3, and a learning rate of 0.0005. For complete hyperparameter search results see the appendix.

Model	Accuracy	F1	Precision	Recall
MRI	0.778	0.647	0.579	0.733
Tabular	0.907	0.815	0.917	0.733
Early Fusion	0.852	0.667	0.889	0.533

Table 3: Unimodal vs. early fusion performance (test set)

While the early fusion model achieves competitive overall accuracy on the test set (85.2%), it exhibits a pronounced trade-off between precision and recall. Specifically, it attains one of the highest precisions among all models, but its recall is substantially lower. This means

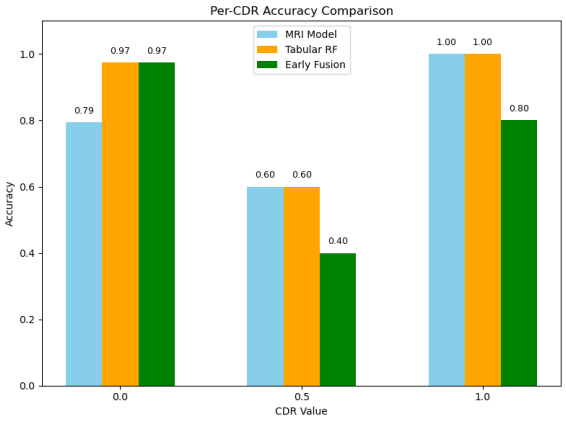

Figure 8: Per-CDR Accuracy Comparison Across Models (test set)

that it has strong confidence in its positive predictions but fails to identify true cases of cognitive impairment.

Fig. 8[4] further illustrates the discrepancy across levels of cognitive impairment. The early fusion model does a good job classifying healthy individuals (CDR = 0), matching the top unimodal baseline accuracy. However, it struggles with borderline cases (CDR = 0.5), with accuracy falling to 40% compared to 60% for both the MRI and tabular models. Although performance improves for clear dementia cases (CDR = 1), it still lags behind unimodal models.

These results suggest that, despite its theoretical appeal, the early fusion model struggles to effectively combine heterogenous feature types under limited data conditions. In particular, its inability to detect borderline cognitive impairment (CDR = 0.5) highlights the limitations of simple concatenation-based fusion. These findings motivate exploration of late fusion strategies, which integrate predictions rather than raw features and may offer more robust performance in small-data regimes this clinical context.

5.2 Late Fusion

In our late fusion approach, we combine modality-specific predictions – CNN probabilities from MRI images and random forest probabilities from tabular data – using a variety of ensemble strategies. This approach avoids directly concatenating raw feature vectors, allowing for more flexible and interpretable decision-making. We implement and compare five late fusion models: logistic regression, random forest, MLP meta-learners, simple averaging, and weighted averaging.

We first extract predicted class probabilities from both unimodal models for the training, validation, and test sets, finding that a weight of 0.95 on the tabular baseline

[4]Note that the unimodal models have the same architecture here but different splits due to the use of different data loaders. This is not great but ultimately we care about performance relative to the fusion models, and the unimodal models perform consistently across different subsets of the sample.

Model	Accuracy	F1	Precision	Recall
Unimodal Baselines				
MRI (VGG-16 Slice 6)	0.778	0.647	0.579	0.733
Tabular (Random Forest)	0.907	0.815	0.917	0.733
Early Fusion				
Early Fusion (MLP)	0.852	0.667	0.889	0.533
Late Fusion				
Logistic Regression Meta-Learner	**0.926**	**0.857**	**0.923**	**0.800**
Random Forest Meta-Learner	**0.926**	**0.857**	**0.923**	**0.800**
MLP Meta-Learner	**0.926**	**0.857**	**0.923**	**0.800**
Simple Average	0.870	0.759	0.786	0.733
Weighted Average	0.907	0.815	0.917	0.733

Table 4: Final model performance comparison (test set)

probabilities and 0.05 on the MRI baseline probabilities yields the highest validation F1 score (0.667). For the MLP meta-learner, we conduct a hyperparameter search over hidden layer sizes and learning rates using the validation set. The best configuration uses a single hidden layer of size 64 with a learning rate of 0.0005, achieving a validation F1 of 0.640. Before comparing

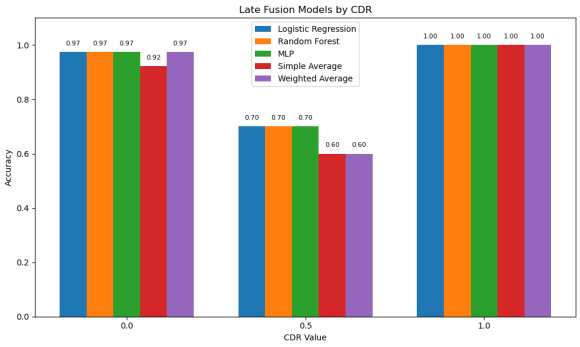

Figure 9: Per-CDR accuracy comparison across all late fusion models (test set)

fusion approaches to unimodal and early fusion models, we examine the relative performance of the late fusion methods. Table 9 presents their final performance on the test set.

