



Randomized Multi-task Feature Learning Approach for Modelling and Predicting Alzheimer's Disease Progression

Xulong Wang¹, Yu Zhang¹, Menghui Zhou¹, Tong Liu¹, Zhipeng Yuan¹,
Xiyang Peng¹, Kang Liu¹, Jun Qi², and Po Yang¹(✉)

¹ Department of Computer Science, University of Sheffield, Sheffield, UK
{x.l.wang, yzhang489, mzhou47, zhipeng.yuan, xpeng24,
kang.liu, po.yang}@sheffield.ac.uk

² Department of Computing, Xi'an JiaoTong-Liverpool University, Suzhou, China
Jun.Qi@xjtlu.edu.cn

Abstract. Multi-task feature learning (MTFL) methods play a key role in predicting Alzheimer's disease (AD) progression. These studies adhere to a unified feature-sharing framework to promote information exchange on relevant disease progression tasks. MTFL not only utilise the inherent properties of tasks to enhance prediction performance, but also yields weights that are capable to indicate nuanced changes of related AD biomarkers. Task regularized priors, however, introduced by MTFL lead to uncertainty in biomarkers selection, particularly amidst a plethora of highly interrelated biomarkers in a high dimensional space. There is little attention on studying how to design feasible experimental protocols for assessment of MTFL models. To narrow this knowledge gap, we proposed a **Randomize Multi-task Feature Learning (RMFL)** approach to effectively model and predict AD progression. As task increases, the results show that the RMFL is not only stable and interpretable, but also reduced by 0.2 in normalized mean square error compared to single-task models like Lasso, Ridge. Our method is also adaptable as a general regression framework to predict other chronic disease progression.

Keywords: Multi-task feature learning · Alzheimer's disease · Randomization · Stability selection

1 Introduction

Alzheimer's disease, as one of the most common forms of dementia, is a neurodegenerative disease that causes problems with progressive cognitive decline and memory loss [8]. With rates projected to increase by 75% in the next quarter of a century [1], AD is a leading contributor to disability amongst older people and causes significant morbidity as well as personal family burden. So far, there is no effective cure for AD where science has not yet identified any treatments that can slow or halt the progression of this disease. Yet, early intervention and timely diagnosis could be still promising and cost-effective. It poses an important

research area that understands how the AD progresses and identify their related pathological biomarkers for the progression. To accelerate AD’s research, the Alzheimer’s Disease Neuroimaging Initiative (ADNI) funded by NHI provided a large boundary of publicly available neuroimaging data including magnetic resonance imaging (MRI), positron emission tomography (PET), other biomarkers and cognitive measures for scientific study. A variety of medical data driven based machine learning techniques [9, 10, 21–23], like deep learning models [5, 11], multi-task feature learning (MTFL) model [12, 24, 26] and survival model [15, 19, 20], have been investigated to deal with these data for better prediction of AD progression. The motivation of those study is to learn a stable set of features across all tasks and share them to improve the accuracy of all tasks. However, before they share feature information, picking out stable and unbiased features is a key challenge.

Randomization as a method of machine learning has been extensively used in theoretical algorithms and real-world applications [18]. It prevents the selection bias and insures against the accidental bias. For example, in ensemble learning approaches, the Random Forest and the Extra-Trees algorithm [13, 16] belong to two averaging algorithms based on randomized decision trees. Both algorithms are perturb-and-combine techniques [2] specifically designed for trees. This means a diverse set of models is created by introducing randomness in the model’s construction. The prediction of the ensemble is given as the averaged prediction of the individual models. Despite the algorithms in ensemble learning have good predictive accuracy, they are black box methods which are unable to explain the reasons behind the result. Particularly in the field of medically assisted diagnosis as well as in finance, the value of model interpretability is much higher than the accuracy of its predictions.

In this paper, we introduce a randomize multi-task feature learning (RMFL) approach for effectively modelling and predicting AD progression. We examine typical MTFL models via randomized structural regularization approaches in AD study and choose two typical single task models: Ridge regression and Lasso regression. Considering that MTFL features shared parameters and representations, we further explore four potential key points affecting evaluation process of RMFL in AD study: 1) evaluation indicators: validating the model’s robustness on different type of square error or correlation coefficient; 2) repeated experimental times (e.g., results of 10 repeated experiments and 100 repeated experiments are different results; 3) size and portion of training data; 4) number of tasks in MTFL (e.g., time points in AD progression). For each point, we design and set up experimental protocols for comparison and exploration, highlighting following multi-fold contributions:

- We introduce a RMFL strategy that is capable of predicting AD progression with high accuracy, while elucidating the structure that can structural nuances indicative of significant biomarkers alterations in AD.
- We provide a solid evidence that whether RMFL model perform well in complex practical experimental settings. One key finding is that MTEN’s superior performance may stem from the stability selection of features across multiple

tasks. This provides a checkpoint for whether the model works well in more complex practical applications.

- By leveraging methodical validation, we demonstrate that some limitations of MTFM models in AD study: 1) the normalized mean square error emerges as the most reliable performance metric, while alternative evaluative indicators lack comparable objectivity. 2) MTEN has a considerable potential for further improvement at late stage prediction of AD progression. 3) The assumption of temporal smoothness in MTFM models for AD study constrains early task performance.

