A Mutual Knowledge Distillation-Empowered AI Framework for Early Detection of Alzheimer's Disease Using Incomplete Multi-Modal Images

Min Gu Kwak[1], Yi Su[2], Kewei Chen[2], David Weidman[2], Teresa Wu[3], Fleming Lure[4], and Jing Li[1]

[1]School of Industrial and Systems Engineering, Georgia Institute of Technology, GA

[2]Banner Alzheimer's Institute, AZ

[3]School of Computing, Informatics and Decision Systems Engineering, Arizona State University, AZ

[4]MS Technologies Corporation, Rockville, MD

Abstract

Early detection of Alzheimer's Disease (AD) is crucial to ensure timely interventions and optimize treatment outcomes for patients. While integrating multi-modal neuroimages, such as MRI and PET, has shown great promise, limited research has been done to effectively handle incomplete multi-modal image datasets in the integration. To this end, we propose a deep learning-based framework that employs Mutual Knowledge Distillation (MKD) to jointly model different sub-cohorts based on their respective available image modalities. In MKD, the model with more modalities (e.g., MRI and PET) is considered a teacher while the model with fewer modalities (e.g., only MRI) is considered a student. Our proposed MKD framework includes three key components: First, we design a teacher model that is student-oriented, namely the Student-oriented Multi-modal Teacher (SMT), through multi-modal information disentanglement. Second, we train the student model by not only minimizing its classification errors but also learning from the SMT teacher. Third, we update the teacher model by transfer learning from the student's feature extractor because the student model is trained with more samples. Evaluations on Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets highlight the effectiveness of our method. Our work demonstrates the potential of using AI for addressing the challenges of incomplete multi-modal neuroimage datasets, opening new avenues for advancing early AD detection and treatment strategies.

1. Introduction

Alzheimer's Disease (AD), a fatal neurodegenerative disorder, is currently affecting many people worldwide. An estimated 6.7 million Americans age 65 and older are living with AD in 2023, and it is about 10.8% of people age 65 and older (Alzheimer's Association, 2023). Despite several decades of unsuccessful drug development, this year has signaled a glimmer of hope with the full FDA approval of a novel drug, Leqembi (Canady, 2023). Moreover, another promising medication, donanemab, is under testing and showing encouraging early results (Sims et al., 2023). Notably, these groundbreaking pharmaceutical developments herald a new era in the fight against AD. Yet, their potential to slow disease progression is contingent upon early administration, during the Mild Cognitive Impairment (MCI) phase before advancing to AD dementia. MCI is known to be heterogeneous, meaning that some individuals with MCI will convert to AD dementia while others' MCI may be due to some other non-AD-related brain diseases or conditions. Therefore, it is important to differentiate which MCI patients will convert to AD (Thung et al., 2016). The objective is to ensure the right patient receives the right treatment at the right time.

Detecting AD at its early stages presents significant challenges. Neuroimaging holds great promise, as indicated by national and international expert consensus groups, such as the working group convened by the National Institute of Aging and the Alzheimer's Association (NIA-AA), and the International Working Group. Accurate detection often requires integrating multi-modality datasets, including neuroimaging data capturing brain structure and function from various perspectives, such as magnetic resonance imaging (MRI), positron emission tomography (PET), etc. (Liu et al., 2018). However, such data integration necessitates highly trained dementia specialists, a resource that remains scarce. Herein lies an excellent opportunity for artificial intelligence (AI) to bridge this gap, aiding clinicians and significantly improving early AD detection by integrating multi-modal datasets.

Nonetheless, integrating multi-modal neuroimages to detect AD faces significant challenges,

primarily due to the variable availability of different modalities among patients. Factors like cost, limited clinic access, or safety concerns may restrict some patients from undergoing specific imaging examinations, creating distinct sub-cohorts of patients with various available image modalities. For example, with two modalities (MRI and PET), there may be two sub-cohorts: one with only MRI and another with both MRI and PET. As the number of modalities increases, the number of sub-cohorts also rises.

To address this challenge comprehensively, our goal is to train a collection of models, each tailored to a specific sub-cohort with the same available modalities. This ensures that, during deployment, an appropriate model is available to predict conversion to AD for each patient with any combination of available modalities. This versatility is crucial for the AI system to be universally applicable, rather than restricted to patients with specific modality combinations.