Simple averaging performs the worst, unable to capture modality-specific importance, particularly in borderline cognitive impairment cases. Weighted averaging improves performance by heavily favoring the tabular modality, but it still applies a static weighting scheme.

Meta-learners (logistic regression, random forest, and MLP) all achieve identical top-line performance, with an F1 score of 0.857 and accuracy of 0.926. While their performance metrics are the same, this is not cause for concern. The models produce different class probabilities, but ultimately make the same classification decisions on the test set. Given the small and unbalanced nature of the data, there are relatively few chances for the models to differ in predicted labels, especially once both modalities are already highly predictive. The strong performance of logistic regression in particular suggests that the relationship between modality-level predictions and the final outcomes is largely linear, making it a compelling option for clinical applications given its simplicity and interpretability. When comparing across all

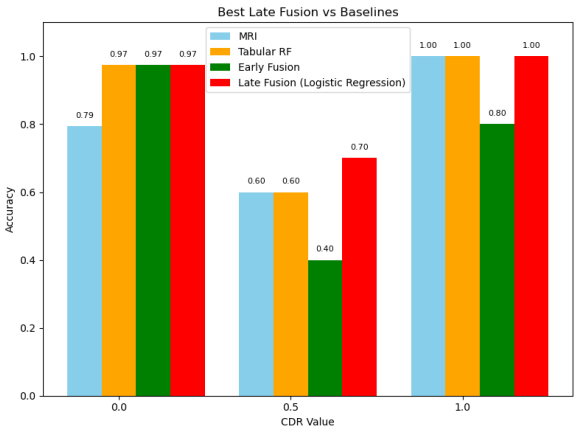

Figure 10: Per-CDR accuracy comparison of best late fusion model vs. best early fusion vs. baselines (test set)

model strategies – see fig. 10 and table 4 – late fusion offers the strongest overall performance particularly in handling borderline cognitive impairment cases (CDR = 0.5). Meta-learners in the late fusion framework outperform both unimodal baselines and early fusion, achieving an F1 score of 0.857 and accuracy of 0.926. This suggests that the decision-level integration of modality-specific predictions capture complementary information more effectively than direct feature concatenation.

Early fusion, while conceptually appealing for capturing cross-modal interactions, underperforms in practice.

Its final F1 score of 0.667 and lower recall indicate difficulty in learning meaningful joint representations from limited data, especially when combining high-dimensional CNN embeddings with tabular features.

Among unimodal baselines, the tabular random forest achieves surprisingly strong performance (F1 = 0.815), outperforming the MRI-based CNN model (F1 = 0.647). This underscores the predictive value of clinical and demographic features, which may capture early indicators of cognitive decline more directly than structural brain changes visible on MRI.

In summary, late fusion meta-learners effectively combine the complementary strengths of each modality, offering the highest predictive accuracy and robustness across cognitive impairment levels. While unimodal models provide valuable baselines and early fusion highlights the challenges of joint representation learning, late fusion emerges as the most effective and clinically actionable strategy for this task.

6 Ablation Study

To assess the contribution of specific features to model performance, we conduct an ablation study by removing one of the more informative predictors in the tabular dataset: age. See the appendix for a detailed table for the test performance of unimodal, early fusion, and late fusion models without the age feature, along with the absolute change in each metric relative to the original full feature model. Interestingly, the removal of age

