Spring 5-8-2024

Evaluating Neuroimaging Modalities in the A/T/N Framework: Single and Combined FDG-PET and T1-Weighted MRI for Alzheimer's Diagnosis

Peiwang Liu
Washington University – McKelvey School of Engineering

WASHINGTON UNIVERSITY IN ST. LOUIS

McKelvey School of Engineering
Department of Electrical & Systems Engineering

Thesis Examination Committee:

Aristeidis Sotiras, Chair
James Feher
Abhinav Kumar Jha

Evaluating Neuroimaging Modalities in the A/T/N Framework: Single and Combined FDG-PET and T1-Weighted MRI for Alzheimer's Diagnosis

by
Peiwang Liu

A thesis presented to
the McKelvey School of Engineering
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Science

May 2024
St. Louis, Missouri

Table of Contents

List of Figures iii
List of Tables iv
Acknowledgments v
Abstract vi
Chapter 1: Introduction 1
Chapter 2: Methods 5
2.1 Datasets and Data Preprocessing 5
2.2 Experimental Setup 6
2.3 Linear Support Vector Machines 7
2.4 Training Strategy 9
2.4.1 **Nested Cross-Validation (NCV)** 9
2.4.2 Addressing the Class Imbalance 10
2.4.3 Fusion Strategies 10
2.5 Evaluation Scheme 11
2.5.1 Evaluation Matrix 11
2.6 Interpretation Methods 12
Chapter 3: Results 15
3.1 Data Summary 15
3.2 Comparative Performance Analysis of Neuroimaging Modalities and Fusion Strategies in AD Diagnosis 15
3.3 Activation maps derivation methodologies comparison results and activation maps 18
Chapter 4: Discussion 21
Chapter 5: Conclusion 23
References 24

List of Figures

Figure 1: Visual Representation of Support Vector Machine …….…….....15

Figure 2: Confusion Matrix ……...……..…18

Figure 3: ROC curve analysis of neuroimaging modalities across

diagnostic groups……...……...……..22

Figure 4: Comparison Between Haufe (2014) Covariance Adjustment Method
With other activation maps derivation methods…….25

Figure 5: Activation maps for experiments with T1w-MRI and FDG-PET modalities ……...26

List of Tables

Table 1: Demographics and clinical characteristics of study participants……21

Table 2: Comparative diagnostic performance metrics across imaging modalities and fusion techniques……..24

Acknowledgments

I would like to extend my deepest gratitude to Professor Aristeidis Sotiras for his invaluable guidance, mentorship, and support throughout the course of my study. His expertise and insightful advice have been fundamental in shaping the direction and success of my research.

I am also immensely grateful to my lab members Tom Earnest, Braden Yang, Pan Xiao and John Lee at the Medical Imaging and Data Science (MINDS) Lab, Washington University School of Medicine in St. Louis. Their collaboration, encouragement, and shared knowledge have been crucial to my work and personal growth during this time.

Special thanks are owed to the Washington University School of Engineering for their generosity in allowing us to utilize their thesis template. This resource has been instrumental in the development of this document.

Lastly, I appreciate the constant support and patience of my friends and family, who have provided me with the motivation and strength to pursue my goals.

Peiwang Liu

Washington University in St. Louis

May 2024

ABSTRACT OF THE THESIS

Evaluating Neuroimaging Modalities in the A/T/N Framework:

Single and Combined FDG-PET and T1-Weighted MRI for Alzheimer’s Diagnosis

by

Peiwang Liu

Master of Science in Engineering Data Analytics and Statistics

Washington University in St. Louis, 2024

Professor Aristeidis Sotiras, Chair

With the escalating prevalence of dementia, particularly Alzheimer's Disease (AD), the need for early and precise diagnostic techniques is rising. This study delves into the comparative efficacy of Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) and T1-weighted Magnetic Resonance Imaging (MRI) in diagnosing AD, where the integration of multimodal models is becoming a trend. Leveraging data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), we employed linear Support Vector Machines (SVM) to assess the diagnostic potential of these modalities, both individually and in combination, within the AD continuum. Our analysis, under the A/T/N framework's 'N' category, reveals that FDG-PET consistently outperforms T1w-MRI across various stages of cognitive impairment. Contrary to expectations and previous studies that suggested enhanced diagnostic accuracy through the fusion of neuroimaging modalities—including CSF markers—our findings do not demonstrate a significant improvement in diagnostic performance from combining FDG-PET and MRI data. This outcome aligns with Narazani et al. (2022), challenging the prevailing assumption about the added value of multimodal data fusion in AD diagnosis. Through the interpretation of activation maps, our study further elucidates the distinct yet complementary roles of FDG-PET and MRI in highlighting the pathological underpinnings of AD, contributing to a nuanced understanding of

neuroimaging biomarkers in clinical settings. Our research underscores the critical need for refined strategies in neuroimaging data integration, advocating for a more discerning application of single and multimodal approaches in the early detection of AD.

Chapter 1: Introduction

Dementia commonly denotes a reduction in cognitive functions, encompassing memory, cognition, and linguistic abilities, to an extent where it significantly impacts daily functioning and may leave the patient completely disabled (Bhushan et al., 2018). It is approximated that currently more than 46.8 million individuals globally are experiencing dementia, and this figure is forecasted to rise to 74.7 million by 2030. Simultaneously, the expenses related to dementia care are anticipated to escalate from US$818 billion to US$2 trillion (APA, 2013). Since Alois Alzheimer pinpointed an aggressive type of dementia, now recognized as Alzheimer's disease, in the previous century, it has been acknowledged as the most common form of dementia in individuals above the age of 65 (Bhushan et al., 2018). Although no specific medications have been formulated to directly address the disease, early detection and the commencement of care management are vital (Borson et al., 2013). It is also imperative to distinguish Alzheimer's disease from other analogous conditions such as depression and delirium.

