Optimal Transport-Based Early Detection of Mild Cognitive Impairment Patients Based on Magnetic Resonance Images

Ziyu Liu
Purdue University Department of Statistics
Travis Johnson
Indiana University School of Medicine
Wei Shao
Indiana University School of Medicine
Min Zhang
Purdue University Department of Statistics
Jie Zhang
Indiana University School of Medicine
Kun Huang (kunhuang@iu.edu)
Indiana University School of Medicine https://orcid.org/0000-0002-8530-370X

Research

Keywords: Transfer Learning, Optimal Transport, Bootstrap Aggregation

Posted Date: February 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-246076/v1

RESEARCH

Optimal transport-based early detection of mild cognitive impairment patients based on magnetic resonance images

Ziyu Liu[1], Travis S. Johnson[2], Wei Shao[2], Min Zhang[1], Jie Zhang[4] and Kun Huang[2,3*]

Abstract

Background: To help clinicians provide timely treatment and delay disease progress, it is crucial to identify dementia patients during the mild cognitive impairment (MCI) stage and stratify these MCI patients into early and late MCI stages before they progress to alzheimer's disease (AD). In the process of diagnosing MCI and AD in living patients, brain scans are regularly collected using neuroimaging technologies such as computed tomography (CT), magnetic resonance imaging (MRI), or positron emission tomography (PET). These brain scans measure the volume and molecular activity within the brain resulting in a very promising avenue to diagnose patients early in a non-invasive manner.

Methods: We have developed an optimal transport based transfer learning model to discriminate between early and late MCI. Combing this transfer learning model with bootstrap aggregation strategy, we overcome the over-fitting problem and improve model stability and prediction accuracy.

Results: With the transfer learning methods that we have developed, we outperform the current state of the art MCI stage classification frameworks and show that it is crucial to leverage alzheimer's disease and normal control subjects to accurately predict early and late stage cognitive impairment.

Conclusions: Our method is the current state of the art based on benchmark comparisons. This method is a necessary technological stepping stone to widespread clinical usage of MRI based early detection of AD.

Keywords: Transfer Learning; Optimal Transport; Bootstrap Aggregation

Background

Alzheimer's disease (AD) is an irreversible, degenerative brain disorder, affecting over six million Americans and is the sixth leading cause of death in the United States [1]. AD is hallmarked by neuron loss [2], inflammation [3], amyloid plaques [4], and tau deposition [5], which lead to progressive tissue loss in the brain and cognitive decline in the patient [6]. Diagnosing AD is largely based on tests of cognitive impairment combined with technologies such as computed tomography (CT), magnetic resonance imaging (MRI), or positron emission tomography (PET) but can only be verified after death on the postmortem brain [7]. Patients who have not yet progressed to AD may also be diagnosed with the AD precursor condition mild cognitive impairment (MCI). To help clinicians to provide timely treatment and delay the disease progress, it is crucial to identify patients during the MCI stage and stratify MCI patients into early and late MCI stages. In the process of diagnosing MCI and AD in living patients, brain scans are also regularly collected using neuroimaging technologies such as CT, MRI, and PET to rule out other potential causes of the disease. These brain scans measure the volume and molecular activity within the brain resulting in a very promising avenue to diagnose patients early in an non-invasive manner.

Specifically, neuroimaging techniques enables us to identify regions of interests related to AD [8] and extract sensitive markers for AD. It has been demonstrated that voxel-based measures extracted from structural MRI (VBM-MRI) and fluorodeoxyglucos PET (FDG-PET) can help us investigate the neurophysiological feature of AD and MCI [9, 10]. These features can be utilized to diagnose the early stage of AD patients and predict whether a MCI patient will progress to AD [11]. We seek to utilize these features

*Correspondence: kunhuang@iu.edu
[2]Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, USA
[3]Regenstrief Institute, Indianapolis, USA
Full list of author information is available at the end of the article

for distinguishing the early stage MCI (E-MCI) versus late stage MCI (L-MCI) as a classification task.

Recent progress in the machine learning (ML) and pattern recognition methods shed light on diagnosis of AD and MCI patients with the help of neuroimaging features. Despite the wide applications of ML models in biomedical problems, there are two major challenges in determining MCI stages, namely that the collection of multiple-modality datasets is costly and time consuming, and that the effect size observed between early and late stages of MCI is much smaller than between AD and normal control (NC) subjects who are not cognitively impaired. Accordingly, it is of great interest to develop ML models for utilizing samples from related and easier to train tasks with data that are more readily available such as AD patients versus NC patients and transfer the knowledge to the more challenging task of predicting MCI stage. Some previous works [12, 11] introduced auxiliary tasks such as AD and NC classification task to identify disease related features and construct the decision function for classification. Transferring knowledge from different but related auxiliary task to increase the prediction accuracy on a more difficult target task is a widely used machine learning strategy called transfer learning (TL). TL uses heterogeneous data and has to face the challenging ML task as the decision function learned from the source (auxiliary) task cannot be directly applied to the target domain. Two heterogeneous datasets will occupy different distributions in the feature spaces, which is termed distributional drift. Traditional TL techniques adopt sample weighting strategies and feature alignment strategies [13] to overcome the distributional drifting problem. Recently, Optimal Transport(OT) theory has been successfully introduced in TL problems [14, 15]. Since OT has shown great promise in tackling the data drifting (target shifting) issue, we adopt it in our model to address the difficulty of utilizing AD and NC samples for tackling our problem.

Our model consists three main components: feature selection, transfer learning, and bootstrap aggregation. We will first use the rMLTFL [11] framework as well as traditional one-way ANOVA to select representative features from voxel-based morphometry MRI (VBM-MRI) and fluorodeoxyglucose PET (FDG-PET) modalities. Then, we will develop the OT transfer learning strategies to train classifiers for L-MCI and E-MCI with the help of AD and NC samples. Finally, we will apply the Bootstrap Aggregation (BAg) strategy to overcome the overfitting problem and improve stability and accuracy.

Method

Data collection and preprocessing

The Alzheimer's Disease Neuroimaging Initiative (ADNI) provides researchers with multi-modal longitudinal data for subjects as they work to define the progression of AD. The ADNI-1 dataset contains 202 subjects with VBM-MRI and FDG-PET brain images. The updated dataset ADNI-2 assessed participants from the ADNI-1 phase besides new participant groups including elderly controls and subjects with significant memory concern, E-MCI, and L-MCI. We summarize the samples used in our study in Table 1.

Table 1 The values are expressed as mean ± standard deviation. AD=Alzheimer's disease, NC=Normal Control, E-MCI=Early Mild Cognitive Impairment, L-MCI=Late Mild Cognitive Impairment, MMSE = the Mini-Mental State Examination, and CDR = the clinical dementia rating.

	NC	E-MCI	L-MCI	AD
Number	211	273	187	160
Gender (M/F)	190/101	153/119	108/76	95/65
Age	76.1±6.5	71.5±7.1	73.9±8.4	75.2±7.9
Education	16.4±2.6	16.1±2.6	16.4±2.8	15.9±2.8
MMSE	29.0±1.2	28.4±1.5	27.7±1.7	24.0±2.6
CDR	0.0±0.1	0.5±0.1	0.5±0.1	0.7±0.3

The feature extraction process includes image registration, region of interests selection, and feature quantification. We specifically use the morphometry features extracted from VBM-MRI and FDG-PET images previously extracted by previous study [16] and denote the two classes of features as VBM and FDG features. The details of feature extraction can be found in [16].

Feature selection

To reasonably utilize informative features from the two data modalities, we use the robust multi-label transfer feature learning (rMLTFL) model [11] to filter out features which are irrelevant to the classification task. In [11], this model was applied to select features to train a support vector machine (SVM) model for distinguishing Progressive MCI and Stable MCI. This framework can helps identify features related to the target task (L-MCI vs E-MCI) that benefit from auxiliary tasks (AD vs NC, AD vs MCI, MCI vs NC). However, it faces a difficult situation that separating E-MCI and L-MCI samples using linear model SVM and logistic regression (LR) is not effective, even with various kernels. Therefore, we only adopt it as a feature selection method and compare it with the traditional one way Analysis of variance (ANOVA) feature selection technique.