Machine learning models addressing incomplete modalities in AD tasks have attracted significant attention from researchers. This challenge is distinct from typical missing data imputation because a missing modality results in the loss of all its encompassed features all at once. An autoencoder-based missing modality completion method with graph regularization was proposed for AD diagnosis (Liu et al., 2021). A framework was proposed for AD diagnosis that utilizes a latent representation space, where complete multi-modality data forms a common representation and incomplete data informs modality-specific representations (Zhou et al., 2019). A pairwise feature-based generation adversarial network was introduced that leverages MRI features to generate corresponding PET features, reinforced by real PET constraints, and incorporates an attention mechanism to retain structural integrity (Ye et al., 2023). However, these existing methods used pre-defined extracted features from images.

We focus on image-based deep learning (DL) models using incomplete multi-modal datasets. Due to the high dimensionality and the spatial organization of image data, existing feature-based methods are not easily adaptable. In image-based DL models, various algorithms have been proposed to integrate multi-modal images (Liu et al., 2018; Song et al., 2021; Zhang et al., 2019), while limited work has been done to tackle incomplete multi-modal image datasets (Chen et al, 2023).

To address the gap in the literature, we propose a DL-based framework that employs mutual knowledge distillation (MKD) to jointly learn predictive models for sub-cohorts with varying availability of image modalities. The core concept behind MKD is the mutual exchange of knowledge between two models: a student model and a teacher model. The student model, with a subset of modalities included in the teacher model, learns from the higher predictive capacity of the teacher. Conversely, the teacher can leverage the encoder of available modalities in the student. This is possible because the student model is trained on fewer modalities compared to the teacher allowing it to have more training samples and thus can learn better feature representations. Figure 1 provides a high-level conceptual depiction of MKD framework in our context. The contribution of this paper is summarized as follows:

- We proposed a novel image-based DL framework using MKD to jointly model sub-cohorts with different missing modality patterns. Different from existing KD methods that are one-directional (from teacher to student) (Garcia et al., 2018; Hu et al., 2020), our method reinforces both teacher and student models in bi-directional manner.
- We designed a novel teacher model that is “student-oriented” through multi-modal information disentanglement, in order to best facilitate the student’s KD process, instead of being a general-purpose model like most existing KD methods.
- We applied the MKD framework to the early detection of AD using multi-modal image datasets with missing modalities and achieved promising results. Our work demonstrated the potential of using AI for addressing the challenges of incomplete multi-modal neuroimage datasets, opening new avenues for advancing early AD detection and treatment strategies.

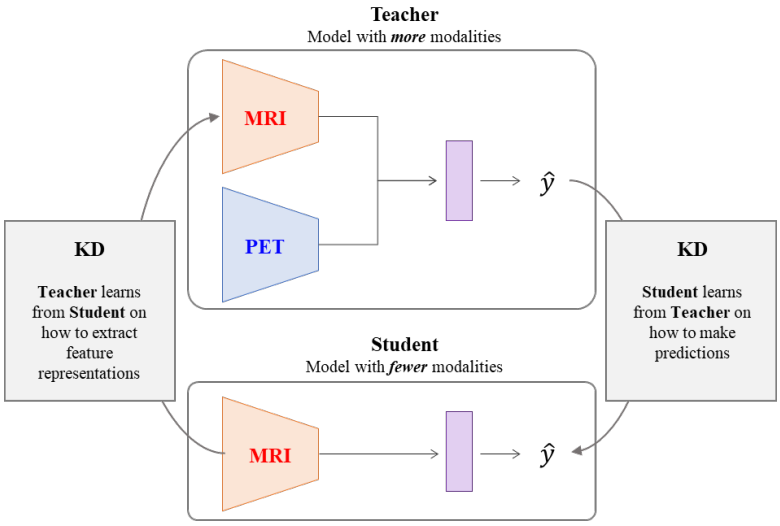

Figure 1. A conceptual depiction of the proposed MKD framework.