Mounting proof proposes that the Alzheimer’s disease does not advance through distinct phases; instead, it is more comparable to a continuum, characterized as a smooth progression in which neighboring elements are not noticeably distinct from each other, with physiological changes occurring many years prior to clinical diagnosis (Aisen et al., 2017). Up to the present time, no direct causative and transitional mechanisms have been pinpointed between healthy individuals and patients; however, numerous risk factors have been correlated with the disease, encompassing age, genetics, and educational levels. Ongoing research has identified several histopathological features of Alzheimer's disease: the buildup of amyloid plaques, which are extracellular accumulations of Aβ protein; the development of neurofibrillary tangles (NFTs),

which are intraneuronal clusters of tau protein; and neurodegeneration, which is the gradual decline of neurons or their processes (Aisen et al., 2017). Notably, Aβ protein aggregation can precede any clinical indications of cognitive abnormalities and there is no distinct boundary between phases, as formerly defined in clinical contexts, but instead a continuous advancement as previously mentioned (Jack et al., 2010). The onset of the disease could be generally identified when the accumulation of Aβ protein becomes apparent, which can be detected through positron emission tomography (PET) scans or cerebrospinal fluid (CSF) analysis. There are seven principal AD biomarkers, which, according to the A/T/N system, are categorized into three binary groups reflecting the nature of the pathophysiology they measure: 'A' represents the value of an amyloid biomarker such as amyloid PET or CSF Aβ_42; 'T' is the value of a tau biomarker like CSF phosphorylated tau or tau PET; and 'N' signifies biomarkers of neurodegeneration or neuronal injury, which include [18F]-fluorodeoxyglucose–PET (FDG-PET), structural MRI, or CSF total tau (Jack et al., 2016; Aisen et al., 2017). To diagnose AD as early as possible, recent research has focused on identifying patients with cognitive impairment who have not yet progressed to severe dementia—such as those with mild cognitive impairment (MCI), a well-defined clinical syndrome that carries a higher risk of progressing to AD dementia. Notably, a proportion of MCI patients remain stable for years or may even revert to normal cognitive functioning. (Grueso & Viejo-Sobera, 2021; Sperling et al., 2011; Ewers et al., 2011; Jack et al., 2010).

FDG-PET is used to assess a decline in synaptic function, and structural volumetric magnetic resonance imaging (MRI) measures changes in the brain's gray matter. Hence, FDG-PET and MRI have become popular biomarkers in 'N' group from A/T/N system in some research (Frisoni et al., 2013; Ewers et al., 2011; Narazani et al., 2022; Samper-González et al., 2018;

Zhang et al., 2011; Song et al., 2021; Grueso & Viejo-Sobera, 2021; Arya et al., 2023; Zhao et al., 2023; Zhou et al., 2019; Hinrichs et al., 2011; Chincarini et al., 2011; Mosconi et al., 2010; Kim et al., 2021; Ding et al., 2019). Some studies focus on a single modality, for instance, Kim et al. (2021) constructed a 2.5-D deep learning architecture using 291 submodules and three-axes FDG-PET images to classify amyloid PET positivity and negativity. Chincarini et al. (2011) sampled the brain at seven relatively small volumes centered on the medial temporal lobe (MTL) and two control regions, using intensity and textural MRI-based features, which were extracted using a Random Forest (RF) algorithm and then processed with a Support Vector Machine (SVM) classifier to predict AD conversion in MCI patients. Others have advocated combining multiple imaging modalities. Zhang et al. (2011) utilized a multiple-kernel SVM to combine three modalities of biomarkers (FDG-PET, MRI, and CSF) to differentiate between AD (or MCI) and healthy controls. Zhou et al. (2019) addressed the challenges of heterogeneous and incomplete multimodality data by constructing a stage-wise deep learning strategy that allowed the model to use the maximum amount of available data from each modality. Specifically, they learned latent representations for each modality in the first stage and then combined the higher-level features from each modality in subsequent stages. In the final stage, they fused the learned joint latent features from the previous stage to learn the diagnostic labels. Both studies found that the performance of combined modalities surpassed that of single modalities (Zhou et al., 2019; Zhang et al., 2011). However, since the volume of brain structures typically decreases with age, it can be challenging to determine whether a person's brain changes observed by MRI are within normal aging or indicative of disease, thus PET often provides greater discriminatory power, as supported by most studies (Song et al., 2021; Zhang et al., 2011; Arya et al., 2023; Grueso & Viejo-Sobera, 2021). This raises the question of why combining MRI and FDG-PET would be

beneficial. According to Narazani et al. (2022), no improvement was observed when combining FDG-PET and MRI imaging modalities with 3D ResNet.

In this study, given that Narazani et al.'s (2022) experiments focused exclusively on deep learning architectures, while prior research often utilized SVM and various other machine learning techniques, we revisited the efficacy of using single modalities (MRI or FDG-PET) versus their combination through two distinct fusion methods with a linear SVM. Our findings, aligning with previous research, highlighted FDG-PET's significant superiority over T1w-MRI. Furthermore, we observed no significant enhancement in model performance when combining FDG-PET and T1w-MRI as inputs.

Chapter 2: Methods

2.1 Datasets and Data Preprocessing

Data for this study were sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, accessible at https://adni.loni.usc.edu. The ADNI dataset was lunched in 2003 by the National Institute on Aging, the National Institute of Biomedical imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organizations with a 5-year public private partnership. With the purpose of exploring the potential of multimodality data, including neuroimaging, clinical, biological, and genetic biomarkers to diagnose AD and its early status. Using the advanced search utility, we procured a total of 1012 Fluorodeoxyglucose (FDG) positron emission tomography (PET) scans, collected between September 22, 2005, and January 4, 2022. These FDG scans, characterized by the descriptor 'Coreg, Avg, Std Img and Vox Siz, Uniform Resolution', were acquired in Neuroimaging Informatics Technology Initiative (NIfTI) format. For each participant with FDG data, we employed the same search methodology to obtain the most contemporaneous T1-weighted (T1w) magnetic resonance imaging (MRI) data, with a maximum allowable interval of 365 days between the FDG and T1w scans. This approach yielded an equivalent dataset of 1012 T1w scans (Table 1) . Initially in Digital Imaging and Communications in Medicine (DICOM) format, these T1w images were converted to NIfTI format using the dcm2niix tool. We then standardized the orientation of all NIfTI images to right-posterior-inferior using AFNI's 3d_resample tool.