2 Methodology

2.1 Subjects

To track the effectiveness of disease progression models, [ADNI-1](#) subjects with all corresponding MRI and cognitive scales are evaluated. The ADNI is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of AD. Since its launch more than a decade ago, the landmark public-private partnership has made major contributions to AD research, enabling the sharing of data between researchers around the world. A total of 800 subjects, approximately 200 normal individuals (NL), 400 subjects with Mild cognitive impairment (MCI) and 200 subjects with early AD, were involved in this study. All participants received standard clinical tests of cognitive function to be followed for 3 years, such as Mini Mental State Exam score (MMSE), Alzheimer’s Disease Assessment Scale cognitive total score (ADAS-cog) and Rey Auditory Verbal Learning Test (RAVLT). The date of the participant’s first visit to the hospital for screening was set as the baseline period in order to facilitate comparison with subsequent changes in the participant’s status. The follow-up points, such as 6 or 12 months after the baseline point, supported the longitudinal disease progression of the subjects. For example, “M12” was defined as the follow-up survey at month 12 after baseline. As the timeline lengthens, the number of subjects who still have follow-up records gradually decreases, but detailed data at the screening stage is useful for early detection of a patient’s potential risk of AD.

2.2 Image Pre-processing

For guarantee high image quality and reliable data handling, the MR images used in the paper were derived from standardized datasets, which provide the intensity normalized and gradient un-warped TI image volumes. Subsequently, the FreeSurfer image analysis suite [4] was performed to feature extraction of the MR, which executes cortical reconstruction and volumetric segmentations for processing and analyzing brain MR images. For each MRI, cortical regions and subcortical regions are generated after this pre-processing suite. For each cortical region, the cortical thickness average, standard deviation of thickness,

surface area, and cortical volume were calculated as features. For each subcortical region, subcortical volume was calculated as feature. Data cleaning operations are performed:

- Removal of individuals who failed cortical reconstruction and failed quality control;
- Removal of features with more than half of the missing values;
- Individual subjects whose removal of baseline did not screen for MRI;
- Using the average of the features to fill in missing data;
- Removal of cognitive function tests in individuals with missing follow-up points in longitudinal studies.

After the pre-processing procedure, there are a total of 429 subjects and 327 MRI features.

2.3 Regression Model via Structural Regularization

Regression model has been widely used in statistical, medical and industrial applications. It is a mathematical and statistical analysis of dependent influences (independent variables) and predictors (dependent variables). Its strength lies in its strong interpretation. By fitting the data, the parameter values corresponding to the independent variable indicate its effect on the dependent variable.

We consider the problem of prediction as a linear model. In order to obtain models with generalizability, loss functions with empirical structural loss risk minimization as the formula:

$$\min_{\beta} L(y, X, \beta) + \lambda R(\beta) \quad (1)$$

where the loss term $L(y, X, \beta)$ measures how well the model fits the data, the regularization term $R(\beta)$ measures model complexity. When $\lambda \geq 0$ denotes the penalty parameters, i.e., balancing the goal of fitting the training with the goal of keeping the parameter values small, come to keep the hypothesis relatively simple in form and avoid overfitting.

In general, the sample contains a large number of possible biomarkers for the patient, such as MRI statistical values for the regional cortex, CSF, biochemical indicators and cognitive scores. They are transformed into features that can be run by the model so that the relatively important subset of features can be filtered out in the subsequent training process.

The regularization term is considered as the addition of a prior, and common paradigms are Ridge regression and Lasso, which respectively add the L_1 and L_2 norm. Statistical theory can prove that Ridge regression specifies a prior that the model obeys a Gaussian distribution and Lasso specifies a prior that the model obeys a Laplace distribution. This regularization term can be expressed as:

$$\min_w L(Y, X, W) + \lambda \|W\|_1 \quad (2)$$

$$\min_w L(Y, X, W) + \lambda \|W\|_2 \quad (3)$$

where ridge regression constrains variables to a smaller range for reducing some factors with little impacts on model’s prediction. Unfortunately, this reduction means that these variables are still considered. To solve this problem, Lasso was proposed as a new sparse representation linear algorithm, which simultaneously performs feature selection and regression. Some variables are set to zero directly to achieve sparsity and dimensionality reduction. In addition, some randomization-based sparse algorithms [17] put in different prior assumptions to achieve the desired effect and kernel extended strategy [3] to cope with nonlinear system in complex space.

2.4 Multi-task Feature Learning

A popular setting of multi-task feature is to treat a regression model as a task. The purpose of multi-task feature learning [6] is to learn a common set of features across all tasks and share them to improve the accuracy of all tasks. Among these learning tasks, a basic assumption of MTL is that one or more subsets are related to each other.

Let $X = [x_1, \dots, x_n]^T \in \mathbb{R}^{n \times d}$ be the data matrix, $Y = [y_1, \dots, y_n]^T \in \mathbb{R}^{n \times k}$ be the predicted matrix, and $W = [w_1, \dots, w_k]^T \in \mathbb{R}^{d \times k}$ be the weight matrix. The process of establishing a MTL model is to estimate the value of W , which is the parameter to be estimated from the training samples.

