Novel Effective Connectivity Inference Using Ultra-Group Constrained Orthogonal Forward Regression and Elastic Multilayer Perceptron Classifier for MCI Identification

Yang Li, Hao Yang, Baiying Lei, Jingyu Liu, and Chong-Yaw Wee

***Abstract*— Mild cognitive impairment (MCI) detection is important, such that appropriate interventions can be imposed to delay or prevent its progression to severe stages, including Alzheimer's disease (AD). Brain connectivity network inferred from the functional magnetic resonance imaging data has been prevalently used to identify the individuals with MCI/AD from the normal controls. The capability to detect the causal or effective connectivity is highly desirable for understanding directed functional interactions between brain regions and further helping the detection of MCI. In this paper, we proposed a novel sparse constrained effective connectivity inference method and an elastic multilayer perceptron classifier for MCI identification. Specifically, a ultra-group constrained structure detection algorithm is first designed to identify the parsimonious topology of the effective connectivity network, in which the weak derivatives of the observable data are considered. Second, based on the identified topology structure, an effective connectivity network is then constructed by using an ultra-orthogonal forward regression algorithm to minimize the shrinking effect of the group constraint-based method. Finally, the effective connectivity network is validated in MCI identification using an elastic multilayer perceptron classifier, which extracts lower to higher level information from initial input features and hence improves the classification performance. Relatively high classification accuracy is achieved by the proposed method when compared with the state-of-the-art classification methods. Furthermore, the network analysis results demonstrate that MCI patients suffer a rich club effect loss and have decreased connectivity among several brain regions. These findings suggest that the proposed method not only improves the classification performance but also successfully discovers critical disease-related neuroimaging biomarkers.**

***Index Terms*— Functional imaging, brain, computer-aided detection and diagnosis, connectivity analysis, machine learning.**

I. Introduction

MILD cognitive impairment (MCI) is identified as the clinical stage between normal forgetfulness and dementia, which commonly suffers from a cognitive decline that does not interfere notably with activities of daily living [1]. However, MCI is associated with increased risk of developing Alzheimer's disease (AD) at a rate of approximately 10%∼15% per year [2] compared with healthy controls who develop dementia at a rate of 1%∼2% per year [3]. By providing appropriate pharmacological treatments and behavioral interventions to MCI, it is possible to delay or prevent the progression of MCI to moderate and severe stages [4]. Thus, there is an urgent need for the accurate MCI detection and diagnosis. However, MCI is difficult to diagnose due to its mild symptoms, especially in high functioning individuals who maintain a positive public profile without showing apparent cognitive impairment [5]. Some anatomical and physiological results suggest that the cognitive process is substantially associated with interactions among brain regions [6].

Constructing brain connectivity from neuroimaging data holds great promise for diagnosing brain diseases and understanding the brain activity interactions. Many functional and effective connectivity modeling approaches based on functional magnetic resonance imaging (fMRI) have been proposed to distinguish diseases, e.g., AD and MCI, from normal controls (NCs) [7]. Among functional connectivity modeling methods, the correlation-based methods are commonly used for MCI identification and can obtain relatively high sensitivity for detecting potentially affected network connections. However, the spurious or insignificant connections in fully

Manuscript received August 28, 2018; revised October 27, 2018; accepted November 12, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61671042 and Grant 61403016 and in part by the Beijing Natural Science Foundation under Grant 4172037. *(Corresponding authors: Hao ~~Yao~~; Chong-Yaw Wee.)*

Y. Li is with the Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China, also with the Beijing Advanced Innovation Center for Big Date-based Precision Medicine, Beihang University, Beijing 100191, China, and also with the School of Automation Sciences and Electrical Engineering, Beihang University, Beijing 100191, China (e-mail: liyang@buaa.edu.cn).

H. Yang and J. Liu are with the School of Automation Sciences and Electrical Engineering, Beihang University, Beijing 100191, China (e-mail: hansyang@buaa.edu.cn).

B. Lei is with the Health Science Center, Shenzhen University, Shenzhen 518060, China (e-mail: leiby@szu.edu.cn).

C. Wee is with the Department of Biomedical Engineering, National University of Singapore, Singapore 117583 (e-mail: cywee2000@gmail.com).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2018.2882189