The subsequent image processing workflow involved several key steps. First, we performed bias field correction in the T1w images using Advanced Normalization Tools (ANTs) N4BiasFieldCorrection. This was followed by the generation of nonlinear warps using ANTs'

antsRegistrationSyNQuick and antsApplyTransforms tools, which allowed for the rigid-body co-registration of FDG to T1w images using 4dfp t4_resolve. Subsequent normalization of the FDG standardized uptake value ratio (SUVR) was performed relative to the pons and cerebellar vermis regions, as delineated by FreeSurfer. This comprehensive process of registration, warping, and the application of inverse transformations enabled the accurate alignment of FDG data to the MNI152 atlas. For high-resolution warping, atlas-registered binary masks were applied to exclude FDG voxels outside the brain.

In the final processing stage for T1w data, we utilized ANTs to calculate the Jacobian determinants and performed intracranial volume (ICV) correction on all T1w data by normalizing the maps with estimated ICV values derived from the T1w data. To optimize classification, normalized values were scaled to a 0-1 range to mitigate the impact of excessively small magnitudes. After completing the registration and normalization processes, we applied grey matter masks to all datasets. These masks were generated using FSL's FAST (FMRIB's Automated Segmentation Tool) from segmented grey matter in the T1w images. By applying these masks to both the FDG PET and T1w datasets, we ensured that subsequent analyses were specifically focused on grey matter regions.

2.2 Experimental Setup

In this research, we conducted a series of binary classification experiments to assess the diagnostic capabilities of two neuroimaging modalities: [18F]-fluorodeoxyglucose (FDG) positron emission tomography (PET) and T1-weighted magnetic resonance imaging (MRI) in Alzheimer's Disease (AD).

Our experimental design is centered around four pairs of group comparisons, with criteria clinical dementia rating (CDR) and Amyloid status:

Comparative analyses were carried out between the CDR=0, Amyloid-negative group (Cognitively unimpaired) and each of the other three groups: CDR=0, Amyloid-positive (Preclinical); CDR=0.5, Amyloid-positive (MCI); and CDR>0.5, Amyloid-positive (AD Dementia).

Additionally, a specific binary classification was conducted between the CDR>0.5, Amyloid-positive (AD Dementia) group and the CDR=0.5, Amyloid-positive group (MCI).

For each of these four group pairs, we conducted four distinct experiments for analysis: two baseline experiments using each modality independently, and two fusion experiments with combined modalities —early fusion and late fusion, as detailed later. This quadripartite experiments for each group pair, involving both FDG-PET and T1w-MRI data, leads to a comprehensive set of sixteen experiments.

2.3 Linear Support Vector Machines

Support Vector Machines (SVM) were used in all sixteen experiments, it classifies data into distinct groups (e.g., baseline/control), by identifying a hyperplane that maximizes the margin between these groups. In the context of Linear SVM, the chosen hyperplane is explicitly linear, ensuring a straightforward and consistent separation between the two classes in the high-dimensional space. Once trained, the SVM model employs this hyperplane as a decision boundary: data points on one side would be classified into one group, while those on the opposite side would be categorized into the other.

In the context of neuroimaging data, an image with D voxels is transformed into a vector whose d^{th} component corresponds to the intensity value at the d^{th} voxel in the image.

The Linear SVM optimization problem can be mathematically represented as follows:

$$\{w^*, b^*\} = \arg\min_{w,b,\xi} \frac{1}{2}|w|^2 + C\sum_{i=1}^{m} \xi_i \tag{1}$$

subj. to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, m$$

$$\xi_i \geq 0 \quad \forall i = 1, \ldots, m$$

Where w is the weight vector defining the orientation of the hyperplane, b is the bias term that determines the position of the hyperplane, C is the regularization parameter controlling the trade-off between maximizing the margin and minimizing classification error, ξ_i are slack variables allowing for misclassification or representing the distance of correctly classified points from the margin boundary for each data point x_i, y_i is the label of the data point x_i, either +1 or -1, indicating its class.

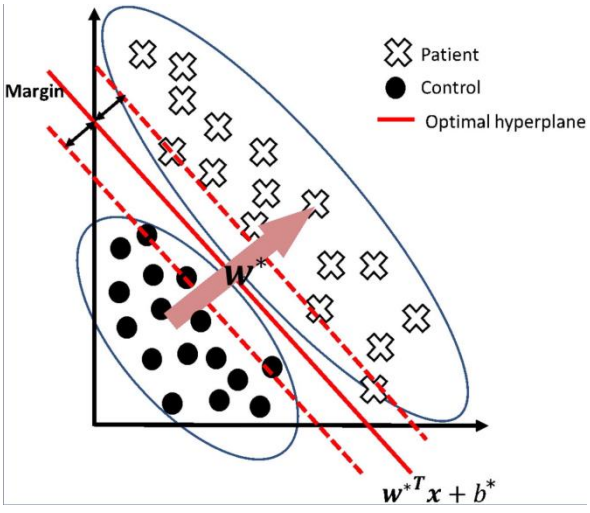

Figure 1: Visual representation of Support Vector Machine

The weight vector w^* and bias b^* together determine the optimal hyperplane in the format $w^*x + b^*$. The position and orientation of this hyperplane in the high-dimensional space is influenced by the neuroimaging data. For any new data point x_{new}, its classification is based on the sign of $w^*x_{new} + b^*$. If the result is positive, the data point belongs to one class, and if it's negative, it belongs to the other class.

2.4 Training Strategy

2.4.1 Nested Cross-Validation (NCV)

The integral part of our methods was the Nested Cross-Validation (NCV). This structure served a dual purpose: to better estimate the model's performance while simultaneously optimizing the hyperparameters.

Essentially, the Nested Cross-Validation (NCV) operates in a layered manner. The entire dataset is first divided by the outer loop, where one portion is designated for training the model and another for testing its generalization. Within each fold of this outer division, the inner cross-validation further splits the training data for hyperparameter tuning. Its primary task is to determine the hyperparameters could generate the best performance for the model. By running multiple iterations on these subsets of the training data, it identifies which hyperparameter configurations yield the best results according to the predefined evaluation metric. Upon determining these optimal settings in the inner loop, they are applied to train the model on the full dataset portion provided by the current fold of the outer loop to estimate the model's performance. This layered approach ensures robust evaluation and optimization of the model while preventing overfitting to the training data.