Two common MTL models are presented to display their properties. Multi-Task lasso is a linear model that estimates sparse coefficients for multiple regression problems jointly. The constraint is that the selected features are the same for all the regression problems, also called tasks. The Fig. 4 compares the location of the non-zero entries in the coefficient matrix W obtained with a simple Lasso or a Multi-task Lasso. Mathematically, it consists of a linear model trained with a L_{21} -norm for regularization. The objective function to minimize is:

$$\min_w \frac{1}{2n} \|XW - Y\|_F^2 + \alpha \|W\|_{21} \quad (4)$$

where $\|\cdot\|_F$ denotes the Frobenius norm $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}$, and $\|W\|_{21}$ denotes $\|W\|_{21} = \sum_{i=1}^d \sqrt{\sum_{j=1}^k W_{i,j}^2}$. The multi-task lasso allows to fit multiple regression problems jointly enforcing the selected features to be the same across tasks. For example, AD cognitive progress sequential measurements, each task is a time instant, and the relevant features vary in amplitude over time while being the same. This makes feature selection by the Lasso more stable. However, when there are correlations between multiple features, the features will be randomly selected, especially when the brain region is regarded as a feature, there are some blocks with high correlation, such as atrophy of the cerebral cortex causes reduction in cortical volume and cortical thickness.

Another approach of MTL is multi-task elastic net (MTEN). It can compensate for the shortcomings generated by multi-task lasso. When multiple features are correlated with one another, MTEN tends to select both features rather than

a random. Mathematically, it consists of a linear model trained with a mixed L_{21} -norm and L_{21} -norm for regularization. The objective function to minimize is:

$$\min_w \frac{1}{2n} \|XW - Y\|_F^2 + \alpha \rho \|W\|_{21} + \frac{\alpha(1-\rho)}{2} \|W\|_F^2 \quad (5)$$

The difference from multi-task lasso is that MTEN adds a constraint on the F-norm of W . α and ρ controls the strictness of model penalties to trading-off the advantages between Lasso and Ridge. When $\rho = 0$, MTEN degrades to multi-task lasso; When $\alpha = 0$, MTEN degrades to traditional linear regression problem.

2.5 Randomize Multi-task Feature Learning

Randomization as a method of machine learning has been extensively used in theoretical algorithms and real-world applications [18]. It prevents the selection bias and insures against the accidental bias. For example, in embedded feature selection schemes, randomization has recently received increasing attention due to the use of randomization-related techniques to select a more stable and less biased feature subsets. Stability selection are one of them.

Stability selection is based on subsampling in combination with (high dimensional) selection algorithms. In previous related studies [24], the stability ranking score gives the probability that it is naturally interpretable. This study propose to extend a strategy of stability selection to multi-task feature study to quantify the importance of the features selected by the MTFE formulations for predicting disease progression. Multi-Task elastic network algorithm was utilized to track the area of the cerebral cortex associated with AD progression.

Let F be the overall set of features and let $f \in F$ be the subset of features by sub-sampling. Let γ denote the iteration number of sub-sampling and $D_i = \{X(i), Y(i)\}$ denote one random sub-sample operation of number $i \in (0, \gamma]$. Each operation size account for $\lfloor \frac{n}{2} \rfloor$. Let Λ be the regularization parameter space. For a $\lambda \in \Lambda$, let $\hat{W}^{(i)}$ denote the model coefficient of MTFE that fitted on a subset of $D(i)$. Then, the subset of features generated in task j by the sparse constraints of the MTFE algorithm can be denote as:

$$S_j^\lambda(D_{(i)}) = \{f : \hat{W}_j^{(i)} \neq 0\}. \quad (6)$$

With stability selection, we do not simply select one model in the parameter space λ . Instead the data are perturbed (e.g. by sub-sampling) γ times at task j and we choose all structures or variables that occur in a large fraction of the resulting selection sets:

$$\hat{\pi}_j^\lambda = \frac{\sum_{i=1}^{\gamma} I(f \in S_j^\lambda(D_{ij}))}{\gamma}. \quad (7)$$

where indicator function $I(\bullet)$ denote $I(x) = \begin{cases} 1, & x = 0 \\ 0, & \text{others} \end{cases}$ and $\hat{\pi}_j^\lambda \in [0, 1]$ denote the stability probability of task j at MTFE approaches which feature selection

is not based on individual operations but on multiple task collaboration constraints.

Repeat the above procedure for all $\lambda \in \Lambda$, we obtain the stability score $S_j(f)$ for each feature f at task j :

$$S_j(f) = \max_{\lambda \in \Lambda} (\hat{\pi}_j^\lambda). \quad (8)$$

Finally, for a cut-off π_{th} with $0 < \pi_{th} < 1$ and a set of regularization parameters Λ , the set of stable variables is defined as:

$$\hat{S}^{\text{stable}} = \{k : S_j(f) \geq \pi_{th}\} = \left\{ k : \max_{\lambda \in \Lambda} (\hat{\pi}_j^\lambda) \geq \pi_{th} \right\}. \quad (9)$$

The embedded multi-task approach ensures that the selected features have the following properties: 1) Stability. A cortical region of the brain that is closely related to the subject’s disease progression. 2) Global significance. MTFL makes sure that the selected features are important for each task. One technique that arises here is to pick the coefficient value for one of the tasks when doing statistics on the stability of the selected features at Eq. 4. Overall, the complete stability selection procedure is shown below:

- Randomized selection of feature subsets;
- Randomly selected data subsets;
- Given a hyperparameter search range and a selected set;
- Training Multi-task model and Obtaining weighting factors;
- Polling statistics to find out the probability of a feature being selected;
- Chosen the maximum value as its final stability probability in each randomization algorithm;
- Feature selection based on a given threshold.