We denote the dataset on the target task (L-MCI vs E-MCI) as $(\mathbf{X}^1, \mathbf{X}^2, \mathbf{y}^t)$. $\mathbf{X}^1, \mathbf{X}^2 \in \mathbb{R}^{460\times116}$ represent the FDG and VBM features respectively while

$\mathbf{y}^t \in \{-1, +1\}$ is the class label. We also construct three auxiliary domains $\{(\mathbf{A}_1^1, \mathbf{A}_1^2, \mathbf{y}_1^a), (\mathbf{A}_2^1, \mathbf{A}_2^2, \mathbf{y}_2^a), (\mathbf{A}_3^1, \mathbf{A}_3^2, \mathbf{y}_3^a)\}$. Each triplet in the bracket represents a task that may be helpful for feature selection. For instance, $(\mathbf{A}_2^1, \mathbf{A}_2^2, \mathbf{y}_2^a)$ denotes the FDG and VBM features along with labels for AD and NC patients. To construct a *multi-bit label coding matrix* for the transfer learning task, we firstly train three logistic regression models on three auxiliary domains. Then, we use these three models to independently estimate three labels for each patient on the target domain. Finally, we concatenate the true label with three predicted labels to form a multi-bit label for each patient and obtain a multi-bit label matrix $\mathbf{Y} = [\mathbf{y}^t, \mathbf{y}_1^p, \mathbf{y}_2^p, \mathbf{y}_3^p] \in \mathbb{R}^{460\times 4}$ (one true label, three predictions). The goal of the rMLTFL algorithms is to learn a weight matrix $\mathbf{W} = [\mathbf{w}^t, \mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3] \in \mathbb{R}^{116\times 4}$ which can be decomposed into two components $\mathbf{P}$ and $\mathbf{Q}$ for feature selection and domain identification respectively. Specifically, the objective function is formulated a following:

$$\min_{\mathbf{W},\mathbf{P},\mathbf{Q}} \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda_1 \|\mathbf{P}\|_{2,1} + \lambda_2 \left\|\mathbf{Q}^T\right\|_{2,1} + \lambda_3 \sum_{i=1}^{3} \left\|((\mathbf{X}\mathbf{w}^t - \mathbf{X}\mathbf{w}_i) - (\mathbf{y}^t - \mathbf{y}_i^p))\right\|_2^2, \quad s.t.\ \mathbf{W} = \mathbf{P} + \mathbf{Q}. \tag{1}$$

The first term is to ensure the similarity between the multi-bit labels $\mathbf{Y}$ and its prediction $\mathbf{XW}$. In the second and the third term, we use the 2,1 norm to capture the shared features cross all tasks and filter out the unrelated task. The 2,1 norm forces some rows of $\mathbf{P}$ and some columns of $\mathbf{Q}$ to be all zero. Non-zeros rows in $\mathbf{P}$ and non-zero columns in $\mathbf{Q}$ corresponds to informative features and tasks respectively. The last term indicates that the distance from predicted target domain label $\mathbf{X}\mathbf{w}^t$ to multi-bit label $\mathbf{X}\mathbf{w}_i^p$ should be similar to the distance from the true label $\mathbf{y}^t$ to the estimated multi-bit label $\mathbf{y}_i^p$.

The above rMLTFL framework to select feature can be illustrated in Figure 1. After we obtain the multi-bit label matrix $\mathbf{Y}$, we use the accelerate gradient decent algorithm to optimize the target function (1). Then, we filter out domains which correspond to all zero columns in $\mathbf{Q}$. After that, we repeat the same process as above without these useless domains. Finally, we select rows which correspond to non-zero rows in $\mathbf{P}$ as features related to the target task. When implementing rMLTFL and one-way ANOVA to select features, we apply each method to the two data modalities separately and simultaneously. Hence, we obtain six sets of sample features. After examining the prediction performance of these feature sets, we can choose the most relevant feature sets and achieve higher prediction accuracy by applying model aggregation techniques.

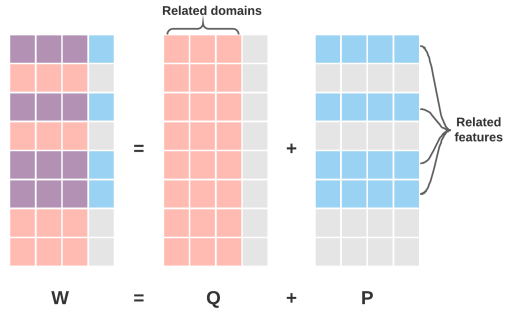

Figure 1 The learnable weighting matrix $\mathbf{W}$ can be decomposed into two matrices, $\mathbf{Q}$ and $\mathbf{P}$. They are responsible for selecting target problem related tasks (AD vs NC, AD vs MCI, MCI vs NC) and features. By enforcing the $l2 - l1$ norm of $\mathbf{Q}^T$ and $\mathbf{P}$ to be small, these group lasso penalty terms on rows on $\mathbf{P}$ and columns of $\mathbf{Q}$ encourage the rows of $\mathbf{P}$ and columns of $\mathbf{Q}$ in (1) to have all zero (rows and columns in grey) or non-zero elements. The first column of $\mathbf{Q}$ corresponds to the L-MCI vs E-MCI stratification task and the rest of them correspond to three auxiliary tasks. We could observe from the plot that the AD v NC and the AD vs MCI tasks are two related domains while the MCI vs NC task could not provide helpful information. Similarly, non-zero rows of $\mathbf{P}$ capture the shared features among useful domains.

Optimal transport for transfer learning

In previous work of MCI stage classification, i.e. classifying progressive MCI versus stable MCI [11] and MCI converters versus MCI non-converters [12], a common assumption is that introducing auxiliary tasks (ie. AD vs NC) can improve the accuracy of classification. It is assumed that at least some of these auxiliary domains can help us understand the target domain, even without feature transformation. From the t-distributed stochastic neighbor embedding (t-SNE), boxplot of principle components, and violin plot of features we conclude that the feature distribution of L-MCI and E-MCI is similar to the pattern of those in the AD and NC subjects. However, the difference between early and late state MCI is much more subtle than the difference between AD and NC samples. Therefore, we must adopt TL strategies to reduce the inter-task discrepancy between AD vs NC task and E-MCI vs L-MCI task while maximizing the intra-task differences. Traditional TL methods using sample weighting or feature alignment strategies to adapt source data samples (i.e. AD and NC samples) to the target domain (i.e. L-MCI and E-MCI samples)[13]. Compared with these previous works, the OT for TL frameworks can capture the intrinsic geometry structure difference of two feature spaces and address the distributional drift problem more efficiently. We illustrate in the our experiments that our proposed method based on OT outperforms the current state-of-the-art methods.

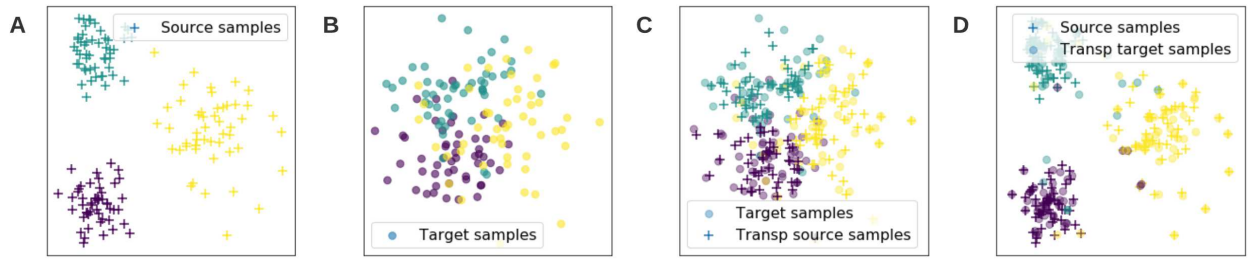

Figure 2 We use a synthetic Gaussian distributed dataset to demonstrate our method. In panel (**A**), we generate three clusters of guassian distributed samples. Their clusters are distinct, hence simple decision boundaries can separate them clearly. This example corresponds to the AD vs NC classification task. In panel (**B**), we also generate three clusters which are not distinctive from one another. In fact, the E-MCI and L-MCI clusters are much less distinct than the samples in panel (**B**). In panel (**C**), we use OT to map the source domain samples onto the target domain. In the last panel (**D**), we use our proposed method adopting OT to map target samples onto the source domain by utilizing sample labels.

OT maps the representations of one data domain to another by minimizing the earth moving distance [17, 14] between their distributions. To better understand the feature distribution within and across classes and to estimate a better transformation, [14, 15, 18] added different regularization terms such as L_1l_2 and L_pl_1 terms to achieve group sparsity. By adding the group sparsity regularization terms, the OT feature mapping strategy only project L-MCI training samples to the AD samples and E-MCI training samples to the NC samples. For computational efficiency, most of the state-of-the-art OT models incorporate an entropy regularization term. This regularized version of earth moving distance [19] is call Sinkhorn distance (SD). In this study, we implemented three OT mapping strategies defined by SD, SD with L_pl_1 regularization term, and SD with L_1l_2 regularization term respectively.

Before introducing the experiment setting of using OT to train classifiers, we want to emphasize the difference between our proposed method and traditional OT methods for TL that used as benchmarks in this study. Traditionally, the source domain features (AD and NC features) are mapped to the target domain (L-MCI vs E-MCI) via an OT strategy. Then, AD and NC labels as well as the transformed features can be used to train a classifier on the target domain that will be directly applied to the L-MCI vs E-MCI stratification task. This strategy is powerful when dealing when few labels are available on the target domain and the decision boundary for the target task is easy to learn. In our problem, the intrinsic difficulty is that the decision boundary is difficult to learn even after using kernel methods. Fortunately, we have plenty of samples (187 L-MCI, 273 E-MCI) on the target domain, which enable us to separate them into training and testing sets. Therefore we instead map training samples on the target domain (L-MCI vs E-MCI) to the source domain (AD vs NC) where the classification boundary is more clearly defined. During this process, we learn a non-linear OT mapping strategy $\mathbf{T}$. Then, we train classifiers to use AD and NC samples as well as E-MCI and L-MCI samples transformed by $\mathbf{T}$. After that, we use the OT mapping $\mathbf{T}$ to project testing samples to the source domain and use the classifier to stratify L-MCI and E-MCI samples. Finally, we evaluate the classification performance using accuracy and area under the receiver operating curve (AUC) score. Figure 2 illustrates the effects of using OT to obtain more distinguishable features in synthesized data.

In our experiments on real AD data, we investigate different OT mapping strategies as well as different classifiers on the source domain. In Figure 3, we illustrate how to adapt MCI samples onto the AD and NC domain. In Figure 5 panel (**A**), we demonstrate how to combine different OT mapping strategies in different learning tasks with different classifiers. Since logistic regression achieves higher prediction accuracy than SVM, we adopt it as a benchmark classifier and combine it with linear and polynomial kernel functions to form kernel based classifiers.