Within this study's NCV framework, the outer loop, configured with 5 folds, primarily evaluated the model's generalization capabilities on the dataset. The inner loop, consisting of 3 folds, was

tailored to fine-tune hyperparameter C as presented in Eq. (1). We utilized the Grid Search method over the search space [1, 0.1, 0.01, 0.001, 0.0001], with the Area Under the Receiver Operating Characteristic Curve (AUC) as the evaluation metric.

2.4.2 Addressing the Class Imbalance

There is class imbalance exist in the dataset which frequently pose challenges in machine learning tasks, potentially skewing model performance. To decrease the influence of this imbalance, we employed Stratified K-Fold Cross-Validation for both inner and outer cross-validation stages. Additionally, we modified SVM model setting to make sure the SVM was trained with class weights that inversely reflect the class frequencies. This approach ensures that the model treats each class instance with equal importance, thereby mitigating biases that could favor predictions towards the majority class.

2.4.3 Fusion Strategies

We conducted two sets of experiments employing distinct fusion strategies to validate our findings further. For early fusion, MRI and FDG data were combined by concatenating them before being passed to the Linear SVM, harnessing the Sparse Random Projection technique to ensure consistent dimensionalities with non-fusion cases. In the context of late fusion, our approach involves employing an optimized weighting scheme derived from the best Area Under the Curve (AUC) values obtained from individual T1w-MRI and FDG-PET models. Specifically, we initiate the process by training two Support Vector Machine (SVM) models on distinct models originating from FDG-PET and MRI modalities. Subsequently, when presented with a new testing sample, each model generates a prediction for it. Finally, we aggregate all predictions based on selected weight to reach a consensus decision on the classification of the new testing sample.

2.5 Evaluation Scheme

We assessed the model's efficacy utilizing metrics such as accuracy, PPV, NPV, sensitivity, specificity, F1 score, and Receiver Operating Characteristics (ROC) and Area Under the ROC Curve (AUC). The determination of 95% CIs was facilitated through the Bootstrap method. This method resampled results from each cross-validation split with replacement 1,000 times to simulate different scenarios. For each resampled dataset, our performance metric was calculated using the predictions and corresponding ground truth values. The confidence interval was then derived from the 2.5th and 97.5th percentiles of the resulting score distribution, providing an estimate of the metric's variability without assuming a specific underlying distribution. The DeLong test (1988) was enlisted for comparative statistical evaluations.

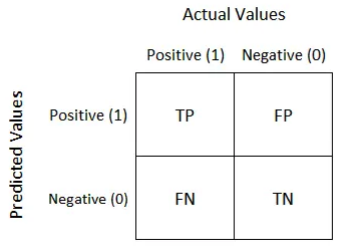

Figure 2: Confusion Matrix

2.5.1 Evaluation Matrix

Accuracy calculated as (TP+TN)/(TP+TN+FN+FP) (Fig. 2), represents the proportion of true results (both true positives and true negatives) among the total number of cases examined. It gives a quick snapshot of the model's overall performance but may not always reflect the effectiveness of the model in datasets with imbalanced classes.

PPV calculated as TP/(TP+FP) (Fig. 2), also known as precision. PPV is the ratio of true positive results to all positive results reported by the model, indicating the likelihood that a positive classification is correct.

NPV calculated as TN/(TN+FN) (Fig. 2). Similar to PPV, NPV is the ratio of true negative results to all negative results reported by the model, reflecting the probability that a negative classification is accurate.

Sensitivity calculated as TP/(TP+FN) (Fig. 2), also known as recall, measures the proportion of actual positives correctly identified by the model, highlighting its ability to detect positive cases.

Specificity calculated as TN/(TN+FP) (Fig. 2), assesses the proportion of actual negatives that are correctly identified, indicating the model's efficacy in recognizing negative cases.

F-1 score calculated as 2*(PPV*Sensitivity)/(PPV+Sensitivity), is the harmonic mean of precision (PPV) and sensitivity (recall), providing a single metric that balances both values, particularly useful for imbalanced datasets.

The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The AUC represents the degree to which the model is capable of distinguishing between classes, with an AUC of 1 indicating perfect prediction and an AUC of 0.5 suggesting no discriminative power.

2.6 Interpretation Methods

In the interpretation of our SVM models, the weight vector w, may not accurately represent the neural sources of data. To obtain meaningful patterns, we compute the activation patterns A using the formula:

$$A = Cov[x(n), \hat{s}(n)] \tag{2}$$

In this equation, $x(n)$ denotes the preprocessed neuroimaging data vector, and $\hat{s}(n)$ is the predicted label vector corresponding to the data. The covariance between the data and the labels, as described by Haufe et al. (2014), provides a correction to the Support Vector Machine weights, yielding activation patterns A that are more representative of the underlying neural activity.

To validate the significance of these activation patterns, we implemented permutation testing. This involves systematically shuffling the labels and recalculating the activation patterns for 1000 iterations. Each permuted set's activation pattern is then compared to the original, unshuffled activation pattern. The comparison yields p-values for each feature, reflecting the likelihood that the observed activation is due to chance. These p-values inform the statistical significance of the weights, enhancing the robustness of the activation maps derived from the model.

To underscore the efficacy of our approach in enhancing SVM weight interpretability, we conducted a comparative analysis through three additional experiments, each tailored to derive activation maps via distinct methodologies:

(1) Direct application of unaltered SVM weights, assessing their intrinsic representational capacity

(2) Augmentation of raw SVM weights with the mentioned permutation testing

(3) Derivation of activation patterns using the methodology proposed above by Haufe et al. (2014), excluding the permutation testing phase.

To facilitate a standardized comparison across these methods, we adjusted the resulting metrics to conform to a 'smaller is better' paradigm. Hence, p-values yielded by permutation testing were directly interpreted, whereas outcomes obtained without permutation were converted via subtraction from unity, ensuring a uniform metric indicating the reliability of the activation patterns across all experimental conditions.