3 Experiment

3.1 Experiment Setup

First, experiments demonstrated that MTFL is superior in following AD progression. Combined with randomization techniques, RMFL is enable to locate the stable and sensitive cortical biomarkers. Our empirical protocol design is based on a pipeline shown in Fig. A5. The total experimental process mainly includes 5 steps: 1) split the data set; 2) select the hyper-parameters; 3) train the model; 4) evaluate the model using the test set; 5) iterate the above operations and 6) randomize multi-task feature selection strategy. Different colors denote the source or generation of different data, arrows indicate the flow of data, and serial numbers indicate the steps of the experiment.

Then, to demonstrate that the MTFL algorithm is more generalizable and stable in a variety of realistic scenarios, Four protocol is set up to explore the potential influence that the error arising from the experimental process itself: 1) evaluation indicators, 2) repeated experimental times; 3) size and portion

of training data; 4) number of tasks in MTFML. In addition, the significance of randomize multi-task feature selection strategy in guiding the search for stable biomarkers was demonstrated in [Experiment II visually stability biomarkers](#).

The evaluation metric of cross-validation is employed to evaluate the performance of AD progression model. When a metric is set in the cross-validation experiment process, a set of hyper-parameters can be obtained. By comparing the pros and cons of the results, the suitable metric for the model is finally determined. The regression performance metric often employed in MTL is normalized mean square error (nMSE) and root mean square error (rMSE) is employed to measure the performance of each specific regression task. In particular, nMSE has been normalized to each task before evaluation, so it is widely used in MTL methods based on regression tasks. Also, weighted correlation coefficient (wR) as employed in the medical literature addressing AD progression problems [7, 14, 25]. nMSE, rMSE and wR are defined as follows:

$$\text{nMSE}(Y, \hat{Y}) = \frac{\sum_{i=1}^t \|Y_i - \hat{Y}_i\|_2^2 / \sigma(Y_i)}{\sum_{i=1}^t n_i} \quad (10)$$

$$\text{rMSE}(y, \hat{y}) = \sqrt{\frac{\|y - \hat{y}\|_2^2}{n}} \quad (11)$$

$$\text{wR}(Y, \hat{Y}) = \frac{\sum_{i=1}^t \text{Corr}(Y_i, \hat{Y}_i) n_i}{\sum_{i=1}^t n_i} \quad (12)$$

3.2 Experiment I Prediction with Cerebral Cortex Features

In many real-world AD application scenarios, clinicians expect the prediction model to be simple and with less input data required for giving timely early screening. In this case, it is hard to acquire both precise MRI and cognitive measures. Normally, clinicians have to spend few hours to measure AD patients' cognitive scores through some tests. Thus, one key application was considered with only MRI data as input data to predict cognitive scores at baseline and future time points. It is necessary for clinicians to perform a cognitive scale assessment, but time-consuming to complete a set of cognitive measures.

The first goal is to show a quantitative analysis of typical MTFML methods (MTEN) in comparing to single task methods (Ridge, Lasso). The external experiment setting remained consistent, with same split ratio of sample data, iteration times and features. Specifically, dataset was randomly split into training and testing sets using a ratio 9:1, i.e., models were built on 90% of the data and evaluated on the remaining 10% of the data. Models parameters were selected by 5-fold cross validation.

The results in Table 1 implies that three selected structural regularization methods are all robust (low variance). Also, MTEN models outperforms single-task learning model (Ridge and Lasso), in terms of prediction accuracy.

Table 1. Validation of AD disease progression based MTFL

	Ridge	Lasso	MTEN
<i>Target: MMSE</i>			
nMSE	2.088 ± 0.359	0.945 ± 0.247	0.745 ± 0.172
wR	0.310 ± 0.070	0.499 ± 0.034	0.568 ± 0.053
BL rMSE	2.841 ± 0.298	2.042 ± 0.496	1.721 ± 0.225
M06 rMSE	3.767 ± 0.408	2.491 ± 0.480	2.197 ± 0.244
M12 rMSE	3.958 ± 0.456	2.717 ± 0.587	2.368 ± 0.535
M24 rMSE	4.633 ± 0.579	3.320 ± 0.713	2.944 ± 0.437
M36 rMSE	5.745 ± 0.708	3.947 ± 0.791	3.820 ± 0.736
<i>Target: ADAS-cog</i>			
nMSE	1.147 ± 0.111	0.729 ± 0.060	0.698 ± 0.063
wR	0.468 ± 0.046	0.542 ± 0.052	0.573 ± 0.044
BL rMSE	5.465 ± 0.599	4.229 ± 0.534	4.117 ± 0.558
M06 rMSE	5.900 ± 0.840	4.590 ± 0.672	4.489 ± 0.675
M12 rMSE	6.074 ± 0.894	4.998 ± 0.754	4.759 ± 0.620
M24 rMSE	7.483 ± 1.200	5.818 ± 1.066	5.761 ± 1.012
M36 rMSE	8.905 ± 1.361	7.981 ± 1.420	7.730 ± 1.221

Key: MMSE, Mini-Mental State Examination; ADAS-cog, Alzheimer’s Disease Assessment Scale Cognitive Subscale; BL Baseline visiting point; All algorithms were repeat 100 times and their means ± variance was counted. Represents that the result in bold is statistically significantly better than other comparison methods.