Bootstrap aggregation to improve model stability

Bootstrap aggregation (BAg) is an algorithm proposed in [20] for both regression and statistical classification. By randomly sampling training sets (bootstrapping) with replacement, one can train several classifiers using the same algorithm. By aggregating model predictions based on majority voting strategies or aggregated prediction probabilities, we raise the stability of our models by reducing inter-model variability from overfitting. When we implement the BAg strategy, we firstly need to decide the number of "bags" to use. Since our study

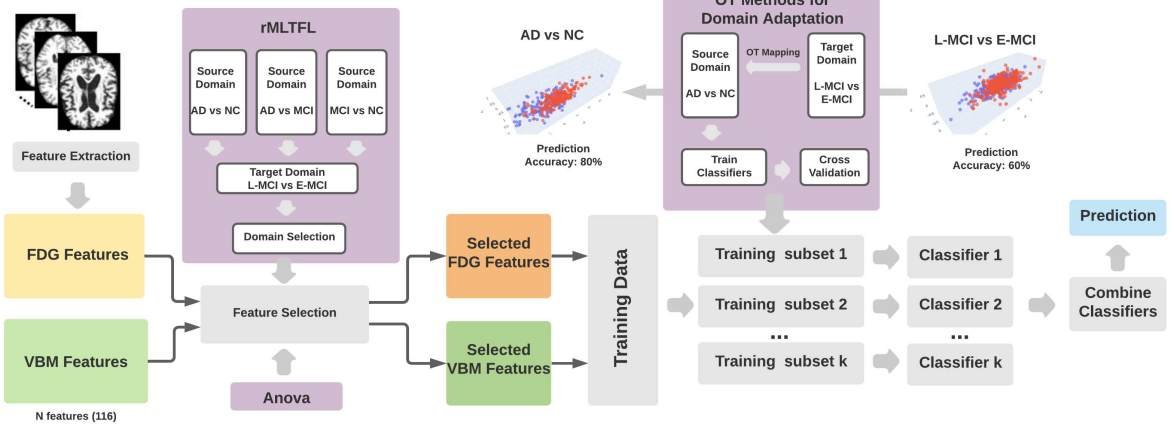

Figure 3 Our novel OT TL framework and pipeline used to train our models. Using the pre-processing workflow, we extract VBM and FDG features from the manually labeled regions-of-interest (ROIs) in MRI images. Then, we use the the rMLTFL framework as well as one-way ANOVA to select features from two modalities (FDG and VBM) both separately and simultaneously. We then separate the target dataset into training (80%) and testing (20%) sets. After that, we sample subsets of training samples, use the regularized OT to mapping selected samples on the AD vs NC dataset, and train classifiers using AD, NC, and transformed samples. Finally, we aggregate these models to form a robust BAg model and make predictions on OT transformed testing samples.

only contains a few hundreds samples, We use 5 bags to train five sub-models. We then separate the dataset into training and testing sets (80% and 20%). On the training set, we implement the Bootstrap strategy in a slightly different manner. During the stage of Bootstrapping, we randomly split the training set into five folds and pick four folds each time to train a classifier using our OT TL strategy. Then, we aggregate the model using a majority vote strategy. The prediction probability is obtained by calculating the mean prediction probability across each sub-model. We illustrate the pipeline in Figure 3. To demonstrate that our OT alignment improves the stratification performance, we also compare our method with different versions of BAg versions using traditional SVM, logistic regression, and rMLTFL models.

Results

Data visualization

Using one-way ANOVA, we calculated the p-value for each features individually. Using the p-value threshold 0.05, we selected 47 out of 116 features from the FDG and the VBM domain respectively. The rMLTFL method captures features by training a model and selecting features based on that trained model. We need to verify the stability of this feature selection procedure. To determine which hyper-parameters to use and whether the collection of useful features were dependent on the training set, we used five-fold cross validation to verify the robustness of the rMLTFL method. We took a grid search approach for the three hyper-parameters over a 1,000 combinations of these parameters and chose the hyper-parameter combination with the highest average prediction accuracy. Using the optimal hyper-parameters, we ran the rMLTFL algorithm on the FDG data modality to filter out useless features and obtained 96 features by merging selected training sets respectively across five folds. For the VBM data modality, the model only filtered out one useless features over all hyper-parameter combinations. Therefore, we kept 115 features from the VBM data modality. To combine the two data modalities, we concatenated the two feature vectors and repeated the same process as described above. We visualized the selected FDG features selected in Figure 4. Panels (**A**) and (**B**) show the t-SNE plots of features selected by ANOVA and rMLTFL respectively. In panel (**A**), we observe that AD patients mainly concentrated on the upper right corner where L-MCI patient is also denser than other areas while E-MCL and NC samples are denser at the lower left corner. We concluded that the pattern of AD vs NC may help us delineate the distributions of L-MCI versus E-MCI. The same pattern can be observed in panel (**B**). Panel (**C**) and (**E**) illustrate distributions of first two principle components of ANOVA and rMLTFL features. From these plots we concluded that the distributional differences between the first principle components of L-MCI and E-MCI patients are more subtle than the differences between AD and NC patients. AD and L-MCI patients tended to have lower PC 1 while E-MCI and NC tend to have higher values of PC 1. We also visualized part

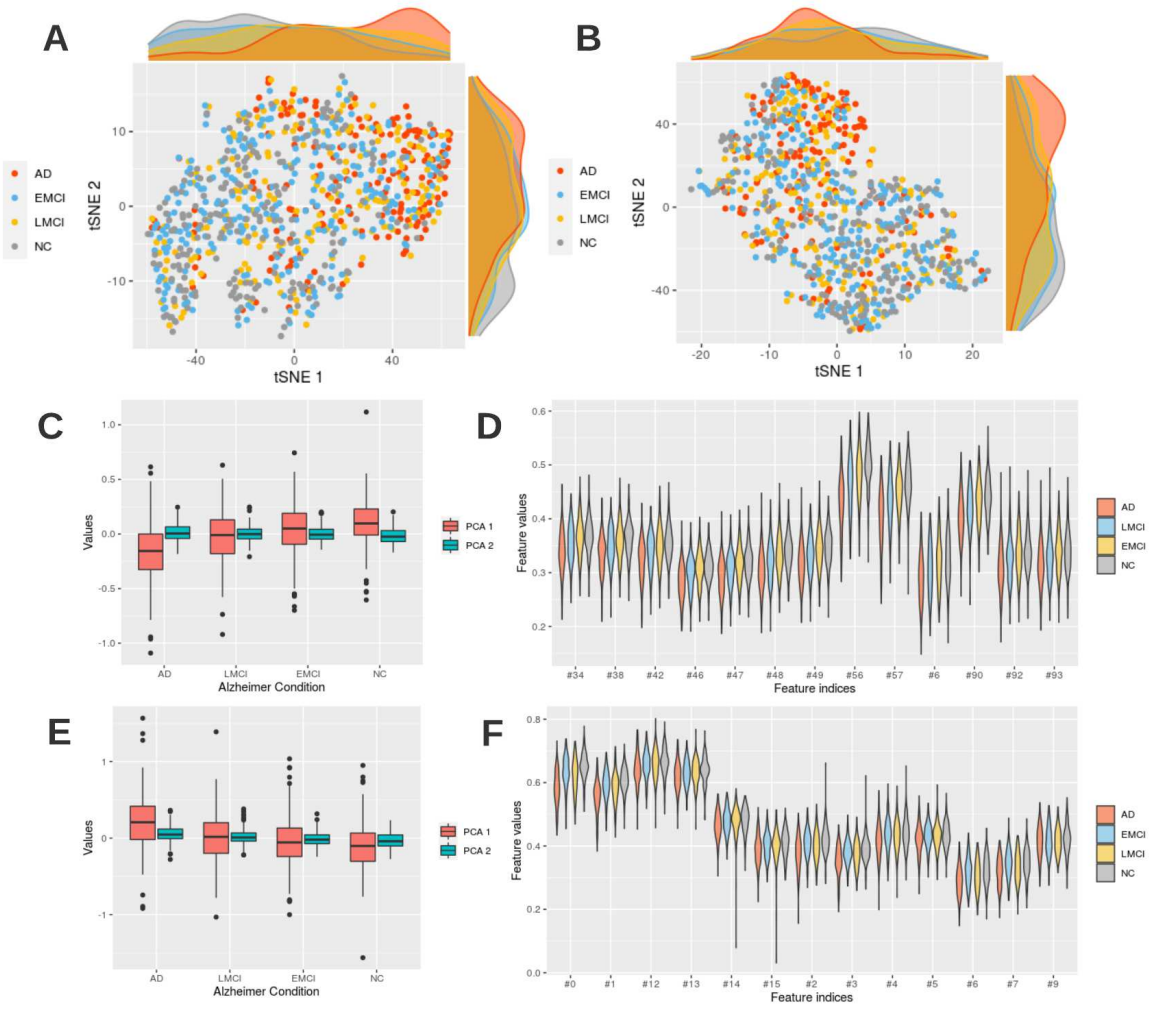

Figure 4 (**A**) and (**B**) represent t-SNE plots and their marginal distributions for **FDG** features selected by ANOVA and rMLTFL respectively. (**C**) and (**E**) are box-plots for first two principle components of these selected features. We also visualize the distribution of part of ANOVA (**D**) and rMLTFL features (**F**) using violin plots.