Chapter 3: Results

3.1 Data Summary

The dataset used in this study includes 1,012 participants segmented by CDR scores and amyloid status. There is data imbalance exists, with the CDR = 0.5, Amyloid + group having the highest number of subjects (n = 458), while the CDR > 0.5, Amyloid + group is the smallest (n = 151). Regarding age, the variation across groups is minimal, with all groups having a mean age within a close range of approximately 73 to 75 years. The gender distribution exhibits a discrepancy; the CDR = 0, Amyloid + group has the highest percentage of females (62.9%), in contrast to the CDR = 0.5, Amyloid + group, which has the lowest (44.3%) (Table 1).

Group	CDR = 0, Amyloid -	CDR = 0, Amyloid +	CDR = 0.5, Amyloid +	CDR > 0.5, Amyloid +	Total
Subject Number	263	140	458	151	1012
Age	73.4±6.5	75.9±5.9	74.2±7.3	75.4±8.0	74.4±7.1
Female	47.5%	62.9%	44.3%	51.0%	48.7%

Table 1: Demographics and clinical characteristics of study participants

3.2 Comparative Performance Analysis of Neuroimaging Modalities and Fusion Strategies in AD Diagnosis

In the comparative analysis across different diagnostic groups, the result indicated that the contrast between CDR = 0, Amyloid-negative (Cognitively Unimpaired) and CDR > 0.5, Amyloid-positive (AD Dementia) groups yielded the most robust diagnostic outcomes, with AUC values ranging from 0.967 to 0.992. Conversely, the comparison between CDR = 0, Amyloid-negative (Cognitively Unimpaired) and CDR = 0, Amyloid-positive (Preclinical) groups produced the least discriminative results, with AUC values ranging from 0.549 to 0.628 (Table 2 and Figure 3). These findings align with our hypothesis that greater clinical stage differences would lead to more distinct diagnostic results.

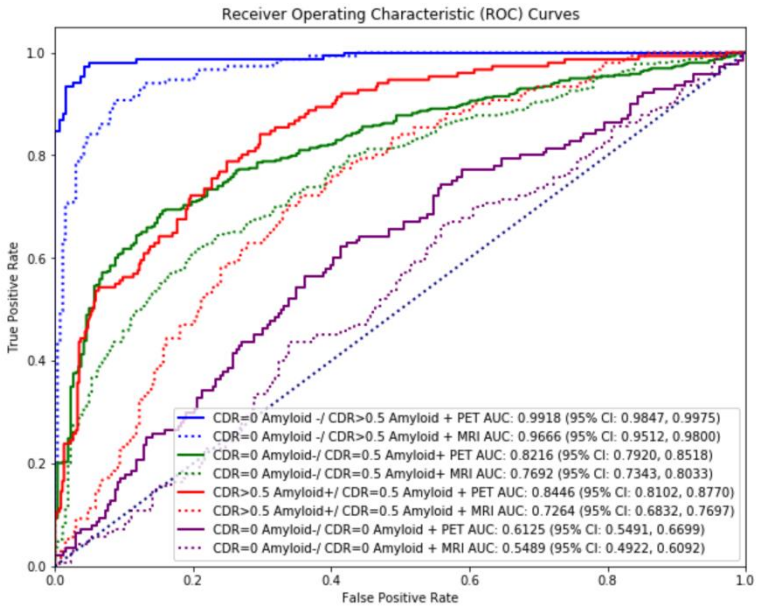

Figure 3: ROC curve analysis of neuroimaging modalities across different diagnostic groups

Analyzing the performance within each comparative pair, FDG-PET consistently outperformed T1w-MRI in terms of AUC, specifically, p-values correlated with the AUC comparisons were $P<0.001$ for the first pair, $P=0.003$ for the second, $P<0.001$ for the third, and $P=0.1$ for the fourth, corresponding sequentially with the experiments listed in Table 2 from top to bottom. With regards to fusion techniques, early fusion generally resulted in intermediate AUC performance between FDG-PET and T1w-MRI modalities, except the group CDR = 0.5, Amyloid-positive (MCI) versus CDR > 0.5 (AD Dementia), suggesting that mere combination and dimensionality reduction of multimodal data may not necessarily yield improved diagnostic performance.

Late fusion, on the other hand, either maintained or slightly improved AUC performance (0.992/0.992, 0.833/0.822, 0.845/0.845, 0.628/0.612; Table 2). However, statistical analysis revealed no significant differences between the models applied in most experiments (P=0.86, 0.03, 0.29, 0.11). Within each late fusion model, FDG-PET input model had higher weights than

T1-weighted MRI input model across the board: 0.664 versus 0.336, 0.64 versus 0.36, 0.86 versus 0.14, and 0.568 versus 0.432, respectively, corresponding sequentially with the experiments listed in Table 2 from top to bottom. It must also be considered that the late fusion technique inherently always optimizes for the best AUC result by combining individual modality outcomes, which could introduce a bias in favor of those models demonstrating superior individual performance.