2. Proposed Method

The proposed MKD framework will be presented in the context of two imaging modalities: MRI and PET. While all MCI patients have MRI as MRI is part of the standard of care in AD-related clinical examinations, some patients may not have PET due to its high cost. Therefore, the teacher model is one that predicts AD conversion using both MRI and PET, while the student model only uses MRI. In this context, the proposed MKD framework includes three steps: First, we design a teacher model that is student-oriented, namely the SMT model. Second, we train the student model by not only minimizing its classification errors but also learning from the SMT teacher. Third, we update the teacher model by leveraging the student's MRI feature extractor because the student is trained with more samples (patients with only MRI). The three steps are presented in Sec. 2.1-2.3, respectively.

2.1. Design a Student-oriented Multi-modal Teacher (SMT) model

Knowledge Distillation (KD) has proven effective in transferring knowledge from one DL model to another, using a teacher-student learning framework. Initially, KD was developed for model compression, aiming to train a lightweight student that matches the performance of a large-scale, sophisticated teacher for deployment benefits (Wang and Yoon, 2021; Hinton et al., 2015). In recent years, KD has been extended to transfer knowledge across different modalities, known as cross-modal KD (Thoker et al., 2019; Valverde et al., 2021; Xue et al., 2021). However, the success of cross-modal KD is highly dependent on training the teacher model to tease out modality-common information, which is transferable to the student (Xue et al., 2022).

To address this challenge, we propose a Student-oriented Multi-modal Teacher (SMT) model, which learns modality-common and modality-specific representations to disentangle information from multi-modal datasets. This helps reduce the burden of the classification task, as it removes redundant and noisy information from the input modalities. Also, the modality-common representation is essential to make the multi-modal teacher an effective teacher for the student.

To learn these representations, we incorporate a combination of losses that include a similarity loss (for helping extract common representations from the different modalities), a difference loss (for helping extract modality-specific representations), a reconstruction loss (for regularizing the representation learning), and a classification loss. Herein, we introduce the details of SMT design.

Consider the case of two modalities for notation simplicity, while the proposed method can be extended to more modalities. Let X_1 and X_2 denote the input modalities such as MRI and PET images, respectively. The task is a binary classification of an MCI patient as AD converter or non-converter.

The architecture of SMT includes several subnetworks for feature extraction, representation encoding,

representation decoding, and classification. In the feature extraction subnetwork, we map each modality X_1 to a latent vector h_1. The subnetwork can use an existing architecture capable of handling image data as backbone, such as ResNet50 (He et al., 2016), followed by Global Average Pooling and a fully connected projector layer. For each input modality X_i, where $\forall i \in \{1,2\}$, the input data x_i is processed through the feature extractor subnetwork, F_i, resulting in latent vector h_i:

$$h_i = F_i(x_i). \tag{1}$$

The latent vector h_i is subsequently passed through a modality-specific encoder E_i^s, yielding:

$$z_i^s = E_i^s(h_i). \tag{2}$$

In a parallel manner, all the latent vectors are fed into a modality-common encoder E^c:

$$z_i^c = E^c(h_i). \tag{3}$$

Note that there is only a single modality-common encoder, and it processes the feature representations from all modalities. These encoders are simple feed-forward neural network layers that transform the modality-wise latent vector, h_i, into modality-common and modality-specific representations, z_i^c and z_i^s, respectively. To make the representation scale from different modalities the same, all the representations are L2-normalized. Then, the summation of z_i^c and z_i^s is passed through a decoder D_i and to reconstruct h_i:

$$\hat{h}_i = D_i(z_i^c + z_i^s), \tag{4}$$

where $\hat{h}_i$ denotes the reconstructed latent vector for modality X_i.

Also, the representations are passed through a classifier C to predict the classification label. Only the modality-common representations are passed for training the best teacher to facilitate the learning of the student by the teacher, which is the SMT model presented in this section. We denote the process as follows:

$$\hat{y} = C(\textstyle\sum_i z_i^c). \tag{5}$$

The overall learning of the model is performed by minimizing the following loss function:

$$\mathcal{L} = \alpha_{\text{sim}}\mathcal{L}_{sim} + \alpha_{diff}\mathcal{L}_{diff} + \alpha_{recon}\mathcal{L}_{recon} + \alpha_{class}\mathcal{L}_{class}, \tag{6}$$

where each α denotes a balancing hyperparameter.