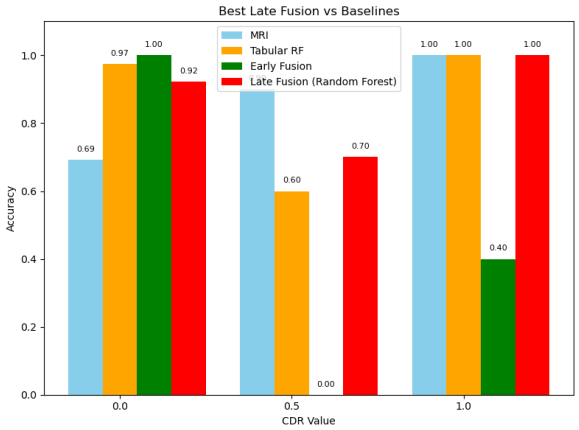

Figure 11: Per-CDR accuracy comparison of best late fusion model vs. best early fusion vs. baselines after removing age feature (test set)

had no measurable impact on the performance of the unimodal tabular model [5]. This appears surprising but likely reflects the small size of the dataset and the dominance of other highly predictive features such as MMSE and education – or possible inconsistencies introduced during reinitialization and retraining (read: I could have also messed up).

In contrast, the early fusion model suffered a dramatic collapse in F1 (-0.432) and recall ($-.400$), revealing heavy reliance on age for identifying positive cases and an inability to generalize without that cue. In particular, its ability to recognize the mild-impairment class (CDR = 0.5) goes to zero. This brittleness likely stems from the challenge of learning stable cross-modal interactions in a low data regime – especially when one modality might carry a disproportionately strong signal. By forcing MRI and tabular inputs into a single learned embedding, the early-fusion network effectively overfits to the dominant CDR 0.0 and CDR 1.0 classes once this feature is removed.

Late fusion models, on the other hand, proved to be remarkably resilient. When we look at per-CDR accuracy across the different models – fig. 14 int he appendix – we see that simple averaging of MRI and tabular probabilities achieves the highest recovery of mild impairment of 90% although it does sacrifice some accuracy on healthy controls. Both the logistic regression and MLP meta learners deliver 95% accuracy on CDR = 0, perfect detection of dementia(100% on CDR = 1), but only detects 50% of the borderline cases. The random forest combiner hits a balanced 92% on controls, 70% on CDR = 0.5, and 100% on CDR = 1.0, losing only 0.937 in accuracy and 0.057 in F1 compared to the full feature model. Weighted averaging sits in between, matching the meta learners at 92% on CDR = 0 and recovering 70% of CDR = 0.5. These results show that fusing at the probability level not only preserves extreme-class performance but also lets us dial in the trade-off between false positives on healthy subjects and true positives on mild cases.

Our ablation study underscores two key insights. First, early-fusion networks are not robust in low-data settings when a dominant feature disappears, as evidenced by their complete collapse on mild-impairment cases. Second, late-fusion strategies shield overall performance from missing features and retain the power to detect clinically important, underrepresented classes (e.g. CDR = 0.5). In practice, this suggests that diagnostic pipelines combining neuroimaging and tabular data should defer fusion until after unimodal probability estimation, then use a learned combiner to flexibly re-weigh each modality's uncertainties and ensure sensitivity across patient classes.

7 Discussion and Directions for Future Work

This project compared early and late fusion strategies for Alzheimer's detection using MRI and clinical tabular data. We found that late fusion, especially stacking with random forest or logistic regression meta-learners, consistently outperformed both unimodal baselines and early fusion. These models not only achieved the highest overall accuracy but also showed strong resilience

[5] Note that due to kernel crash, a new data loader was instantiated for this ablation study and models were retrained. As a result, unimodal baseline performance may differ from earlier reported values.

when key features like age were removed. In contrast, early fusion struggled to generalize, collapsing entirely on borderline cases (CDR = 0.5) under ablation. This highlights the brittleness of feature-level integration in low-data settings, where one modality can dominate and destabilize the learned representation.

We hypothesize that late fusion's robustness stems from its architecture: by deferring integration until after unimodal probability estimation, it preserves modality-specific signal and uncertainty. Even simple averaging recovered borderline cases better than early fusion, though learned combiners offered a stronger balance between precision and recall. These findings suggest that decision-level fusion is not only more stable but also more clinically practical, especially when dealing with heterogeneous data sources.

Future work should extend this pipeline to larger and/or longitudinal datasets like OASIS-3 or ADNI to predict progression over time. Additional improvements could involve incorporating Freesurfer-derived anatomical features in our tabular model, applying interpretability tools like Grad-CAM to the MRI network, and optimizing for computational efficiency via parallel training or model distillation. Overall, our results affirm that late fusion is a reliable and interpretable strategy for multimodal dementia classification – and a strong foundation for more complex, clinically aligned models going forward.

References

George Barnum, Sabera Talukder, and Yisong Yue. 2020. On the benefits of early fusion in multimodal representation learning. In *2nd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM)*.