3.3 Experiment II Visually Stability Biomarkers

Experiment screened all MRI features using stability selection strategy and obtained 126 stable features, which were stable scores ≥ 0.96 . Then, this feature set was put back into the MTEN algorithm to obtain a 35 stable sub-features, which can be used to track cortical biomarkers associated with AD progression. The stability vectors of stable MRI features for MMSE are shown in Fig. 1. Experiment finds that the imaging biomarkers identified by RMFL yielded promising patterns that are expected from prior knowledge on neuroimaging and cognition. Some important features are selected, such as Inferior Parietal, Hippocampus, Middle Temporal Gyri and Fusiform, are relevant to the cognitive function.

3.4 Experiment III Evaluation Indicators

In MTFL for AD study, cross-validation with evaluation metric is widely utilised to select suitable model hyper-parameters. Fair hyper-parameters could make MTFL models have better generalization performance. When an evaluation indicator is set in cross-validation experiment process, a set of hyper-parameters can

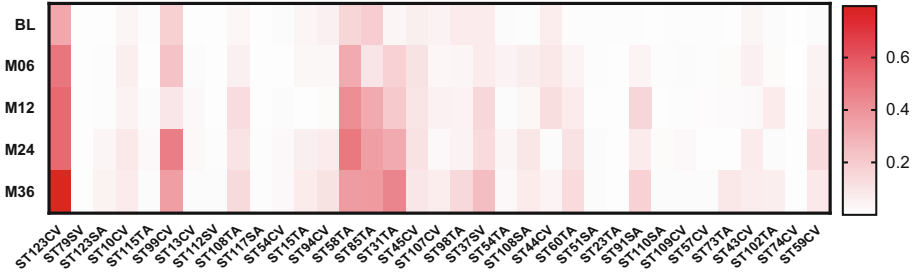


Fig. 1. Thermogram of MRI stability features by multi-task elastic net. Each column represents a cortical region of the brain selected by randomization technique.

be obtained. By comparing the pros and cons of the results, the suitable metric for the model is finally determined. However, different metrics have different preferences and emphasis on the model. It has become a consensus to employ metrics to evaluate the pros and cons of models.

Three models (Lasso, TGL and MTEN) are selected for evaluation. Dataset was randomly split into training and testing sets using a ratio 9:1. Models parameters were selected by 5-fold cross validation. The mean and standard deviation based on 20 iterations of experiments. The experimental results in Table 2 showed that selection of evaluation metrics significantly affect performance assessment of MTFE models.

According to our results, therefore, it can be seen that 1) the results obtained by metrics such as square error (MSE, rMSE, nMSE) are basically the same; 2) nMSE is the best indicator to evaluate these models due to relatively stable performance. The reason is that data distribution of each task is not the same, sharing with each other will have the effect of noise. Therefore, using the variance of tasks in nMSE will reduce the impact of task differences, and the results can better take into account each other’s tasks.

3.5 Experiment IV Repeated Experimental Times

In MTFE for AD study, one typical consensus is that one experiment result is usually accidental and unreliable. To reduce experiment accidental errors, repeated experiments are required. Therefore, we evaluate the performance of four MTFE models under different repeated experimental times. We conducted 6 sets of experiments, and the number of iterations in each set was 5, 10, 20, 30, 40, 50, 100. Also, in each set of experiments, other conditions remained the same. The final result is shown in Fig. 6. The horizontal axis represents iteration, the vertical axis represents the nMSE value of each algorithm, and different colors represent algorithm. In Fig. 6, it appears that the effect of different experiments on three algorithms are visually observed. MTEN models maintains good performance in each set of experiments. From the fluctuation range of the model mean: Ridge not only performs poorly overall, but also has a large range of

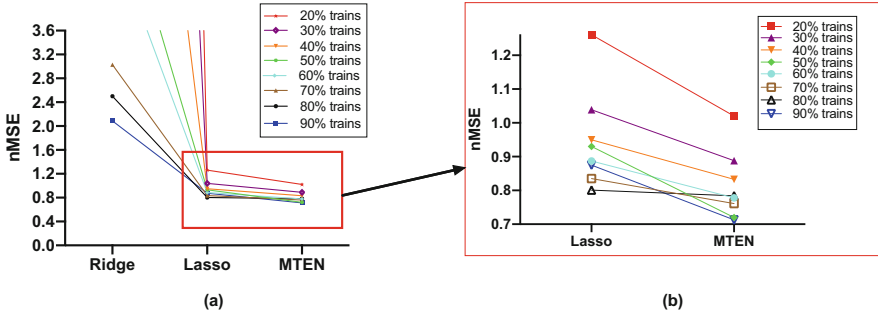


Fig. 2. nMSE values for predicting MMSE cognitive scores under different data size. Each colour label indicates the proportion of the training set in the overall.

fluctuations, which may be the reason for the under-fitting. As the number of iterations increased, three algorithms are fluctuating to varying degrees. Lasso and MTNE are relatively less affected, which implies that sparsity plays a key role in real-world scenarios.