$\mathcal{L}_{sim}$ aims to minimize the difference between the representations of each modality output from the common encoder. This helps align representations from different modalities in a shared subspace. We used cosine similarity to measure this difference as follows:

$$\mathcal{L}_{sim} = 1 - z_1^c \cdot z_2^c. \tag{7}$$

Herein, the inner dot product of z^c can be considered as cosine similarity because z^c is L2-normalized. We added one to the negative cosine similarity to make the minimum value of $\mathcal{L}_{sim}$ as zero for simplicity.

$\mathcal{L}_{diff}$ ensures that the modality-common and -specific representations capture different aspects of the input. This non-redundancy is achieved by enforcing a soft orthogonality constraint between the common- and specific-representations within each modality as well as between the specific representations across the modalities as follows:

$$\mathcal{L}_{diff} = 3 + z_1^c \cdot z_1^s + z_2^c \cdot z_2^s + z_1^s \cdot z_2^s. \tag{8}$$

$\mathcal{L}_{rec}$ is to help avoid extracting trivial features by the encoders that do not contain representative information of each modality. We used mean squared error (MSE) between h_i and $\hat{h}_i$ as the reconstruction loss. Finally, we applied cross-entropy loss to calculate $\mathcal{L}_{class}$. Figure 2 provides a

graphical overview of the SMT model.

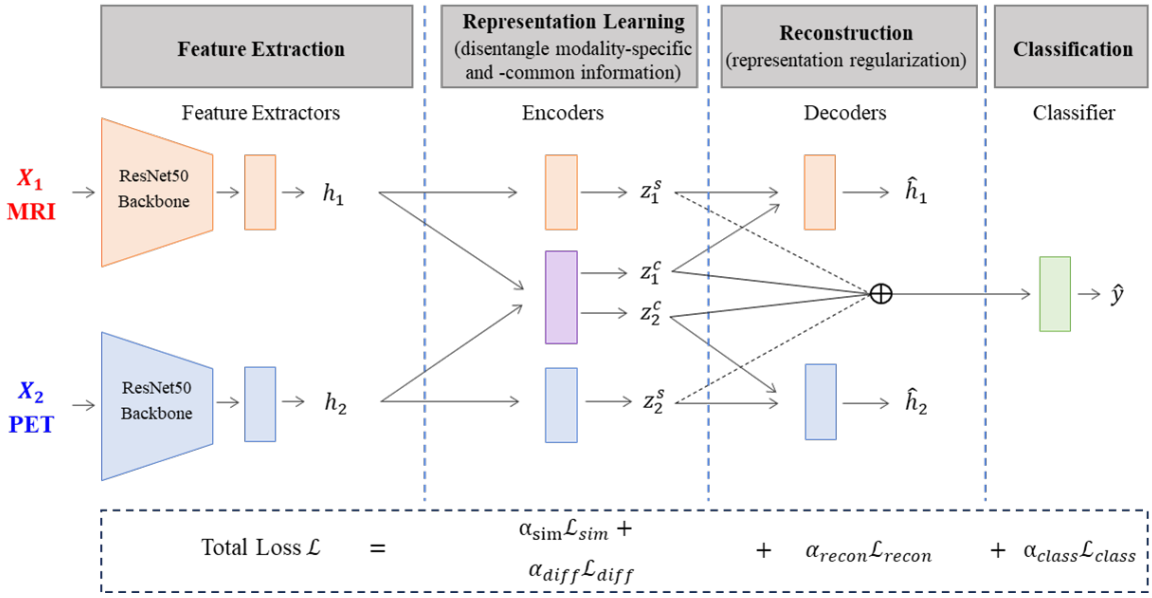

Figure 2. Graphical overview of the proposed SMT model.