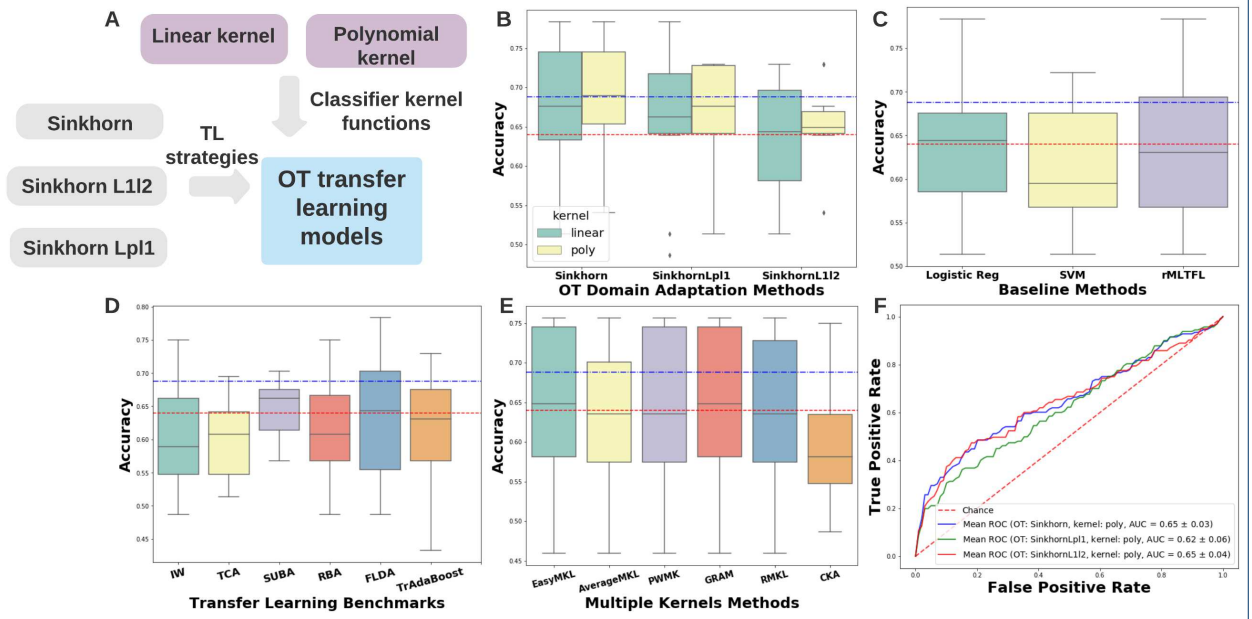

Figure 5 Results of ten folds cross validation using our method and other benchmark methods on FDG features. Panel (**A**) is the working pipeline of our OT TL model. We combine linear and polynomial kernelized logistic regression classifier with different OT mapping strategies. In (**B**), we represent the accuracy score of different OT and kernel combinations. The blue and red horizontal lines represent the average accuracy of our best model and the logistic regression model respectively. In panel (**C**) we demonstrate the performance of two baseline methods, e.g. logistic regression and SVM, and the rMLTFL model. In (**D**) and (**E**), we visualize the performance of TL benchmarks and Multi-kernel learning strategies. In (**F**), we plot the AUC curve of our model across ten folds.

of features selected by ANOVA and rMLTFL in (**D**) and (**F**). From them we observed the same pattern as the boxplots.

Results and benchmark studies

We firstly applied our framework on each data modality individually. Then, for each data modality, we applied three different OT mapping strategies: OT estimated by Sinkhorn distance(SD), SD regularized by $L_p l_1$ norm and SD regularized by $L_1 l_2$ norm. The usage of these regularization norms is to enforce intra-class similarity. When we mapped L-MCI and E-MCI samples to the domain of AD and NC samples, we utilized the labels of training samples. The results of cross validation in Table 2 and 3 demonstrate that our framework outperformed all baseline methods and the original rMLTFL model. Based on FDG features, our model achieved $68.76 \pm 7.53\%$ accuracy and 0.66 ± 0.08 AUC score across ten folds cross validation. The SVM and logistic regression baseline methods achieved $61.20 \pm 7.22\%$ and $64.40 \pm 7.60\%$ accuracy respectively. Our model also outperformed them in the VBM data modality. Comparing the performance of features selected by rMLTFL and ANOVA we observed that the rMLTFL features are always superior than ANOVA features. This indicates that even features that are not significant statistically may be helpful to model complex nonlinear differences between sample classes. Combining two data modalities by directly concatenating features did not help us in distinguishing L-MCI and E-MCI patients.

Besides two baseline methods and the rMLTFL framework, we also compared our model with other TL benchmarks and multiple kernel learning strategies. For TL benchmarks, we compared our method with: Importance-weighting with logistic discrimination (IW) [21], Transfer Component Analysis (TCA) [22], Semi-supervised Subspace Alignment (SUBA) [23], Feature-Level Domain Adaptation (FLDA) [24], and Boosting for Transfer learning (TrAdaBoost)[25]. We also compared with multiple kernel learning strategies including: the simple average of base kernels (AverageMKL), margin-based combination of kernels (EasyMKL) [26], radius-margin ratio optimization for dot-product boolean kernel learning (GRAM) [27], margin and radius based multiple kernel learning (RMKL) [28], simple but effective methods for combining kernels in computational biology (PWMK) [29], and centered kernel alignment optimization in closed form (CKA) [30]. Based on these comparisons, our method proved superior to all of these benchmarks (Table 3, Figure 5). One notable fact is that most of

Table 2 Accuracy (ACC) and AUC score of models based on features selected by rMLTFL and ANOVA (p-value threshold=0.05) respectively. The values are denoted as mean±standard deviation. We investigated different OT mapping strategies, e.g. using Sinkhorn distance, Sinkhorn Distance with $L_p l_1$ regularization term, and Sinkhorn distance with $L_p l_1$ regularization term. Accuracy and AUC scores are calculated by averaging over performances of ten folds cross validation on the training set. We exam the model performance on FDG and VBM features separately and simultaneously.

		Sinkhorn Distance		Sinkhorn Distance + $L_p l_1$		Sinkhorn Distance + $L_1 l_2$	
		ACC	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**
FDG	**rMLTFL**	68.76 ± 7.53	0.66 ± 0.08	66.04 ± 7.53	0.65 ± 0.08	65.48 ± 5.04	0.64 ± 0.07
	ANOVA	66.07 ± 6.96	0.65 ± 0.07	63.63 ± 6.01	0.64 ± 0.07	59.50 ± 6.53	0.62 ± 0.08
VBM	**rMLTFL**	62.37 ± 6.88	0.62 ± 0.11	62.74 ± 0.08	0.62 ± 0.11	57.86 ± 6.32	0.60 ± 0.07
	ANOVA	58.94 ± 7.82	0.59 ± 0.12	58.68 ± 0.08	0.58 ± 0.12	56.79 ± 0.11	0.577 ± 0.13
FDG + VBM	**rMLTFL**	62.26 ± 6.48	0.63 ± 0.05	66.61 ± 6.29	0.65 ± 0.06	66.05 ± 5.91	65.30 ± 0.08
	ANOVA	61.44 ± 6.23	0.64 ± 0.05	63.87 ± 5.75	0.63 ± 0.06	61.15 ± 8.01	0.64 ± 0.07

Table 3 Accuracy of baseline, transfer learning and Multi-kernel benchmark methods. The values are denoted as mean±standard deviation.

Methods	**FDG**	**VBM**
SVM	61.20 ± 7.22	57.64 ± 5.89
Logistic Reg	64.40 ± 7.60	58.72 ± 6.98
rMLTFL	63.33 ± 9.02	62.53 ± 9.08
IW	60.10 ± 8.41	59.56 ± 7.49
TCA	59.83 ± 6.02	57.02 ± 8.27
SUBA	64.68 ± 4.34	52.44 ± 8.33
RBA	61.46 ± 8.21	58.17 ± 8.02
FLDA	63.90 ± 10.00	60.11 ± 9.05
TrAdaBoost	61.45 ± 8.56	59.98 ± 4.73
Easy MKL	64.72 ± 9.75	60.38 ± 7.46
Average MKL	63.34 ± 9.08	60.11 ± 7.14
PWMK	64.19 ± 9.80	60.11 ± 7.14
GRAM	64.72 ± 9.75	/
RMKL	63.91 ± 9.53	60.11 ± 7.14
CKA	59.56 ± 7.49	59.56 ± 7.49

Table 4 Accuracy (ACC) and AUC score of BAg results. The OT method and kernel function combination is denoted as OT method/kernel function. l and p represent linear and polynomial kernel respectively.

Methods	**FDG**		**VBM**		**FDG + VBM**	
	ACC	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**
SD/l	72.82	0.79	68.48	0.67	70.65	0.74
SD/p	73.91	0.76	64.13	0.67	69.56	0.71
SD $L_1 l_2$/l	**75.00**	**0.77**	67.39	0.67	69.56	0.77
SD $L_1 l_2$/p	71.74	0.76	63.04	0.67	65.21	0.70
SD $L_p l_1$/l	71.74	0.76	59.78	0.68	73.91	0.76
SD $L_p l_1$/p	71.74	0.74	67.39	0.69	66.30	0.71
SVM	57.61	0.47	57.61	0.29	57.61	0.55
logistic	68.47	0.67	58.70	0.70	67.39	0.66
rMLTFL	63.04	0.69	60.87	0.70	63.04	0.67

them did not even beat the baseline method logistic regression with linear kernel function. Therefore, traditional TL techniques such as sample weighting and feature alignment strategies may not be effective for us to delineate the distribution patterns of L-MCI and E-MCI. Since our method compares distributions directly, we can glean more information from AD and NC patients as well as MCI patients in the training set. We also found that Easy MKL, average KL, and PWMK methods yielded relatively high performance on both domains. We conclude that combining multiple kernel functions in an appropriate manner can improve the classification performance.