Metric	FDG-PET	T1w-MRI	Early Fusion	Late Fusion
AUC	0.992 (0.985, 0.997)	0.967 (0.951,0.980)	0.989 (0.980, 0.996)	0.992 (0.983, 0.998)
F1 score	0.972 (0.958, 0.985) / 0.949 (0.920, 0.972)	0.922 (0.898,0.944) / 0.844 (0.800,0.889)	0.963 (0.945, 0.978) / 0.932 (0.899, 0.961)	0.972 (0.957, 0.985) / 0.949 (0.922, 0.972)
NPV	0.966 (0.934, 0.993)	0.935 (0.886,0.975)	0.958 (0.922, 0.987)	0.979 (0.952, 1.000)
PPV	0.963 (0.939, 0.985)	0.879 (0.840,0.915)	0.948 (0.919, 0.971)	0.956 (0.928, 0.978)
Sensitivity	0.934 (0.890, 0.969)	0.768 (0.701,0.833)	0.907 (0.863, 0.953)	0.921 (0.874, 0.957)
Specificity	0.981 (0.965, 0.996)	0.970 (0.947,0.989)	0.977 (0.958, 0.996)	0.989 (0.974, 1.000)
CDR = 0, Amyloid - / CDR > 0.5 Amyloid +				
AUC	0.822 (0.792, 0.852)	0.769 (0.734, 0.803)	0.811 (0.778, 0.839)	0.833 (0.802, 0.861)
F1 score	0.678 (0.632, 0.722) / 0.789 (0.758, 0.817)	0.527 (0.469, 0.579) / 0.787 (0.757, 0.814)	0.649 (0.605, 0.692) / 0.774 (0.745, 0.807)	0.626 (0.573, 0.676) / 0.819 (0.790, 0.844)
NPV	0.833 (0.797, 0.866)	0.729 (0.691, 0.766)	0.810 (0.773, 0.844)	0.774 (0.737, 0.811)
PPV	0.628 (0.570, 0.681)	0.638 (0.566, 0.707)	0.608 (0.551, 0.662)	0.710 (0.649, 0.769)
Sensitivity	0.749 (0.709, 0.788)	0.854 (0.822, 0.885)	0.742 (0.705, 0.783)	0.869 (0.837, 0.898)
Specificity	0.738 (0.681, 0.786)	0.449 (0.385, 0.511)	0.696 (0.641, 0.752)	0.559 (0.496, 0.617)
CDR = 0, Amyloid - / CDR = 0.5 Amyloid +				
AUC	0.845 (0.810, 0.877)	0.726 (0.683, 0.770)	0.855 (0.821, 0.887)	0.845 (0.810, 0.877)
F1 score	0.847 (0.819, 0.872) / 0.599 (0.534, 0.654)	0.834 (0.808, 0.858) / 0.353 (0.275, 0.425)	0.853 (0.827, 0.876) / 0.624 (0.561, 0.684)	0.890 (0.868, 0.910) / 0.538 (0.458, 0.609)
NPV	0.543 (0.468, 0.614)	0.449 (0.352, 0.549)	0.557 (0.489, 0.629)	0.759 (0.667, 0.850)
PPV	0.882 (0.847, 0.913)	0.791 (0.755, 0.824)	0.894 (0.864, 0.925)	0.833 (0.801, 0.866)
Sensitivity	0.669 (0.597, 0.745)	0.291 (0.216, 0.365)	0.709 (0.634, 0.778)	0.417 (0.342, 0.496)
Specificity	0.814 (0.780, 0.848)	0.882 (0.850, 0.910)	0.814 (0.779, 0.850)	0.956 (0.936, 0.974)
CDR = 0.5, Amyloid + / CDR > 0.5 Amyloid +				
AUC	0.612 (0.549, 0.670)	0.549 (0.492, 0.609)	0.602 (0.543, 0.660)	0.628 (0.567, 0.683)
F1 score	0.707 (0.659, 0.748) / 0.440 (0.367, 0.513)	0.757 (0.719, 0.793) / 0.184 (0.110, 0.259)	0.706 (0.659, 0.746) / 0.454 (0.387, 0.523)	0.791 (0.756, 0.823) / 0.081 (0.026, 0.141)
NPV	0.445 (0.355, 0.522)	0.378 (0.234, 0.521)	0.451 (0.366, 0.529)	0.667 (0.333, 1.000)
PPV	0.703 (0.645, 0.760)	0.656 (0.606, 0.707)	0.709 (0.650, 0.763)	0.660 (0.612, 0.706)
Sensitivity	0.436 (0.355, 0.518)	0.121 (0.073, 0.179)	0.457 (0.370, 0.545)	0.043 (0.014, 0.082)
Specificity	0.711 (0.656, 0.762)	0.894 (0.854, 0.927)	0.703 (0.648, 0.758)	0.989 (0.974, 1.000)
CDR = 0, Amyloid - / CDR = 0 Amyloid +				

Table 2: Comparative diagnostic performance metrics across imaging modalities and fusion techniques. In each experiment, the group with less severe dementia (indicated by a lower

Clinical Dementia Rating, or CDR) was designated as the negative class. In cases where the CDR values were equivalent, the group without amyloid pathology was assigned as the negative class. F1 score would show in sequence negative class/positive class

3.3 Activation maps derivation methodologies comparison results and activation maps

With the comparison of four Activation maps derivation methodologies (Figure 4), the direct application of unmodified SVM weights and the activation patterns derived via Haufe et al.'s (2014) methodology without permutation testing both yielded few results on the activation maps. This underscores the critical role of permutation procedures in enhancing interpretability. Upon integrating raw SVM weights with permutation testing, the resultant activation maps exhibited a proliferation of noise points in contrast to the refined patterns obtained through the Haufe et al. (2014) methodology coupled with permutation testing. Consequently, these results unequivocally demonstrate that the application of the Haufe er al.’s (2014) transformation in conjunction with permutation testing confers superior performance in discerning meaningful and concise activation patterns.

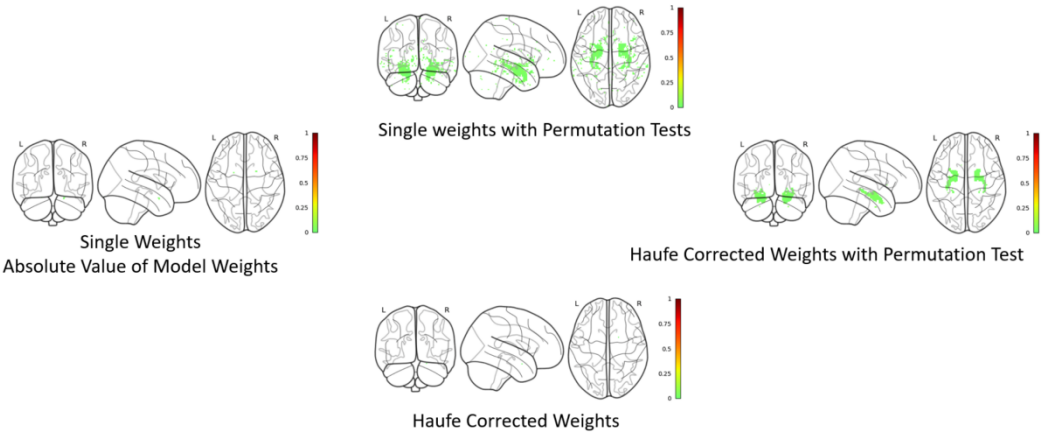

Figure 4: Comparison Between Haufe (2014) Covariance Adjustment Method with other activation maps derivation methods. The results are derived from experiments with the CDR = 0, amyloid- and CDR > 0.5, amyloid+ groups, using T1-weighted MRI as the input modality.