connected correlation-based network lead to difficult interpretation of the network structure. In order to address this problem, the least absolute shrinkage and selection operator (Lasso) method has been proposed for constructing the sparse connectivity network with only a few of significant connections [8]. The strength of insignificant or spurious connections is forced to be zero and hence make the interpretation of constructed sparse connectivity relatively easier. Meanwhile, since the Lasso method applies l_1-norm penalization at an individual level, the network topological structures vary between subjects [9]. This inter-subject variability inevitably cause comparison difficult and thus possibly degrades the classification performance. The group constrained sparse learning method thus enforces sparsity at the group level via $l_{2,1}$-norm penalization rather than at the level of the individual covariates. Through multi-task learning approach, it minimizes inter-subject variability and keeps the network topological structures identical among subjects [10]. However, the group constrained method above has its own weakness. For example, it heavily shrinks large regression coefficients which lead to the bias problem and it treats the datum points individually, omitting the interconnections among them [11]. These interconnections determine many critical characteristics of a constructed model and the absence of this information may lead to an over-fitting problem [11]. Moreover, the group constrained method focus only on constructing functional connectivity networks, generally ignoring the time-lagged relationships between different brain regions.

Recently, some machine learning approaches have been successfully deployed in the automated classification of connectivity networks related to MCI/AD [12]. These algorithms learn a relationship between the input data attributes and the target attribute by minimizing a loss function defined on the pairs of input data and the associated target attribute. Typically, a k nearest neighbors (KNN) stores all available cases and classifies new cases based on a similarity measure [13]. The random forest is an ensemble classification method that operates by constructing a host of decision trees at training time and outputting the class of the new cases as the mode of the classes of the individual trees [14]. However, these conventional methods only extract the low-level information from the fMRI data and thus cannot reveal the inherent and essential information of the connectivity network [15]. With the recent rejuvenation of neural networks, multilayer perceptron classifiers (MLPCs) have become a robust feature extraction method for MCI classification [16]. Different from the traditional classifiers, MLPCs extract the high-level information from the data by using model architectures of multiple nonlinear transformations. Though the stacked nonlinear transformation layers bring the MLPC excellent power to fit the training data, it also raises the concern of overfitting problem that the MLPC may perform excellently on the training data while works badly on test data, especially for small dataset [17].

To address these problems, in this paper, we propose a novel effective connectivity network modeling method and employed an elastic multilayer perceptron classifier (EMLPC) for MCI identification. Specifically, the effective network topological structure is first detected via an ultra-group constrained structure detection algorithm, which considers not only the discrepancy between observed original data and the model prediction but also the discrepancy between the associated weak derivatives. Thus it increases the noise- resistibility and robustness of the modeling and this helps to characterize the model more accurately. Second, based on the detected structure, an ultra-orthogonal forward regression (UOFR) algorithm is employed to estimate the strength of the effective connectivity networks. The UOFR algorithm alleviates the limitation of shrinking effect via a step forward regression process. Then several topological attributes are further extracted to characterize the effective connectivity networks. Following the feature extraction, an EMLPC is used as the classifier to identify MCI, where the EMLPC combines both the l_1-norm and l_2-norm regularization to obtain a trade-off between fitting ability to the training data and generalization ability on unknown data. The proposed framework is validated via a leave–one-out (LOO) cross-validation strategy for MCI identification. Experiment results illustrate that the proposed approach obtains a competitive performance with a relatively high classification accuracy compared to state-of-the-art methods, hence allows a more accurate detection of brain abnormalities which can useful for better interpretation of the pathological underpinnings of MCI.

II. Methodology

The proposed framework for diagnosing MCI essentially involves data acquisition, the effective connectivity network construction, feature extraction and classification, which is graphically shown in Fig. 1. Procedures of the proposed framework are summarized as follows: 1) Generate ultra-time series using the original regional mean time series and the weak derivatives of the fMRI data; 2) Detect the topology of effective connectivity networks via ultra-group constrained structure detection algorithm; 3) Apply an UOFR algorithm to estimate the effective connectivity strength; 4) Extract topological features from effective connectivity networks; 5) Split the dataset into train data and test data with a LOO cross-validation strategy, and optimize the parameters of the classifier via grid search with the train data; 6) With the optimal parameters, train a EMLPC and test it with the test data. The performance of the proposed framework is evaluated by the average classification accuracy of all the LOO loops.

A. Data Acquisition and Preprocessing

The present study involved 73 participants (36 MCI patients and 37 socio-demographically matched healthy controls) selected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (http://adni.loni.usc.edu/), and the subject IDs are provided in the supplementary material. Data acquisition was performed using a 3 Tesla Siemens scanner with the following parameters: flip angle = 90°; matrix size = 64×64; voxel thickness = 3.4 mm; 197 volumes; 48 slices; TR = 3000 ms and TE = 30 ms. Demographic and clinical information of the participants is provided in Table I. There is no difference between healthy subjects and MCI patients in terms of age, Mini-Mental State Examination (MMSE) score and head-motion measured with mean framewise displacement.