Iman Beheshti, Hayrettin Demirel, and Hiroshi Matsuda. 2017. Classification of alzheimer's disease and prediction of mild cognitive impairment-to-alzheimer's conversion from structural magnetic resonance imaging using feature ranking and a genetic algorithm. *Computers in Biology and Medicine*, 83:109–119.

Rohit Bhattacharya. 2025. Introduction to neural networks. Class Lecture, Williams College.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Mark Hopkins. 2024. Convolutional neural networks. Class Lecture, CSCI 381, Williams College.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. 2020. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *npj Digital Medicine*, 3(1):136.

Weizhong Lin, Tong Tong, Qiang Gao, Danni Guo, Xiahai Du, Yanfeng Yang, Gongde Guo, and Alzheimer's Disease Neuroimaging Initiative. 2018. Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment. *Frontiers in neuroscience*, 12:777.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 689–696.

Oren Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. 2017. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.

C. Zach et al. 2020. Simplified coronal slice protocol for hippocampal atrophy. *International Journal of Alzheimer's Disease*, 2020:5894021.

A Appendix

n_estimators	max_depth	max_features	Accuracy	Precision	Recall	F1
100	5	sqrt	0.980	0.982	0.933	0.957
200	5	None	0.976	0.982	0.917	0.948
200	10	sqrt	1.000	1.000	1.000	1.000
50	15	sqrt	1.000	1.000	1.000	1.000
200	5	sqrt	0.980	0.982	0.933	0.957

Table 5: Training set metrics for top random forest configurations

n_estimators	max_depth	max_features	Accuracy	Precision	Recall	F1
100	5	sqrt	0.849	0.667	0.667	0.667
200	5	None	0.868	0.778	0.583	0.667
200	10	sqrt	0.830	0.615	0.667	0.640
50	15	sqrt	0.849	0.667	0.667	0.667
200	5	sqrt	0.849	0.667	0.667	0.667

Table 6: Validation set metrics for top random forest configurations

n_estimators	max_depth	max_features	Accuracy	Precision	Recall	F1
100	5	sqrt	0.907	0.917	0.733	0.815
200	5	None	0.907	0.917	0.733	0.815
200	10	sqrt	0.907	0.917	0.733	0.815
50	15	sqrt	0.907	0.917	0.733	0.815
200	5	sqrt	0.907	0.917	0.733	0.815

Table 7: Testing set metrics for top random forest configurations

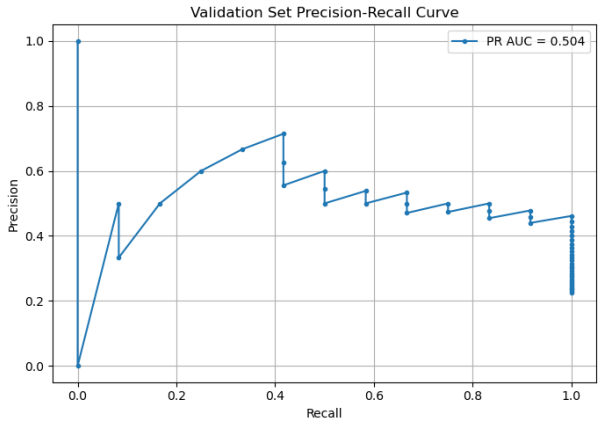

Figure 12: Precision-recall curve for VGG-16 trained on slice 6 (Validation Set)

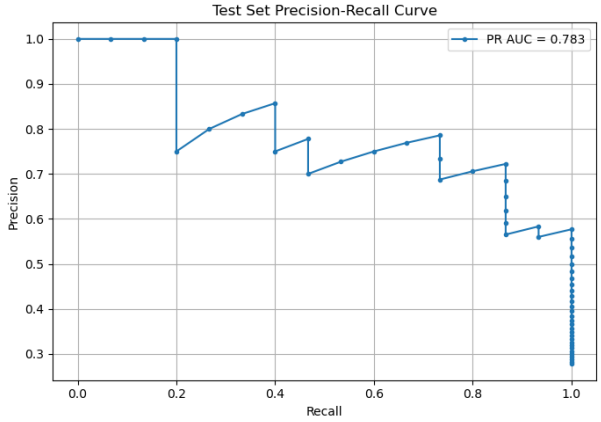

Figure 13: Precision-recall curve for VGG-16 trained on slice 6 (Test Set)