3.6 Experiment V Size and Portion of Training Data

One significant advantage of MTFML is to deal with the issue of missing data and reduce the risk of overfitting. To prove this assumption, we evaluate different portion of training AD data over these MTFML models. Experiment train four MTL models with datasets of different data sizes with 8 groups of experiment performances. Data was split into training and test sets according to the ratio (2: 8, 3: 7, 4: 6, 5: 5, 6: 4, 7: 3, 8: 2, and 9: 1) respectively. For example, in order to compare the experimental results, the other condition settings of each group of experiments are kept consistent: datasets with MMSE scores as learning labels are conducted, with 429 and 425 samples respectively. The same data set was used to predict the trend of cognitive scores of the MMSE and ADAS-cog scales at baseline and in the next three years. The result based on 50 iterations of experiments on different splits of data using 5-fold cross validation. Each group of experiments uses 3 algorithms (Ridge, Lasso, and MTEN) for comparison. The results are shown in the Fig. 2 (a). The finding shows that: Ridge and Lasso have high overfitting risks but MTEN show advantages. In addition, to clarify the difference in performance between Lasso and MTEN, Fig. 2 (b) is the comparison of Lasso and MTEN in detail, connecting the mean two points with a straight line whose slope is less than zero, implying that MTEN is optimal for global training processes.

3.7 Experiment VI Number of Tasks in MTFML

Final key issue to MTFML models is to explore the sharing knowledge between multiple tasks. The common method is to propose an assumption and then transform into a constraint and put into an optimization function. But whether this

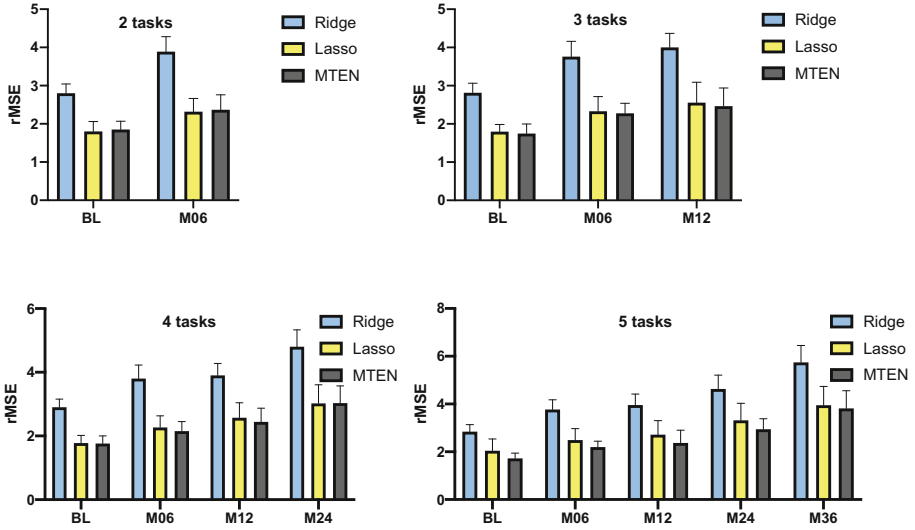


Fig. 3. Histograms of the effect of different numbers of tasks on model performance.

assumption relationship is worth scrutinizing needs to be paid more attention. Therefore, several sets of experiments were designed to test the validity of this relationship. We carried out four sets of experiments using from two to five tasks together to build MTL model. The purpose of the experiment is to find whether the performance of the model can be improved under a certain task relationship. The results were based on 50 iterations of experiments on different splits of data with 9:1 using 5-fold cross validation. Three algorithms (Ridge, Lasso, MTEN) were conducted in each group for comparison. The results are shown in the Fig. 3. The finding shows that:

- As the number of tasks in MTL increases, the accuracy gains of MTL models in AD progression prediction become more obvious. This proves the effectiveness of multi-task learning.
- At 3 or 4 tasks were considered, the errors of the Lasso and MTEN are small. This may be due to the fact that the core element of structure-based regularization of MTL is the use of L1norm. Due to the high similarity between tasks, there is thus less complementary information between tasks, i.e., fewer tasks do not yield significant performance gains.
- The discrepancy results is most obvious when the five tasks is considered simultaneously in one model. Result implies that the sharing knowledge between multiple tasks are effective. Noting that the tasks error also increase, this may be due to a non-linear relationship of MRI features and cognitive scores in the late stage of AD progression.

4 Conclusion

Early intervention of AD may enable clinicians to better monitor disease progression and extend patient longevity. In this study, we introduce RMFL approach to effectively model and predict AD progression. The model is capable of predicting AD progression with high accuracy, even in scenarios characterized by missing data, data scarcity, or reliance on single MRI inputs. We further corroborate the efficacy of the RMFL through rigorous validation across various complex experimental settings. The results show that the RMFL retains stability and interpretability while exhibiting superior performance as the number of tasks increases. This method offers new insights into the role of modeling chronic disease progression and thus may assist in the discovery of more significant biomarkers in future research.

Acknowledgements. The authors wish to acknowledge the support from the Young Scientists Fund of the National Natural Science Foundation of China (Grant No.62301452), China Scholarship Council (No.202107030007) and Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Partnership (EP/T517835/1).

A Lasso and multi-task lasso

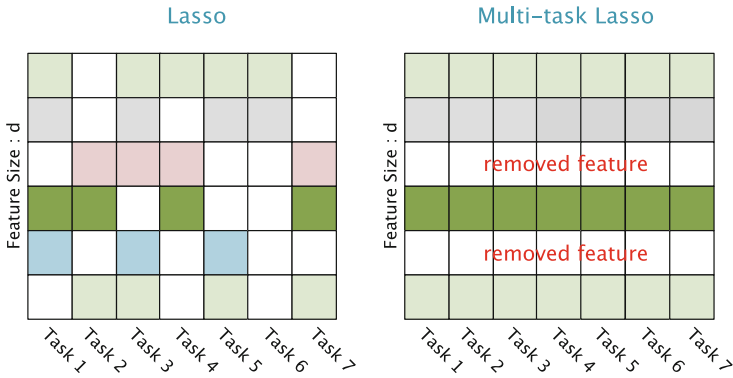


Fig. 4. A comparison of models built by Lasso or a Multi-task Lasso. White block indicates that the parameter value of the position is zero, otherwise, non-zero positions indicated by different colors are used.