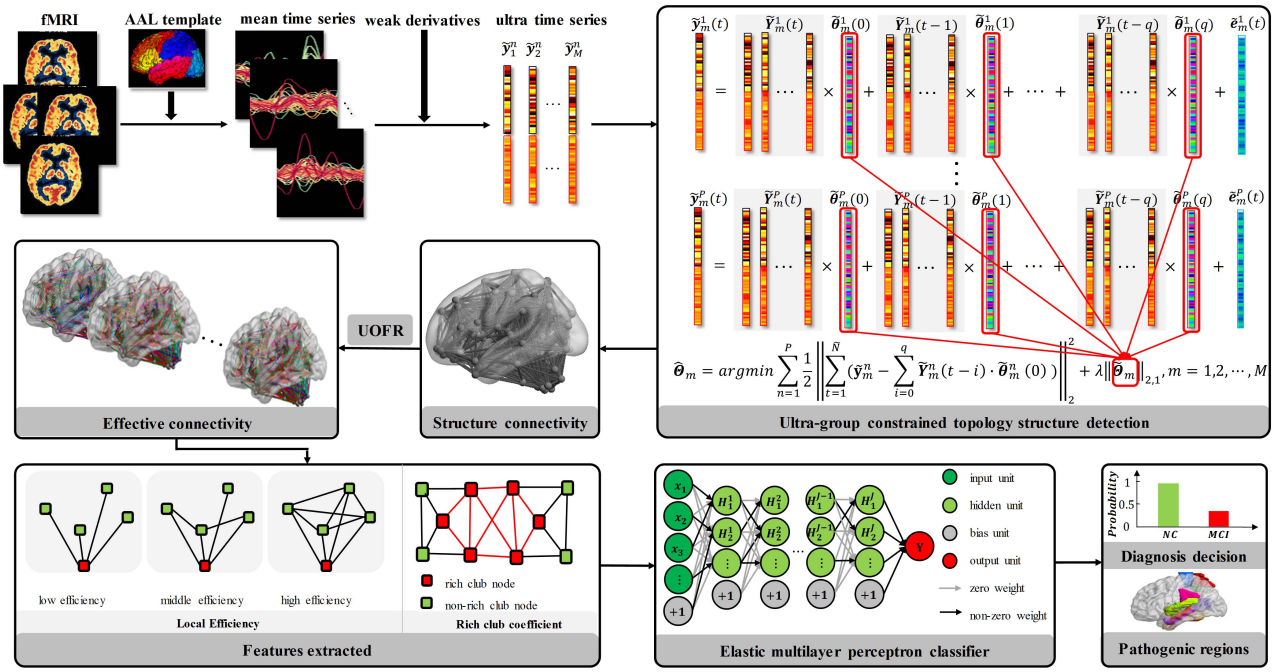

Fig. 1. The flowchart of the proposed method. Specifically, the ultra-mean time series are first computed. Then, a connectivity network topological structure detecting process and connectivity strength estimation process are performed. Finally, connectivity features are extraction and classifiers are trained and tested.

TABLE I
DEMOGRAPHIC AND CLINICAL INFORMATION OF THE PARTICIPANT

Group	MCI	NC	*p*-value
Male/Female	23/13	14/23	-
Age (mean±SD)	76.08±8.30	75.00±7.26	0.5560
MMSE (mean±SD)	27.31±3.65	28.52±2.83	0.1202
MFD (mean±SD)	0.14±0.08	0.14±0.09	0.9356

MMSE: Mini-mental state examination.
MFD: Mean framewise displacement.

Standard data preprocessing was performed using the Statistical Parametric Mapping 12 (SPM12) [18] software package and the Data Processing Assistant for Resting-State fMRI (DPARSF) toolbox [19]. Specifically, we used only the last 180 volumes of the acquired R-fMRI series for the following preprocessing steps to ensure magnetization equilibrium. The preprocessing of R-fMRI images included four main steps: 1) Slice timing correction: correct the remaining 180 volumes for the staggered order of slice acquisition during echo-planar scanning by matching all time points to intermediate time points; 2) Head motion correction: reduce the head-motion artifacts in the R-fMRI time-series by registering all volumes using to the first volume of the remaining R-fMRI time-series of in each subject as a reference to register all the following volume; 3) Spatial normalization: normalize the volumes to the Montreal Neurological Institute (MNI) atlas space and then resample the images to the voxel size of $3 \times 3 \times 3$ mm^3; 4) Spatial smoothing: smooth the dataset with a 4 mm full width half maximum Gaussian kernel. Afterward, we used the cerebrospinal fluid template and the white matter template to extract the values of the associated voxels, *i.e.*, the signal of whole brain is regressed against the average cerebrospinal fluid and white-matter signals as well as the six parameters from motion correction. The brain space was parcellated into 90 regions of interesting (ROIs) by warping the Automated Anatomical Labeling (AAL) template [14] to the subject space using the deformation fields estimated via a deformable registration method called HAMMER [20]. Finally, we obtained the mean time series of each ROI of each subject via averaging time series over gray matter voxels in that particular ROI followed by a band-pass filter of frequency interval $(0.01 \text{ Hz} \leq f \leq 0.08 \text{ Hz})$ to minimize the effects of low-frequency drift and high-frequency noise.