Bootstrap Aggregation result

In Table 4, we list the aggregated model performance of the testing set for different models and different data modalities. Besides our OT mapping strategies, we also implemented the BAg using two baseline methods and the rMLTFL benchmark method. The performance of our model was significantly superior than SVM, logistic regression, and rMLTFL (Figure 6 and 7). By choosing different training sets, our model captured heterogeneous patterns. When we aggregated them using a voting strategy, most models could correctly prediction the testing samples. Hence, the accuracy as well as AUC score was much higher than the single model case. On the other hand, the logistic regression, SVM, and rMLTFL models were quite stable with regard to the training set (Figure 6 and 7). The patterns they learned are quite homogeneous. We conclude that learning sub-models does not improve model performance for these baseline and benchmark methods.

Discussion

We present our novel method which uses optimal transport to improve the performance discriminating between early and late stage MCI (E-MCI vs L-MCI) using MRI images. We found that by using OT theory to project the more difficult task, E-MCI vs L-MCI, onto the easier task of distinguishing AD and NC, we were able to achieve higher performance than by using MCI samples alone. This represents not only a significant advance in OT and TL methods but also has clear clinical implications.

Indeed, identifying cognitively impaired individuals early will likely their health outcomes because of early access to treatment and monitoring [31, 32]. These early detection systems are most frequently focused

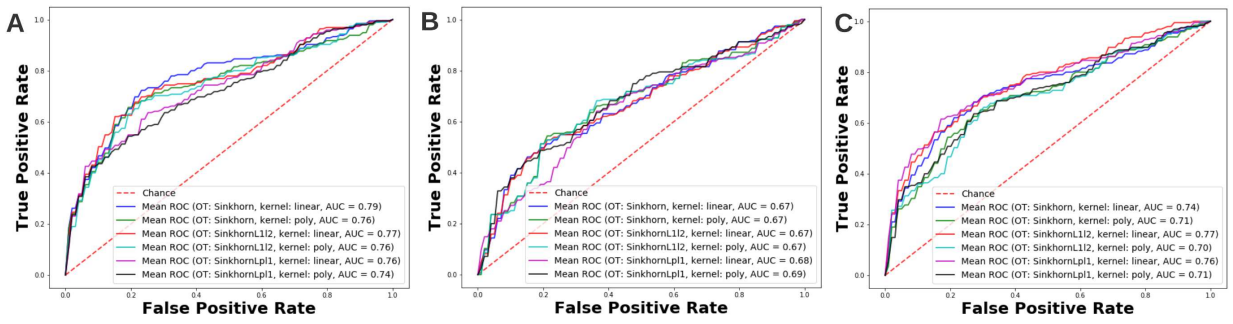

Figure 6 AUC curves of BAg of our OT transfer learning framework. Panel (**A**), (**B**), and (**C**) correspond to results on the **FDG**, **VBM**, and combination of two data modalities respectively. The highest AUC curve is achieved by using SD $L_1 l_2$ as OT mapping cost function and linear kernel logistic regression as classifier.

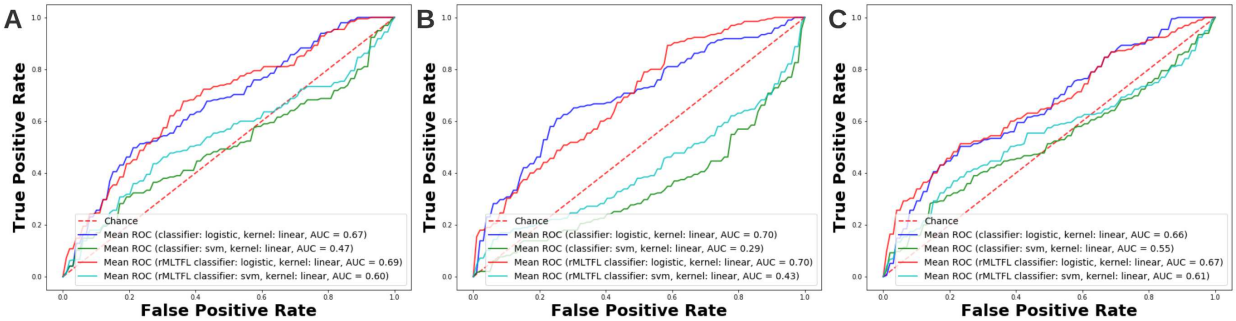

Figure 7 AUC curves of BAg of logistic regression, SVM, and rMLTFL (logsitic regression and SVM as classifier). Panel (**A**), (**B**), and (**C**) correspond to results on the **FDG**, **VBM**, and combination of two data modalities respectively.

on the readily available and minimally-invasive medical imaging procedures like MRI and PET scans. Ideally, at risk patients could regularly be tested for AD and it's precursors like MCI by their physicians. These imaging technologies offer a potential avenue to a minimally invasive test for cognitive impairment. These clinical tests however are dependent on accurate ML models which can effectively discriminate between cognitively normal, end stage Alzheimers, and the entire spectrum in between.

By using OT to map E-MCI and L-MCI samples to the auxiliary domain, we reduce the inter-task discrepancy between AD vs NC task and E-MCI vs L-MCI task while maximizing the intra-task differences. This TL technique enable us to train LR classifiers which can stratify E-MCI and L-MCI patients more accurately. We then aggregate sub-TL models using a majority voting strategy to improve the model stability and avoid the over-fitting issue.

With the novel methods that we have developed, we outperform the current state-of-the-art TL methods and show that it is crucial to leverage AD and NC data to accurately predict early and late stage cognitive impairment. Such continued improvements are necessary to improve the personal, healthcare, and economic costs [33] associated with over six million AD patients in the United States alone.

Limitations

When compared with other benchmark works, our model yields a high prediction accuracy and AUC score. We also acknowledge several limitations. Our model selection method rMLTFL depends on three hyper-parameters. It's of crucial importance to select correct combination hyper-parameters. Although we grid search them over 1000 combinations, there is still lack of evidence that the selected combination is an optimal choice. Furthermore, we have not considered its performance in other challenging MCI classification tasks such as the P-MCI and S-MCI classification task [11].

Conclusion

We have developed an optimal transport based transfer learning model to discriminate between early and late mild cognitive impairment. Our methods are both novel and the current state of the art based on benchmark comparisons. This method is a necessary technological stepping stone to widespread clinical usage of MRI based early detection of AD.

Appendix

Acknowledgements
Not applicable.

Funding
This work is partially supported by Indiana University Precision Health Initiative (to LZ, TJ) and NIH U54AG065181 grant (to KH, ZJ, WS).

Abbreviations
AD: Alzheimer's disease; MCI: Mild cognitive impairment; NC: Normal control; L-MCI: Late stage mild cognitive impairment; E-MCI; early stage mild cognitive impairment; CT: computed tomography; MRI: magnetic resonance imaging; PET: positron emission tomography; VBM: voxel based measure; FDG: fluorodeoxyglucose TL: transfer learning; ADNI: The Alzheimer's Disease Neuroimaging Initiative; MMSE: the mini-mental state examination; CDR: the clinical dementia rating; BAg: Bootstrap aggregation; OT: optimal transport; rMLTFL: multi-label transfer feature learning; SVM: support vector machine; LR: logistic regression; ANOVA: Analysis of variance; t-SNE: t-distributed stochastic neighbor embedding; SD: Sinkhorn distance; AUC: area under the receiver operating curve; IW: Importance-weighting; TCA: Transfer component analysis; SUBA: Semi-supervised subspace alignment; FLDA: Feature-level domain adaptation; TrAdaBoost: Boosting for transfer learning; MKL: Multiple kernel learning; AverageMKL: simple average of base kernels; EasyMKL: radius based combination of kernels; GRAM: radius-margin ratio optimization for dot-product boolean kernel learning; RMKL: radius based multiple kernel learning; PWMK: simple but effective methods for combining kernels in computational biology; CKA: centered kernel alignment optimization in closed form.

Availability of data and materials
The dataset(s) supporting the conclusions of this article is(are) included within the article (and its additional file(s).

Ethics approval and consent to participate
Not applicable.

Competing interests
There is no conflict of interest.

Consent for publication
Not applicable.

Authors' contributions
Study design: Travis S. Johnson, Ziyu Liu, Shao We, Jie Zhang and Kun Huang; data cleaning and pre-processing: Wei Shao; modeling and computational methods: Ziyu Liu and Wei Shao; paper writing: Travis S. Johnson, Ziyu Liu, and Kun Huang; paper review and supervision: Min Zhang, Jie Zhang, and Kun Huang. The authors read and approved the final manuscript.

Author details
[1]Department of Statistics, Purdue University, West Lafayette, USA. [2]Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, USA. [3]Regenstrief Institute, Indianapolis, USA. [4]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, USA.