2.2. Design the Student Model by Learning from the SMT

Recall that the student model takes a subset of input modalities from the teacher model. For example, considering two modalities X_1 and X_2 such as MRI and PET, as input modalities of the teacher, the student uses only X_1 (MRI) as input. The task of the student model is to predict MCI conversion solely with X_1. The student architecture is designed to resemble the branch of the teacher involving X_1, facilitating KD. Specifically, the student model involves subnetworks for feature extraction, representation encoding, and classification ($F_1, E^c,$ and C, respectively). There are three remarks we want to highlight: (1) The student model has only one encoder and does not conduct disentanglement between modality-common and -specific representations because it receives only X_1. (2) The student architecture does not include a decoder for reconstruction. While adding a decoder is straightforward, we found in our experiments that a decoder does not help improve the classification performance of the student model. Our interpretation is that there is a trade-off between reconstruction and classification tasks of student model. Thus, we did not include it in the student design for simplicity. (3) Because we intended to make the student take advantages of classification performance of teacher (multiple modalities), we designed the student architecture to resemble the teacher. This provides the opportunity to use the pre-trained weights in the corresponding teacher network.

During the training phase of the teacher model (i.e., SMT), it adds modality-common and -specific representations and subsequently processes this representation through classifier. Conversely, the student model does not differentiate between these two representations. To leverage the pre-trained weights as proper initial weights, it is imperative that the representations fed into the classifier maintain similar scales. Both z_1^c and z_2^c in the teacher model are L2-normalized, and they are trained to have high cosine similarity. It ensures their scales are consistent and their directional attributes are analogous. Therefore, to maintain this consistency in representation scale for the student model, we doubled the value of $z_1 = E^c(F_1(x_1))$ when using it as an input for the student classifier.

The overall learning of the student model is performed by minimizing the following loss function:

$$\mathcal{L} = \alpha_{class}\mathcal{L}_{class} + \alpha_s\mathcal{L}_{KL}(\hat{y}_s, \hat{y}_t; \tau_s). \quad (9)$$

$\mathcal{L}_{KL}$ is the Kullback-Leibler (KL) divergence loss that measures the difference between predicted logits of student and teacher (i.e., SMT), $\hat{y}_s$ and $\hat{y}_t$, respectively. τ_s is the scaling hyperparameter for logits and it allows the student to learn the dark knowledge of the teacher (Hinton et al., 2015). α_s is a hyperparameter of balancing the KL divergence loss for training the student.

2.3. Update the Teacher Model by Learning from the Student

The student model is exclusively trained on X_1, allowing it to utilize more samples than the teacher. While the inability to use X_2 might compromise its classification performance, it excels at extracting feature representations from X_1. This capability can be exploited to improve the performance of SMT. For SMT, the weights of F_1 and E_1^c are initialized using the weights from the corresponding subnetworks of the student model, while the weights for the remaining subnetworks are randomly initialized. The training procedure adheres to the one described in Section 2.1, but with an added KD loss to enable the transfer of X_1 feature extraction knowledge from the student to SMT. For performing representation-level KD, we employed the following loss function:

$$\mathcal{L}_{KD}^{repr} = \frac{1-z_{1,s}^c \cdot z_{1,t}^c}{2\tau_t^2}, \tag{10}$$

where $z_{1,s}^c, z_{1,t}^c$, and τ_t denote X_1 representation of student and teacher, and scaling hyperparameter, respectively. In this step, we additionally include modality-specific representations for predicting classification label. This inclusion in the classification helps improve the performance of the updated teacher model. $\mathcal{L}_{KD}^{repr}$ is added to Equation (6) with a balancing hyperparameter of α_t.

3. Application in Early Detection of AD

3.1. Data

To evaluate the proposed method, we conducted experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. ADNI is one of the largest datasets for AD studies to date, with the primary goal being to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the MCI conversion. We downloaded 3D 857 MRI and 614 AV45-PET images from 961 MCI patients. The patients who converted to AD within 36 months were assigned as converters, otherwise as non-converters. There were 614 pairs of MRI and PET, and 243 MRI-only images. We randomly split the paired images into 80% for training, 10% for validation and 10% for testing while preserving the class distribution. We also ensured that the patients in the training set were not present in validation and testing sets.

The MRI scans were spatially normalized using the Computational Anatomy Toolbox 12 (CAT12) (Gaser et al., 2022) with Statistical Parametric Mapping (SPM12) (Ashburner et al., 2014) and a standard brain atlas from the Montreal Neuroimaging Institute (MNI). Then, each AV45-PET image was co-registered to the corresponding MRI. Then, we applied zero padding and resizing to both MRI and PET images to apply widely used data augmentation techniques and reduce the computational cost. We obtained images with a size of $72 \times 72 \times 72$.