Hidden Dims	**Dropout**	**LR**	**Accuracy**	**F1**	**Precision**	**Recall**
(64, 32)	0.3	0.0005	0.811	0.583	0.583	0.583
(128, 64)	0.3	0.0005	0.811	0.583	0.583	0.583
(64,)	0.3	0.001	0.792	0.560	0.538	0.583
(64,)	0.3	0.0005	0.792	0.560	0.538	0.583
(64,)	0.5	0.001	0.792	0.560	0.538	0.583
(64,)	0.5	0.0005	0.792	0.560	0.538	0.583
(64, 32)	0.3	0.001	0.792	0.560	0.538	0.583
(64, 32)	0.5	0.001	0.792	0.560	0.538	0.583
(128, 64)	0.3	0.001	0.792	0.560	0.538	0.583
(128, 64)	0.5	0.001	0.792	0.560	0.538	0.583
(256, 128)	0.3	0.001	0.792	0.560	0.538	0.583
(256, 128)	0.3	0.0005	0.792	0.560	0.538	0.583
(256, 128)	0.5	0.001	0.792	0.560	0.538	0.583
(256, 128)	0.5	0.0005	0.792	0.560	0.538	0.583
(128, 64)	0.5	0.0005	0.792	0.353	0.600	0.250
(64, 32)	0.5	0.0005	0.792	0.154	1.000	0.083

Table 8: Early fusion hyperparameter search results (validation set)

Model	Accuracy	F1	Precision	Recall
Unimodal Baselines				
MRI (VGG-16 Slice 6)	0.759 (-0.019)	0.683 (+0.036)	0.538 (-0.041)	0.933 (+0.200)
Tabular (Random Forest)	0.907 (±0.000)	0.815 (±0.000)	0.917 (±0.000)	0.733 (±0.000)
Early Fusion				
Early Fusion (MLP)	0.759 (-0.093)	0.235 (-0.432)	1.000 (+0.111)	0.133 (-0.400)
Late Fusion				
Logistic Regression Meta-Learner	0.870 (-0.056)	0.741 (-0.116)	0.833 (-0.090)	0.667 (-0.133)
Random Forest Meta-Learner	0.889 (-0.037)	0.800 (-0.057)	0.800 (-0.123)	0.800 (±0.000)
MLP Meta-Learner	0.833 (-0.093)	0.741 (-0.116)	0.833 (-0.090)	0.667 (-0.133)
Simple Average	0.833 (-0.037)	0.757 (-0.002)	0.636 (-0.150)	0.933 (+0.200)
Weighted Average	0.889 (-0.018)	0.800 (-0.015)	0.800 (-0.117)	0.800 (+0.067)

Table 9: Test set performance after removing age feature (difference from original in parentheses)

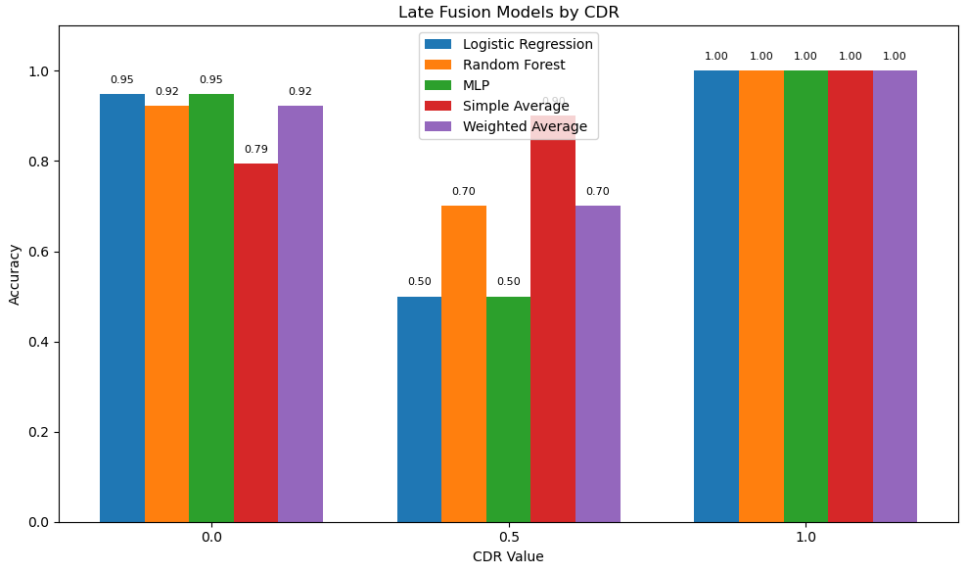

Figure 14: Per-CDR accuracy comparison across all late fusion models after removing age feature (test set)