References

1. Alzheimer's Disease Fact Sheet. U.S. Department of Health and Human Services. `https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet`
2. Niikura, T., Tajima, H., Kita, Y.: Neuronal cell death in alzheimer's disease and a neuroprotective factor, humanin. Current neuropharmacology **4**(2), 139–147 (2006)
3. Kinney, J.W., Bemiller, S.M., Murtishaw, A.S., Leisgang, A.M., Salazar, A.M., Lamb, B.T.: Inflammation as a central mechanism in alzheimer's disease. Alzheimer's & Dementia: Translational Research & Clinical Interventions **4**, 575–590 (2018)
4. Murphy, M.P., LeVine III, H.: Alzheimer's disease and the amyloid-β peptide. Journal of Alzheimer's disease **19**(1), 311–323 (2010)
5. Park, J.-C., Han, S.-H., Yi, D., Byun, M.S., Lee, J.H., Jang, S., Ko, K., Jeon, S.Y., Lee, Y.-S., Kim, Y.K., *et al.*: Plasma tau/amyloid-β1–42 ratio predicts brain tau deposition and neurodegeneration in alzheimer's disease. Brain **142**(3), 771–786 (2019)
6. Mattson, M.P.: Pathways towards and away from alzheimer's disease. Nature **430**(7000), 631–639 (2004)
7. How Is Alzheimer's Disease Diagnosed? U.S. Department of Health and Human Services. `https://www.nia.nih.gov/health/how-alzheimers-disease-diagnosed`
8. Ahmed, O.B., Benois-Pineau, J., Allard, M., Catheline, G., Amar, C.B., Initiative, A.D.N., *et al.*: Recognition of alzheimer's disease and mild cognitive impairment with multimodal image-derived biomarkers and multiple kernel learning. Neurocomputing **220**, 98–110 (2017)
9. Cohen, A.D., Klunk, W.E.: Early detection of alzheimer's disease using pib and fdg pet. Neurobiology of disease **72**, 117–122 (2014)
10. Zhang, Y., Dong, Z., Phillips, P., Wang, S., Ji, G., Yang, J., Yuan, T.-F.: Detection of subjects and brain regions related to alzheimer's disease using 3d mri scans based on eigenbrain and machine learning. Frontiers in computational neuroscience **9**, 66 (2015)
11. Cheng, B., Liu, M., Zhang, D., Shen, D.: Robust multi-label transfer feature learning for early diagnosis of alzheimer's disease. Brain imaging and behavior **13**(1), 138–153 (2019)
12. Cheng, B., Liu, M., Suk, H.-I., Shen, D., Zhang, D.: Multimodal manifold-regularized transfer learning for mci conversion prediction. Brain imaging and behavior **9**(4), 913–926 (2015)
13. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE **109**(1), 43–76 (2020)
14. Flamary, R., Courty, N., Rakotomamonjy, A., Tuia, D.: Optimal transport with laplacian regularization. In: NIPS 2014, Workshop on Optimal Transport and Machine Learning (2014)
15. Courty, N., Flamary, R., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. IEEE transactions on pattern analysis and machine intelligence **39**(9), 1853–1865 (2016)
16. Hao, X., Bao, Y., Guo, Y., Yu, M., Zhang, D., Risacher, S.L., Saykin, A.J., Yao, X., Shen, L., Initiative, A.D.N., *et al.*: Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of alzheimer's disease. Medical image analysis **60**, 101625 (2020)
17. Levina, E., Bickel, P.: The earth mover's distance is the mallows distance: Some insights from statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, vol. 2, pp. 251–256 (2001). IEEE
18. Perrot, M., Courty, N., Flamary, R., Habrard, A.: Mapping estimation for discrete optimal transport. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 4204–4212 (2016)
19. Peleg, S., Werman, M., Rom, H.: A unified approach to the change of resolution: Space and gray-level. IEEE Transactions on Pattern Analysis and Machine Intelligence **11**(7), 739–742 (1989)
20. Breiman, L.: Bagging predictors. Machine learning **24**(2), 123–140 (1996)
21. Bickel, S., Brückner, M., Scheffer, T.: Discriminative learning under covariate shift. Journal of Machine Learning Research **10**(9) (2009)
22. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks **22**(2), 199–210 (2010)
23. Yao, T., Pan, Y., Ngo, C.-W., Li, H., Mei, T.: Semi-supervised domain

adaptation with subspace learning for visual recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2142–2150 (2015)
24. Kouw, W.M., Van Der Maaten, L.J., Krijthe, J.H., Loog, M.: Feature-level domain adaptation. The Journal of Machine Learning Research **17**(1), 5943–5974 (2016)
25. Yao, Y., Doretto, G.: Boosting for transfer learning with multiple sources. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1855–1862 (2010). IEEE
26. Aiolli, F., Donini, M.: Easymkl: a scalable multiple kernel learning algorithm. Neurocomputing **169**, 215–224 (2015)
27. Lauriola, I., Polato, M., Aiolli, F.: Radius-margin ratio optimization for dot-product boolean kernel learning. In: International Conference on Artificial Neural Networks, pp. 183–191 (2017). Springer
28. Do, H., Kalousis, A., Woznica, A., Hilario, M.: Margin and radius based multiple kernel learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 330–343 (2009). Springer
29. Tanabe, H., Ho, T.B., Nguyen, C.H., Kawasaki, S.: Simple but effective methods for combining kernels in computational biology. In: 2008 IEEE International Conference on Research, Innovation and Vision for the Future in Computing and Communication Technologies, pp. 71–78 (2008). IEEE
30. Cortes, C., Mohri, M., Rostamizadeh, A.: Two-stage learning kernel algorithms (2010)
31. Sabbagh, M.N., Boada, M., Borson, S., Chilukuri, M., Doraiswamy, P., Dubois, B., Ingram, J., Iwata, A., Porsteinsson, A., Possin, K., *et al.*: Rationale for early diagnosis of mild cognitive impairment (mci) supported by emerging digital technologies. The Journal of Prevention of Alzheimer's Disease **7**, 158–164 (2020)
32. Rasmussen, J., Langerman, H.: Alzheimer's disease–why we need early diagnosis. Degenerative neurological and neuromuscular disease **9**, 123 (2019)
33. Wong, W.: Economic burden of alzheimer disease and managed care considerations. The American Journal of Managed Care **26**(8 Suppl), 177–183 (2020)

Additional Files

ADdata_FDG.csv
This csv file contains the pre-processed FDG features from [16]. Label 1, 3, 4, 5 correspond to NC, E-MCI, L-MCI and AD subjects respectively.

ADdata_VBM.csv
This csv file contains the pre-processed VBM features from [16]. Label 1, 3, 4, 5 correspond to NC, E-MCI, L-MCI and AD subjects respectively.