3.2. Model Architecture and Training Hyperparameters

The architecture of our proposed network is comprised of several distinct components. The feature extractor leverages a ResNet-50 backbone, followed by a single-layer module characterized by a 128-dimensional output, Leaky ReLU activation, and subsequent Layer Normalization. The encoder is designed with a sequence of a 64-dimensional layer, Layer Normalization, a sigmoid activation, followed by another 64-dimensional layer. The decoder consists of a 64-dimensional layer employing a sigmoid activation, followed by Layer Normalization, and concludes with a 128-dimensional layer. We employed a straightforward single-layer classifier.

The training hyperparameters were determined through grid-search on validation set, ensuring optimal performance. Across all models, the AdamW optimizer (Loshchilov & Hutter, 2017) was employed, characterized by a weight decay of 0.0001 and accompanied by a half-cosine learning rate scheduling. A batch size was set to 16. We trained SMT model for 100 epochs, adopting a learning rate of 0.001. On the other hand, the student model was trained for 30 epochs with a more conservative learning rate set at 0.0001. Updating the SMT by the student adhered to the same hyperparameters as training the SMT. Furthermore, we introduced balancing hyperparameters: $\alpha_{class} = 1.0, \alpha_{sim} = 10.0, \alpha_{diff} = 5.0, \alpha_{recon} = 0.1, \alpha_s = 100.0$, and $\alpha_t = 500.0$. Both τ_s and τ_t were set to 5.0.

For data augmentation, we employed random flipping and random rotation. Due to the inherent computational challenges associated with 3D images and the demands of extensive sample sizes, we pivoted to a strategy of generating 2D slices (Liu et al., 2018). These slices were derived from sagittal, coronal, and axial orientations. Starting from the central point of each orientation, we consistently and evenly extracted slices at 3-voxel intervals. Including the center, we obtained 11 slices for each orientation, amounting to 33 unique 2D slices from a single 3D image. Throughout the training phase, all operations, from feature extraction to predictions, were performed on these 2D slices. For inference, however, we used the average value of the logit for appropriate evaluation.

3.3. Experimental Results

We first conducted an experiment to validate the effectiveness of solely leveraging the modality-common representation for training a teacher model and its subsequent impact on KD. We examined four model combinations: a teacher trained with both modality-common and -specific representations, its derived student, a teacher exclusively trained on modality-common representation, and its corresponding student. The teacher models are tested with pairs of MRI and PET, while the student models are tested with only MRI. For all experiments, we conducted ten repeated trials with different random seeds and reported the average values. We used the area under the receiver operating characteristics (AUROC), accuracy, sensitivity, and specificity as evaluation metrics.

As shown in Table 1, the teacher model utilizing both modalities yielded the highest AUROC of 0.8802, indicating the benefits of a comprehensive modality training approach. The SMT model, focusing solely on the modality-common representation, obtained a lower AUROC of 0.8727. However, when KD is conducted, the student model of SMT outperformed its counterpart with AUROC of 0.7956. This finding underscores that while a holistic modality training can be advantageous, the student-oriented training approach that focuses on modality-common representation exhibits pronounced merits, especially when predicting the conversion of sub-cohorts with only MRI scans.

Table 1. Comparison of classification performance for teacher and student models with and without using modality-specific representation. The best performing teacher and student is in **bold**.

Representation	Model	AUROC	Accuracy	Sensitivity	Specificity
Common and Specific	Teacher	**0.8802**	**0.8482**	**0.7892**	**0.8798**
	Student	0.7685	0.7415	0.6625	0.7605
Common	Teacher (SMT)	0.8727	0.8319	0.7855	0.8479
	Student	**0.7956**	**0.7639**	**0.6803**	**0.7864**

To evaluate the robustness and effectiveness of MKD in our proposed framework, we conducted experiments under various data missing rates. Although our base training dataset has a missing rate of 0.33, we increased this to 0.50 and 0.70 on purpose to test the model when the number of PET images is substantially fewer compared to MRIs. This adjustment was achieved by arbitrarily removing PET images from the training set.