The activation maps from above experiments' SVM models (Figure 5) provide a visual representation of the underlying brain changes. T1w-MRI captures structural brain changes, notably hippocampal atrophy, which correlate with the memory deficits observed in early AD stages. FDG-PET, in contrast, detects functional changes such as decreased glucose metabolism in the temporal regions, often preceding the structural alterations visible on MRI. Together, these weights underscore and are consistent with the prior knowledge that T1-weighted MRI and FDG-PET modalities complement each other in providing a comprehensive understanding of Alzheimer's Disease pathology.

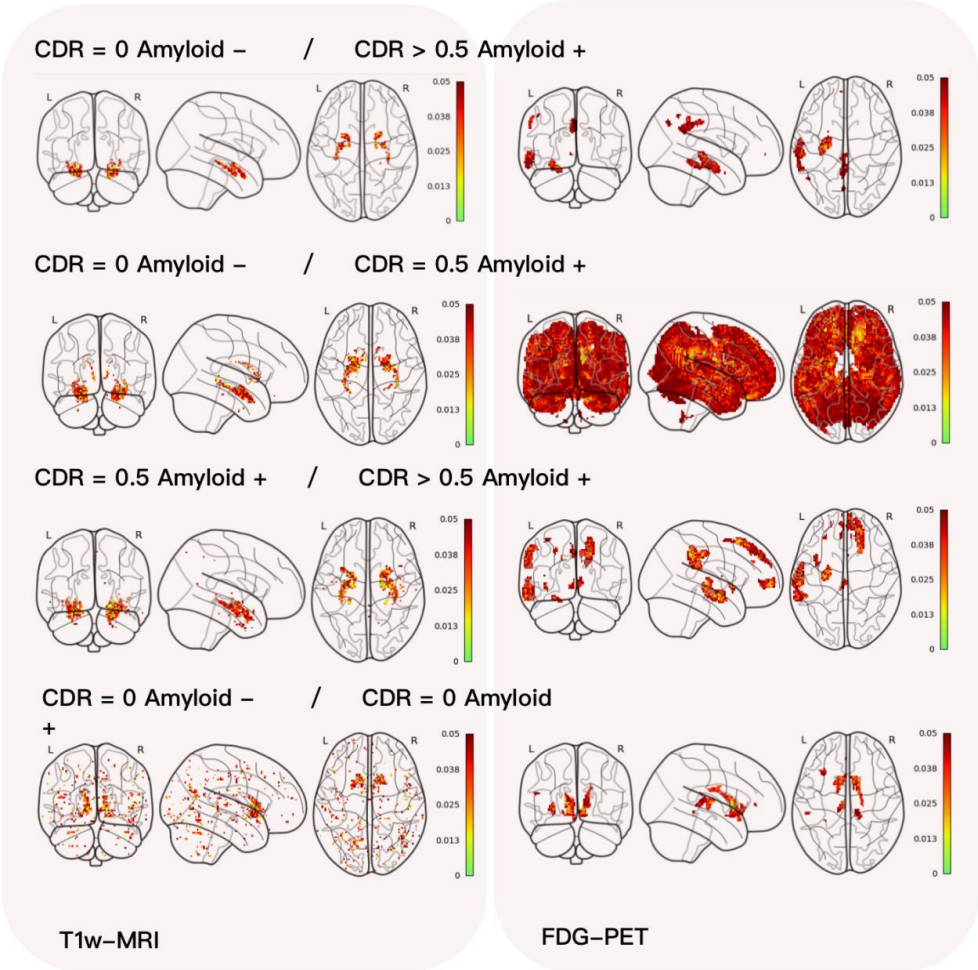

Figure 5: Activation maps for experiments with T1w-MRI and FDG-PET modalities.

Chapter 4: Discussion

We conducted four groups of experiments comparing different CDR values with positive or negative amyloid status. We utilized various imaging modalities as input, such as single modality (FDG-PET, T1w-MRI only) and multi-modal approaches (early fusion and late fusion). Consistent with state-of-the-art researches, our results demonstrate that under single modality cases, using the PET modality consistently outperforms the T1w-MRI modality. Contrary to the findings of Samper-González et al. (2018), our study did not observe an enhancement in AUC using multi-modality data across all SVM models, regardless of the fusion technique employed. This aligns with the findings of Narazani et al. (2022). However, we did observe improvements in the NPV (Negative Predictive Value) and specificity of the models, highlighting the potential of multimodal data fusion to enhance the accuracy of ruling out diseases in future applications.

However, there are limitations within our experimental pipeline that must be acknowledged. Firstly, the performance of the late fusion technique we used is highly contingent on the choice of performance metrics. In these experiments, we selected the AUC value as the performance metric to determine the optimal weights for combining two modalities, potentially introducing bias and inaccuracy into the late fusion results when comparing other performance metrics. Secondly, the late fusion technique employed differs from the multi-kernel method used by Samper-González et al. (2018), which involves training on multiple kernels before combining and normalizing them to derive the final prediction. This divergence in methodology could also contribute to the observed discrepancy wherein multi-modality imaging data input models did not surpass the performance of single-modality input models.

Moreover, while most activation maps align with our a priori hypotheses, we observed an anomalous activation map in the cognitively unimpaired and MCI group using FDG-PET input. This group displayed a notably dense cluster of activation points, indicating almost the whole brain contribute to the classification results. Given that this was an isolated observation and other experiments using the same data set yielded normal behavior, we can tentatively rule out data set issues. Further investigation is needed to understand the underlying cause of this anomaly.

Chapter 5: Conclusion

In this study, we sought to further compare the classification power of utilizing either single or combined PET and MRI imaging data under the A/T/N framework's "N" category in linear SVM. We conducted sixteen experiments across four different groups, differentiated by Clinical Dementia Rating (CDR) values and amyloid status (positive or negative). Our results consistently showed that FDG-PET outperformed T1w-MRI in scenarios where a single modality was used. However, with the two fusion techniques we employed, combining multimodalities did not improve the general performance of the models. There are still some limitations and areas for future work. The late fusion techniques we used are easily biased by the performance metrics selected for choosing weights. Multi-kernel SVM or other fusion techniques might offer more effective strategies and could potentially improve performance. Furthermore, the abnormal activation maps we observed need further exploration.