Figures

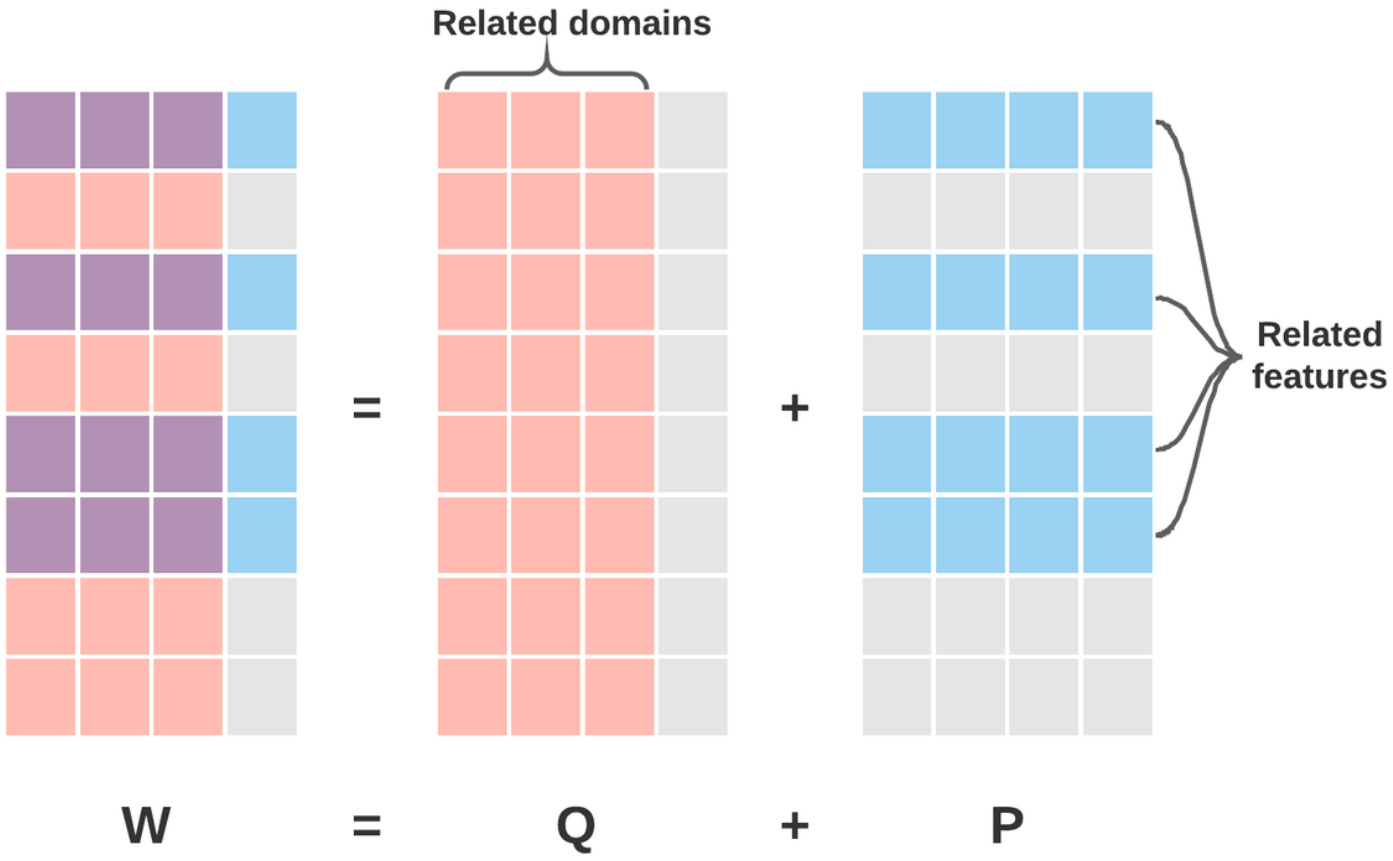

Figure 1

The learnable weighting matrix W can be decomposed into two matrices, Q and P. They are responsible for selecting target problem related tasks (AD vs NC, AD vs MCI, MCI vs NC) and features. By enforcing the l2 - l1 norm of QT and P to be small, these group lasso penalty terms on rows on P and columns of Q encourage the rows of P and columns of Q in (1) to have all zero (rows and columns in grey) or non-zero elements. The rst column of Q corresponds to the L-MCI vs E-MCI stratication task and the rest of them correspond to three auxiliary tasks. We could observe from the plot that the AD v NC and the AD vs MCI tasks are two related domains while the MCI vs NC task could not provide helpful information. Similarly, non-zero rows of P capture the shared features among useful domains.

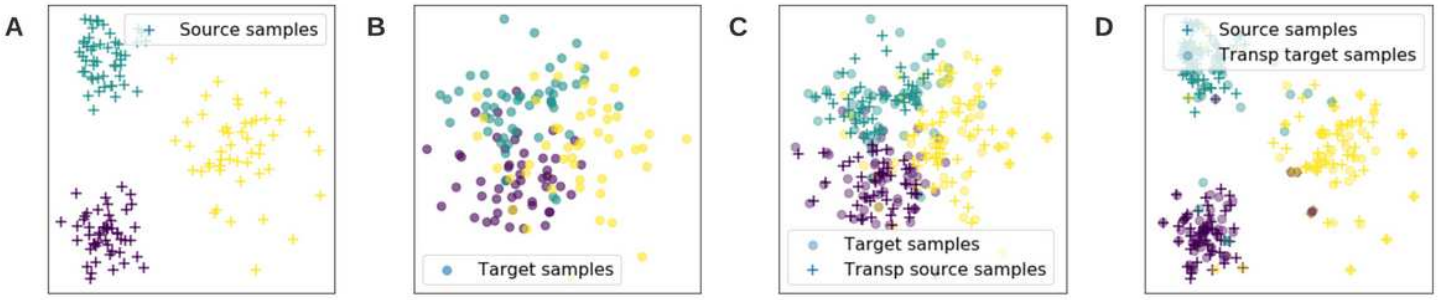

Figure 2

We use a synthetic Gaussian distributed dataset to demonstrate our method. In panel (A), we generate three clusters of guassian distributed samples. Their clusters are distinct, hence simple decision boundaries can separate them clearly. This example corresponds to the AD vs NC classication task. In panel (B), we also generate three clusters which are not distinctive from one another. In fact, the E-MCI and L-MCI clusters are much less distinct than the samples in panel (B). In panel (C), we use OT to map the source domain samples onto the target domain. In the last panel (D), we use our proposed method adopting OT to map target samples onto the source domain by utilizing sample labels.

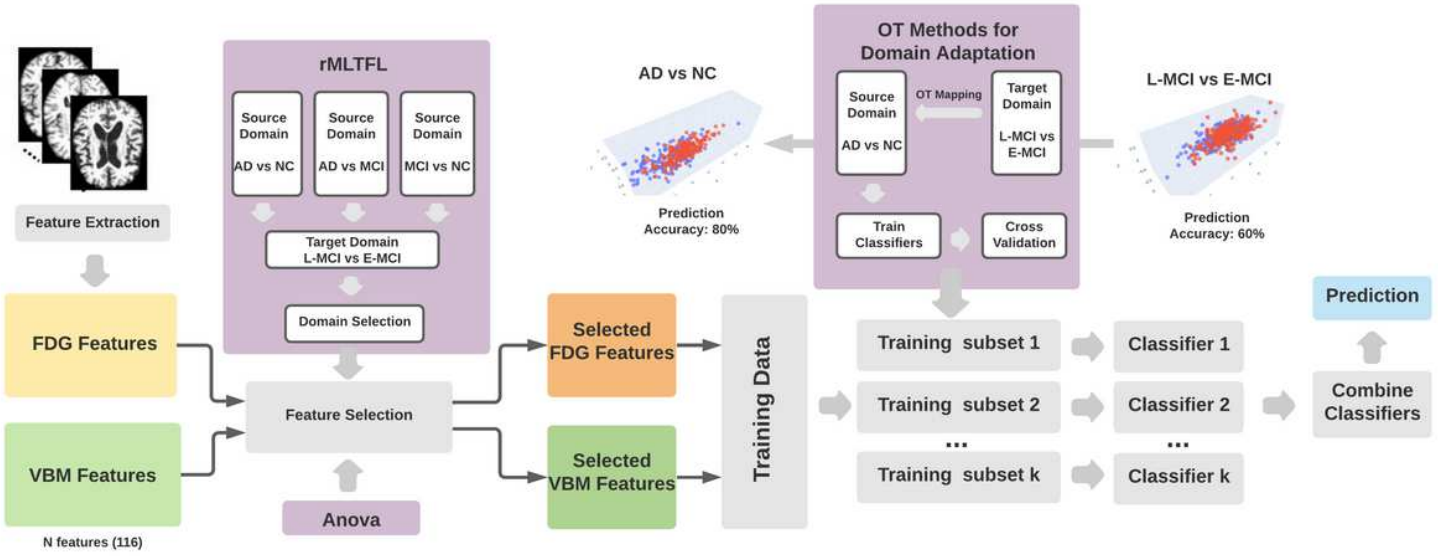

Figure 3

Our novel OT TL framework and pipeline used to train our models. Using the pre-processing work ow, we extract VBM and FDG features from the manually labeled regions-of-interest (ROIs) in MRI images. Then, we use the the rMLTFL framework as well as one-way ANOVA to select features from two modalities (FDG and VBM) both separately and simultaneously. We then separate the target dataset into training (80%) and testing (20%) sets. After that, we sample subsets of training samples, use the regularized OT to mapping selected samples on the AD vs NC dataset, and train classiers using AD, NC, and transformed samples. Finally, we aggregate these models to form a robust BAg model and make predictions on OT transformed testing samples.

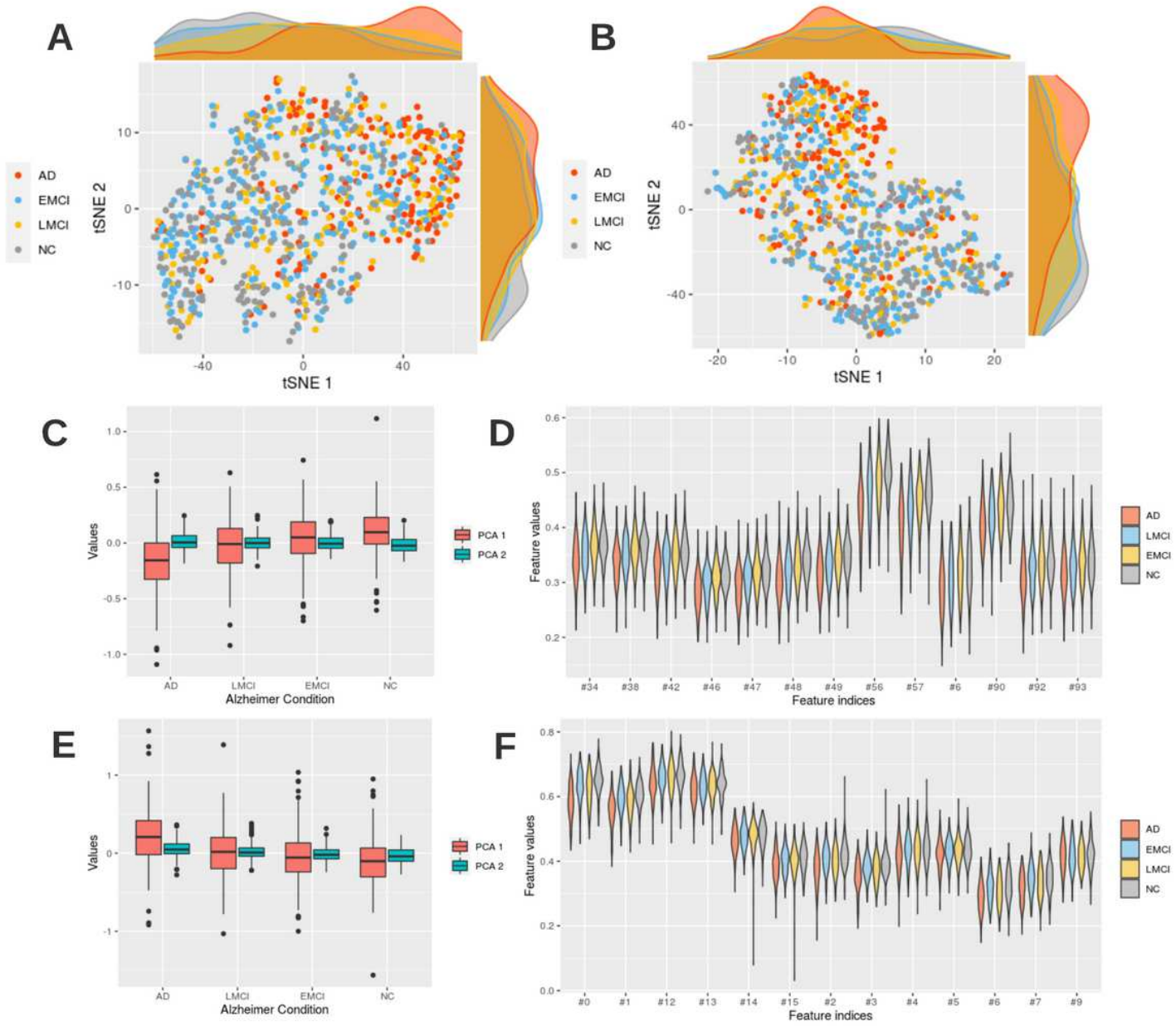

Figure 4

(A) and (B) represent t-SNE plots and their marginal distributions for FDG features selected by ANOVA and rMLTFL respectively. (C) and (E) are box-plots for rst two principle components of these selected features. We also visualize the distribution of part of ANOVA (D) and rMLTFL features (F) using violin plots.