B Pipeline

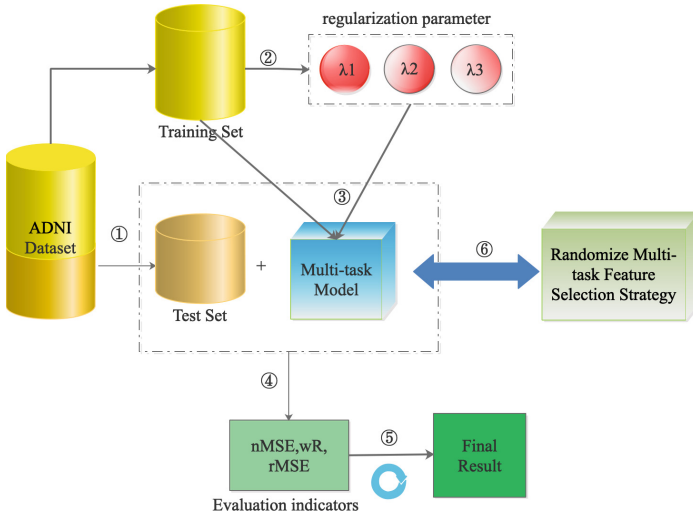


Fig. 5. Pipeline of empirical protocol design.

C Repeated experiments times

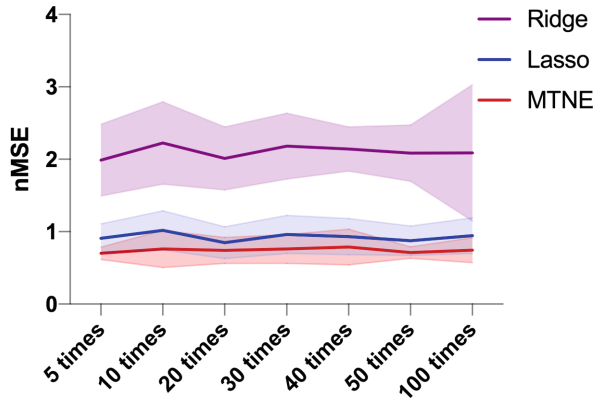


Fig. 6. Evaluation results of repeated experiments times.

D Evaluation indicators

Table 2. The result based on different evaluation indicators

	Ridge	Lasso	MTEN
<i>cv: nMSE</i>			
nMSE	2.779 ± 0.177	0.718 ± 0.137	0.629 ± 0.077
wR	0.516 ± 0.043	0.630 ± 0.049	0.677 ± 0.049
BL rMSE	1.805 ± 0.232	1.803 ± 0.251	1.816 ± 0.286
M06 rMSE	2.345 ± 0.337	2.132 ± 0.293	1.962 ± 0.182
M12 rMSE	2.393 ± 0.537	2.393 ± 0.385	1.966 ± 0.312
M24 rMSE	3.087 ± 0.633	3.087 ± 0.572	2.345 ± 0.400
M36 rMSE	4.924 ± 0.751	3.924 ± 0.683	3.232 ± 0.550
<i>cv: wR</i>			
nMSE	2.783 ± 0.072	0.712 ± 0.192	0.750 ± 0.269
wR	0.514 ± 0.050	0.667 ± 0.043	0.710 ± 0.041
BL rMSE	1.702 ± 0.225	1.813 ± 0.291	2.112 ± 0.329
M06 rMSE	2.293 ± 0.218	2.109 ± 0.312	2.059 ± 0.309
M12 rMSE	2.385 ± 0.425	2.040 ± 0.296	2.092 ± 0.330
M24 rMSE	3.975 ± 0.648	2.570 ± 0.470	2.579 ± 0.809
M36 rMSE	4.635 ± 0.577	3.741 ± 1.118	3.528 ± 0.888
<i>cv: rMSE</i>			
nMSE	2.788 ± 0.091	0.684 ± 0.194	0.630 ± 0.007
wR	0.522 ± 0.044	0.648 ± 0.062	0.691 ± 0.042
BL rMSE	1.776 ± 0.229	1.823 ± 0.293	1.879 ± 0.277
M06 rMSE	2.275 ± 0.348	1.996 ± 0.262	1.943 ± 0.208
M12 rMSE	3.523 ± 0.543	2.133 ± 0.272	1.907 ± 0.243
M24 rMSE	4.180 ± 0.411	2.424 ± 0.544	2.563 ± 0.515
M36 rMSE	4.788 ± 0.556	3.345 ± 0.596	3.149 ± 0.584
<i>cv: MSE</i>			
nMSE	2.765 ± 0.057	0.650 ± 0.087	0.613 ± 0.132
wR	0.527 ± 0.032	0.658 ± 0.039	0.684 ± 0.039
BL rMSE	1.806 ± 0.218	1.748 ± 0.148	1.738 ± 0.252
M06 rMSE	2.304 ± 0.354	1.952 ± 0.234	2.059 ± 0.267
M12 rMSE	2.338 ± 0.486	2.083 ± 0.261	1.992 ± 0.236
M24 rMSE	3.138 ± 0.759	2.689 ± 0.541	2.472 ± 0.576
M36 rMSE	3.876 ± 0.597	3.391 ± 0.645	3.228 ± 0.579