References

Aisen, P.S., Cummings, J., Jack, C.R. Jr., Morris, J.C., Sperling, R., Frölich, L., Jones, R.W., Dowsett, S.A., Matthews, B.R., Raskin, J., Scheltens, P., Dubois, B.: On the path to 2025: understanding the Alzheimer's disease continuum. Alzheimers Res Ther 9(1), 60 (2017)

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (5th edition). Arlington, Va.: American Psychiatric Publishing, 2013.

Arya, A.D., Verma, S.S., Chakarabarti, P., Chakrabarti, T., Elngar, A.A., Kamali, A.M., Nami, M.: A systematic review on machine learning and deep learning techniques in the effective diagnosis of Alzheimer's disease. Brain Inform 10(1), 17 (2023)

Bhushan, I., Kour, M., Kour, G., Gupta, S., Sharma, S., & Yadav, A. (2018). Alzheimer's disease: Causes & treatment – A review. Annals of Biotechnology, 1(1). MedDocs Publishers LLC. http://meddocsonline.org/annals-of-biotechnology/

Borson, S., Frank, L., Bayley, P.J., Boustani, M., Dean, M., Lin, P.J., McCarten, J.R., Morris, J.C., Salmon, D.P., Schmitt, F.A., Stefanacci, R.G., Mendiondo, M.S., Peschin, S., Hall, E.J., Fillit, H., Ashford, J.W.: Improving dementia care: the role of screening and detection of cognitive impairment. Alzheimers Dement 9(2), 151-159 (2013)

Chincarini, A., Bosco, P., Calvini, P., Gemme, G., Esposito, M., Olivieri, C., Rei, L., Squarcia, S., Rodriguez, G., Bellotti, R., Cerello, P., De Mitri, I., Retico, A., Nobili, F., et al.: Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease. Neuroimage 58(2), 469-480 (2011)

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 44(3), 837-845.

Ding, Y., Sohn, J.H., Kawczynski, M.G., Trivedi, H., Harnish, R., Jenkins, N.W., Lituiev, D., Copeland, T.P., Aboian, M.S., Mari Aparici, C., Behr, S.C., Flavell, R.R., Huang, S.Y.,

Zalocusky, K.A., Nardo, L., Seo, Y., Hawkins, R.A., Hernandez Pampaloni, M., Hadley, D., Franc, B.L.: A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain. Radiology 290(2), 456-464 (2019)

Ewers, M., Sperling, R.A., Klunk, W.E., Weiner, M.W., Hampel, H.: Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia. Trends Neurosci 34(8), 430-442 (2011)

Frisoni, G.B., Bocchetta, M., Chételat, G., Rabinovici, G.D., de Leon, M.J., Kaye, J., Reiman, E.M., Scheltens, P., Barkhof, F., Black, S.E., Brooks, D.J., Carrillo, M.C., Fox, N.C., Herholz, K., Nordberg, A., Jack, C.R. Jr., Jagust, W.J., Johnson, K.A., Rowe, C.C., Sperling, R.A., Thies, W., Wahlund, L.O., Weiner, M.W., Pasqualetti, P., Decarli, C., et al.: Imaging markers for Alzheimer disease: which vs how. Neurology 81(5), 487-500 (2013)

Grueso, S., Viejo-Sobera, R.: Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review. Alzheimers Res Ther 13(1), 162 (2021)

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage, 87, 96-110.

Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., et al.: Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. Neuroimage 55(2), 574-589 (2011)

Jack, C.R. Jr., Bennett, D.A., Blennow, K., Carrillo, M.C., Feldman, H.H., Frisoni, G.B., Hampel, H., Jagust, W.J., Johnson, K.A., Knopman, D.S., Petersen, R.C., Scheltens, P., Sperling, R.A., Dubois, B.: A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers. Neurology 87(5), 539-547 (2016)

Jack, C.R. Jr., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q.: Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol 9(1), 119-128 (2010)

Kim, S., Lee, P., Oh, K.T., et al.: Deep learning-based amyloid PET positivity classification model in the Alzheimer's disease continuum by using 2-[18F]FDG PET. EJNMMI Res 11, 56 (2021)

Mosconi, L., Berti, V., Glodzik, L., Pupi, A., De Santi, S., de Leon, M.J.: Pre-clinical detection of Alzheimer's disease using FDG-PET, with or without amyloid imaging. J Alzheimers Dis 20(3), 843-854 (2010)

Narazani, M., Sarasua, I., Pölsterl, S., Lizarraga, A., Yakushev, I., Wachinger, C.: Is a PET all you need? A multi-modal study for Alzheimer's disease using 3D CNNs. arXiv preprint arXiv:2207.02094 (2022).

Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., Bertrand, A., Bertin, H., Habert, M.O., Durrleman, S., Evgeniou, T., Colliot, O., et al.: Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. NeuroImage 183, 504–521 (2018)

Song, J., Zheng, J., Li, P., Lu, X., Zhu, G., Shen, P.: An Effective Multimodal Image Fusion Method Using MRI and PET for Alzheimer's Disease Diagnosis. Front Digit Health 3, 637386 (2021)

Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Iwatsubo, T., Jack, C.R. Jr., Kaye, J., Montine, T.J., Park, D.C., Reiman, E.M., Rowe, C.C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M.C., Thies, B., Morrison-Bogorad, M., Wagster, M.V., Phelps, C.H.: Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 7(3), 280-292 (2011)

Zhao, Z., Chuah, J.H., Lai, K.W., Chow, C.O., Gochoo, M., Dhanalakshmi, S., Wang, N., Bao, W., Wu, X.: Conventional machine learning and deep learning in Alzheimer's disease diagnosis using neuroimaging: A review. Front Comput Neurosci 17, 1038636 (2023)

Zhou, T., Thung, K.H., Zhu, X., Shen, D.: Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Hum Brain Mapp 40(3), 1001-1016 (2019)

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage, 55(3), 856