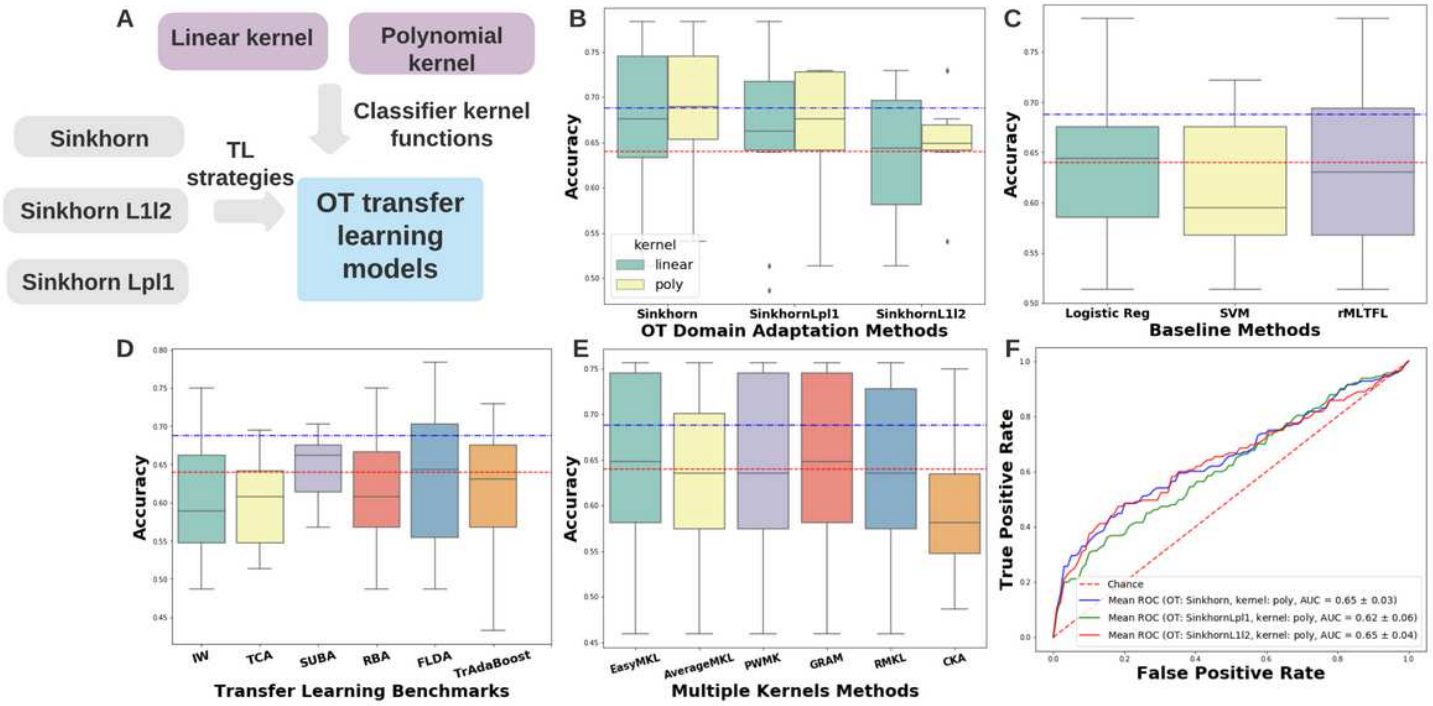

Figure 5

Results of ten folds cross validation using our method and other benchmark methods on FDG features. Panel (A) is the working pipeline of our OT TL model. We combine linear and polynomial kernelized logistic regression classier with di erent OT mapping strategies. In (B), we represent the accuracy score of different OT and kernel combinations. The blue and red horizontal lines represent the average accuracy of our best model and the logistic regression model respectively. In panel (C) we demonstrate the performance of two baseline methods, e.g. logistic regression and SVM, and the rMLTFL model. In (D) and (E), we visualize the performance of TL benchmarks and Multi-kernel learning strategies. In (F), we plot the AUC curve of our model across ten folds.

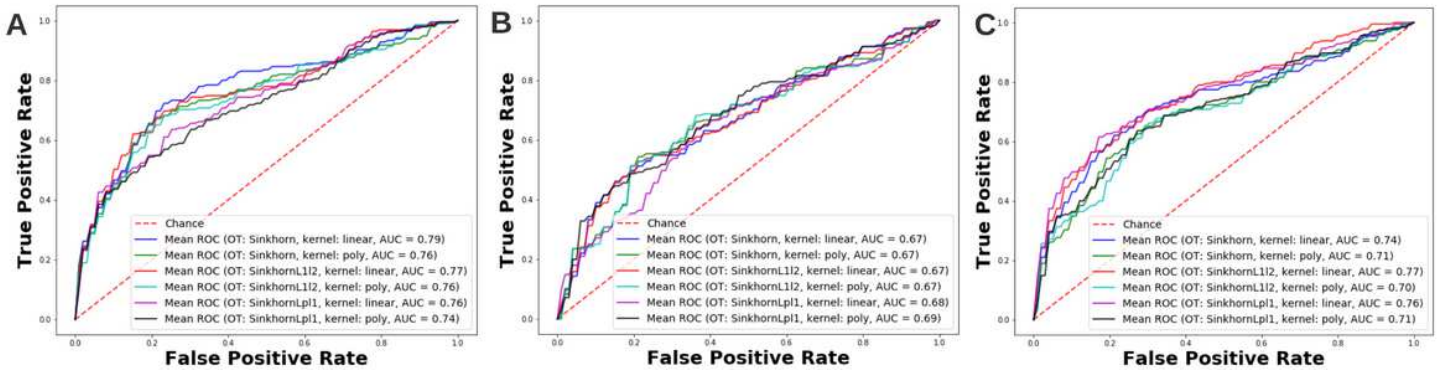

Figure 6

AUC curves of BAg of our OT transfer learning framework. Panel (A), (B), and (C) correspond to results on the FDG, VBM, and combination of two data modalities respectively. The highest AUC curve is achieved

by using SD L1l2 as OT mapping cost function and linear kernel logistic regression as classier.

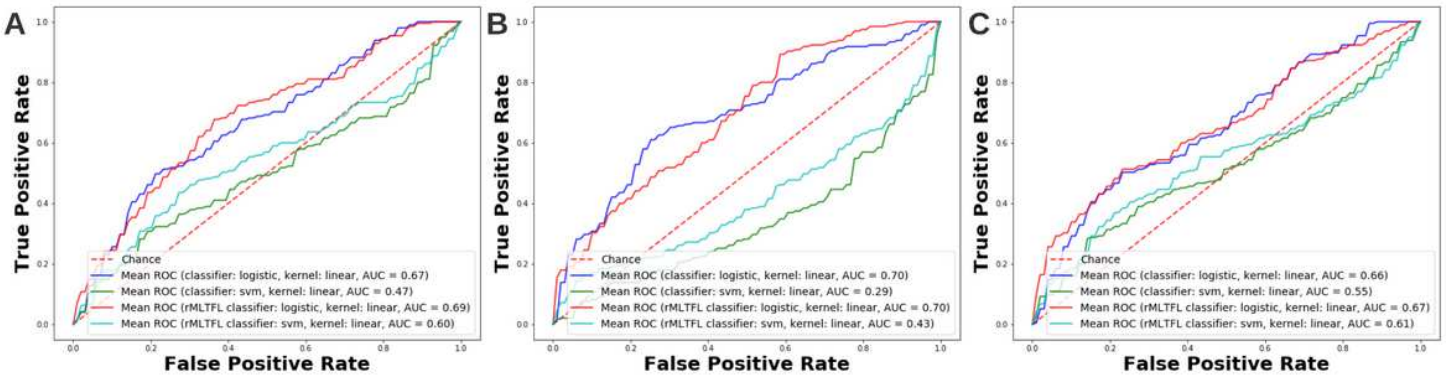

Figure 7

AUC curves of BAg of logistic regression, SVM, and rMLTFL (logsitic regression and SVM as classier). Panel (A), (B), and (C) correspond to results on the FDG, VBM, and combination of two data modalities respectively.