References

1. Association, A.: 2019 Alzheimer's disease facts and figures. *Alzheimer Dement.* **15**(3), 321–387 (2019)
2. Breiman, L.: Arcing classifiers. Tech. rep., Technical report, University of California, Department of Statistics (1996)
3. Chakravorti, T., Satyanarayana, P.: Non linear system identification using kernel based exponentially extended random vector functional link network. *Appl. Soft Comput.* **89**, 106117 (2020)
4. Fischl, B.: Freesurfer. *Neuroimage* **62**(2), 774–781 (2012)
5. Ghazi, M.M., et al.: Training recurrent neural networks robust to incomplete data: application to Alzheimer's disease progression modeling. *Med. Image Anal.* **53**, 39–46 (2019)
6. Gong, P., Ye, J., Zhang, C.s.: Multi-stage multi-task feature learning. In: *Advances in Neural Information Processing Systems 25* (2012)
7. Ito, K., et al.: Disease progression model for cognitive deterioration from Alzheimer's disease neuroimaging initiative database. *Alzheimer. Dement.* **7**(2), 151–160 (2011)
8. Khachaturian, Z.S.: Diagnosis of Alzheimer's disease. *Arch. Neurol.* **42**(11), 1097–1105 (1985)
9. Liu, K., Wang, R.: Antisaturation adaptive fixed-time sliding mode controller design to achieve faster convergence rate and its application. *IEEE Trans. Circuits Syst. II Exp. Briefs* **69**(8), 3555–3559 (2022)
10. Liu, K., Yang, P., Wang, R., Jiao, L., Li, T., Zhang, J.: Observer-based adaptive fuzzy finite-time attitude control for quadrotor UAVs. *IEEE Trans. Aerosp. Electron. Syst.* (2023)
11. Nguyen, M.: Predicting Alzheimer's disease progression using deep recurrent neural networks. *Neuroimage* **222**, 117203 (2020)
12. Peng, J., Zhu, X., Wang, Y., An, L., Shen, D.: Structured sparsity regularized multiple kernel learning for Alzheimer's disease diagnosis. *Pattern Recogn.* **88**, 370–382 (2019)
13. Rao, C., Liu, M., Goh, M., Wen, J.: 2-stage modified random forest model for credit risk assessment of p2p network lending to “three rurals” borrowers. *Appl. Soft Comput.* **95**, 106570 (2020)
14. Stonnington, C.M., et al.: Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. *Neuroimage* **51**(4), 1405–1413 (2010)
15. Sun, M., Baytas, I.M., Zhan, L., Wang, Z., Zhou, J.: Subspace network: deep multi-task censored regression for modeling neurodegenerative diseases. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2259–2268 (2018)
16. Utkin, L.V., Kovalev, M.S., Coolen, F.P.: Imprecise weighted extensions of random forests for classification and regression. *Appl. Soft Comput.* **92**, 106324 (2020)
17. Wang, G., Ma, J., Chen, G., Yang, Y.: Financial distress prediction: regularized sparse-based random subspace with ER aggregation rule incorporating textual disclosures. *Appl. Soft Comput.* **90**, 106152 (2020)
18. Wang, S., Wang, Y., Wang, D., Yin, Y., Wang, Y., Jin, Y.: An improved random forest-based rule extraction method for breast cancer diagnosis. *Appl. Soft Comput.* **86**, 105941 (2020)
19. Wang, X., Qi, J., Yang, Y., Yang, P.: A survey of disease progression modeling techniques for Alzheimer's diseases. In: *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, vol. 1, pp. 1237–1242. IEEE (2019)

20. Yang, P., Bi, G., Qi, J., Wang, X., Yang, Y., Xu, L.: Multimodal wearable intelligence for dementia care in healthcare 4.0: a survey. *Inf. Syst. Front.* 1–18
21. Yang, P., et al.: Feasibility study of mitigation and suppression intervention strategies for controlling Covid-19 outbreaks in London and Wuhan (2020)
22. Yang, P., Yang, C., Lanfranchi, V., Ciravegna, F.: Activity graph based convolutional neural network for human activity recognition using acceleration and gyroscope data. *IEEE Trans. Ind. Inf.* **18**(10), 6619–6630 (2022)
23. Yang, P., et al.: DUAPM: an effective dynamic micro-blogging user activity prediction model towards cyber-physical-social systems. *IEEE Trans. Ind. Inf.* **16**(8), 5317–5326 (2019)
24. Zhang, Y., Lanfranchi, V., Wang, X., Zhou, M., Yang, P.: Modeling Alzheimer’s disease progression via amalgamated magnitude-direction brain structure variation quantification and tensor multi-task learning. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2735–2742. IEEE (2022)
25. Zhou, J., Yuan, L., Liu, J., Ye, J.: A multi-task learning formulation for predicting disease progression. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 814–822 (2011)
26. Zhou, M., Zhang, Y., Yang, Y., Liu, T., Yang, P.: Robust temporal smoothness in multi-task learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 11426–11434 (2023)