Table 2 demonstrates the comparative performance of MRI-only models when trained with MKD and when trained without MKD under varying missing rate scenarios. The MRI-only models with MKD are trained through KD from SMT, which leverages both MRI and PET datasets. In contrast, the MRI-only model without MKD is solely trained from scratch using just the MRI dataset. In result, models with MKD consistently outperformed the model without MKD regardless of the missing rate. Specifically, the MRI-only model with MKD achieved an AUROC of 0.7336 at a missing rate of 0.70. This performance is markedly better than the 0.7110 AUROC of the MRI-only model trained without MKD. This highlights the efficacy of MKD in predicting MCI conversion when only MRI data is available for patients. As such, performance degradation exists in MRI-only testing as the missing rate increases. Yet, using MKD still provides a notable advantage, proving its worth in AD tasks.

Table 2. Classification performance of MRI-only models with or without MKD under various missing rates. The

MRI-only model without MKD is trained exclusively on MRIs, and consequently has no missing rate.

Model	Missing Rate	AUROC	Accuracy	Sensitivity	Specificity
MRI-only with MKD	0.33	0.7956	0.7639	0.6803	0.7864
	0.50	0.7795	0.7234	0.6775	0.7545
	0.70	0.7336	0.7063	0.6609	0.7214
MRI-only without MKD	NA	0.7110	0.6999	0.6457	0.7183

Table 3 presents the performance results of models trained with both MRI and PET data, comparing outcomes when using MKD versus not using MKD. The MRI & PET model with MKD utilizes the MRI feature extraction capability from the previously trained MRI-only with MKD model. On the other hand, the MRI & PET model without MKD is solely trained on the paired MRI and PET data, with representations from each modality combined without disentanglement.

Across all missing rates, the models employing MKD consistently outperformed their counterparts. The improvement gap is less pronounced compared to the MRI-only model results because PET data carries more critical information than MRI. Even though MRI feature extraction capabilities were enhanced, its impact on classification performance remained relatively small than MRI-only case. Nonetheless, the highest performance gain was observed at a missing rate of 0.70. This aligns with the MRI-only results in Table 2, suggesting that our proposed framework is particularly beneficial for AD applications with a high rate of missing modalities.

Table 3. Classification performance of MRI & PET models with or without MKD under various missing rates.

Model	Missing Rate	AUROC	Accuracy	Sensitivity	Specificity
MRI & PET with MKD	0.33	0.8821	0.8436	0.7933	0.8503
	0.50	0.8767	0.8358	0.7910	0.8522
	0.70	0.8642	0.8186	0.7602	0.8334
MRI & PET without MKD	0.33	0.8817	0.8411	0.7915	0.8512
	0.50	0.8702	0.8322	0.7815	0.8501
	0.70	0.8522	0.8087	0.7580	0.8296

4. Conclusions

In this paper, we addressed the challenge of early detection of AD, focusing on the variability in image modality availability among patients. Our innovative DL-based framework employs MKD to jointly model different sub-cohorts based on their respective available image modalities. The bi-directional nature of our method ensures a mutual exchange of knowledge between student and teacher models. Furthermore, the student-oriented design of our teacher model, i.e., SMT, emerged from the necessity to best facilitate the KD process. It strategically emphasized modality-common information, facilitating the learning of the student from the teacher . We also reinforced the teacher model by utilizing the representation extraction capability of the student model. The experimental results showed that the proposed method contributed to considerable classification performance gain in multi-modal and single-modal scenarios with varying missing rates. To the best of our knowledge, this is the first study to propose MKD for jointly modeling sub-cohorts of patients with varying available image modalities in early detection of AD.

While our model showed promising results, there are several limitations and intriguing directions to extend our work. First, there are many hyperparameters that require time-intensive tuning to find their optimal values. This indicates a need for simplification, possibly by streamlining the model structure or combining the similarity and difference loss components. It can also be beneficial when multiple missing modalities exist. Second, we can investigate the potential of integrating other imaging modalities and clinical data into our framework, aiming for a more holistic patient assessment. Lastly, we plan to analyze using modalities that possess similarly decisive information for the task. We can

more clearly assess the effectiveness of our model in future studies.

Acknowledgments

This research is supported by NIH grant 2R42AG053149-02A1 and NSF grant DMS-2053170.

References

2023 Alzheimer's disease facts and figures. Alzheimers Dement. 2023 Apr;19(4):1598-1695. doi: 10.1002/alz.13016. Epub 2023 Mar 14. PMID: 36918389.

Ashburner, J., Barnes, G., Chen, C. C., Daunizeau, J., Flandin, G., Friston, K., ... & Penny, W. (2014). SPM12 manual. Wellcome Trust Centre for Neuroimaging, London, UK, 2464(4).

Canady, V. A. (2023). FDA approves new treatment for Alzheimer's disease. Mental Health Weekly, 33(3), 6-7.

Chen, Y., Pan, Y., Xia, Y., & Yuan, Y. (2023). Disentangle First, Then Distill: A Unified Framework for Missing Modality Imputation and Alzheimer's Disease Diagnosis. IEEE Transactions on Medical Imaging.

Garcia, N. C., Morerio, P., & Murino, V. (2018). Modality distillation with multiple stream networks for action recognition. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 103-118).

Gaser, C., Dahnke, R., Thompson, P. M., Kurth, F., Luders, E., & Alzheimer's Disease Neuroimaging Initiative. (2022). CAT–A computational anatomy toolbox for the analysis of structural MRI data. biorxiv, 2022-06.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., & Gori, P. (2020). Knowledge distillation from multi-modal to mono-modal segmentation networks. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23 (pp. 772-781). Springer International Publishing.

Liu, M., Cheng, D., Wang, K., Wang, Y., & Alzheimer's Disease Neuroimaging Initiative. (2018). Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis. Neuroinformatics, 16, 295-308.

Liu, M., Cheng, D., Yan, W., & Alzheimer's Disease Neuroimaging Initiative. (2018). Classification of Alzheimer's disease by combination of convolutional and recurrent neural networks using FDG-PET images. Frontiers in neuroinformatics, 12, 35.

Liu, X., Chen, K., Wu, T., Weidman, D., Lure, F., & Li, J. (2018). Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. Translational Research, 194, 56-67.

Liu, Y., Fan, L., Zhang, C., Zhou, T., Xiao, Z., Geng, L., & Shen, D. (2021). Incomplete multi-modal representation learning for Alzheimer's disease diagnosis. Medical Image Analysis, 69, 101953.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Sims, J. R., Zimmer, J. A., Evans, C. D., Lu, M., Ardayfio, P., Sparks, J., ... & Kaul, S. (2023). Donanemab in early symptomatic Alzheimer disease: the TRAILBLAZER-ALZ 2 randomized clinical trial. JAMA.

Song, J., Zheng, J., Li, P., Lu, X., Zhu, G., & Shen, P. (2021). An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis. Frontiers in digital health, 3, 637386.

Thoker, F. M., & Gall, J. (2019, September). Cross-modal knowledge distillation for action recognition. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 6-10). IEEE.

Thung, K. H., Wee, C. Y., Yap, P. T., & Shen, D. (2016). Identification of progressive mild cognitive impairment patients using incomplete longitudinal MRI scans. Brain Structure and Function, 221, 3979-3995.

Valverde, F. R., Hurtado, J. V., & Valada, A. (2021). There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11612-11621).
Wang, L., & Yoon, K. J. (2021). Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE transactions on pattern analysis and machine intelligence, 44(6), 3048-3068.
Xue, Z., Gao, Z., Ren, S., & Zhao, H. (2022, September). The Modality Focusing Hypothesis: Towards Understanding Crossmodal Knowledge Distillation. In The Eleventh International Conference on Learning Representations.
Xue, Z., Ren, S., Gao, Z., & Zhao, H. (2021). Multimodal knowledge expansion. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 854-863).
Ye, H., Zhu, Q., Yao, Y., Jin, Y., & Zhang, D. (2023). Pairwise feature-based generative adversarial network for incomplete multi-modal alzheimer's disease diagnosis. The Visual Computer, 39(6), 2235-2244.
Zhang, F., Li, Z., Zhang, B., Du, H., Wang, B., & Zhang, X. (2019). Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease. Neurocomputing, 361, 185-195.
Zhou, T., Liu, M., Thung, K. H., & Shen, D. (2019). Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data. IEEE transactions on medical imaging, 38(10), 2411-2422.