B. Ultra-Group Constrained Topology Structure Detection

Inspired by associated works [8], [9], in our study, a sparse linear regression model is used to detect the structure of connectivity network. Sparse network construction methods, such as the Lasso and group constrained sparse learning method, have been employed to construct functional connectivity networks [21]. These methods use the information of individual datum point and discarding the relationship between them. However, the datum points of a dynamic system are time-dependent and connected with each other through the derivatives of the time continuous function. These interconnections determine many essential characteristics of a system and the absence of the interconnection information may lead to the overfitting problem of the identified model [22]. Moreover, the previous methods only measure the correlation relationship between the brain regions while ignoring their hysteresis effects [23]. Namely, the current brain region

activities are commonly influenced not only by the current activities of other regions but also by the previous activities [23]. To investigate the directional interactions among brain regions and to better understand the interaction mechanism within the brain networks, an ultra-group constrained structure detection algorithm is used to identify the topological structure of the effective connectivity networks.

The weak derivative information of the original fMRI time series in each ROI is first incorporated into the original time series to build the ultra-time series to extract dependent relation information of the associated fMRI time series. Since the original time series are discrete datum points and are not always differentiable, the signal is first smoothed with a test function and then the weak derivatives of original time series are calculated with the smoothed signals. For detailed derivation procedure, please refer to [22]. Suppose there are M ROIs for each subject and P subjects in total, with N being the number of time points, the mean time series of the m-th ROI for n-th subject is represented as $\boldsymbol{y}_m^n(t)$ $(t = 1, 2, \cdots, N)$. Given a normalized test function $\varphi(t)$, the l-th order weak derivative of $\boldsymbol{y}_m^n(\mathrm{t})$ can be defined as

$$\boldsymbol{y}_m^{n(l)}(t) = \sum_{\tau=t}^{t+N_0} y_m^n(\tau)\varphi^{(l)}(\tau - t) \tag{1}$$

where N_0 is the support of the discrete test function with $t = 1, 2, \cdots, N - N_0$. The cubic B-spline basis function is employed as the test function to measure the agreement of data at a local level, and the first and second order derivatives of the smoothed signals are considered as in [11]. Therefore, the ultra-time series generated from y_m^n is defined as $\tilde{\boldsymbol{y}}_m^n = [y_m^n(1), \cdots, y_m^n(N), y_m^{n(1)}(1), \cdots, y_m^{n(1)}(N - N_0), y_m^{n(2)}(1), \ldots, y_m^{n(2)}(N - N_0)]^T$. The length of ultra-time series is $\tilde{N} = 3N - 2N_0$. Note that the ultra-time series $\tilde{\boldsymbol{y}}_m^n$ consist of two parts: the first part $[y_m^n(1), \cdots, y_m^n(N)]^T$ is the original time series that emphases the original datum points, while the second part $[y_m^{n(1)}(1), \cdots, y_m^{n(1)}(N-N_0), y_m^{n(2)}(1), \cdots, y_m^{n(2)}(N-N_0)]^T$ is the weak derivatives that essentially focuses on the dependent relationship of the associated original time series.

With the ultra-time series, the topological structure of the effective connectivity network is then identified via an ultra-group constrained structure detection algorithm which reveals the linear relationship between the current and previous ultra-time series and explores the underlying neuronal interactions. For the n-th subject, the m-th ROI activity $\tilde{\boldsymbol{y}}_m^n(t)$ is represented by a linear combination of current and previous activities of other ROIs, which is defined as

$$\tilde{\boldsymbol{y}}_m^n(t) = \sum_{i=0}^{q} \tilde{\boldsymbol{Y}}_m^n(t-i) \cdot \tilde{\boldsymbol{\theta}}_m^n(i) + \tilde{\boldsymbol{e}}_m^n(t) \tag{2}$$

where q is the model order with $i = 0, 1, \cdots, q$, $\tilde{\boldsymbol{Y}}_m^n(t-i) = [\tilde{\mathbf{y}}_1^n(t-i), \tilde{\mathbf{y}}_2^n(t-i), \cdots, \tilde{\mathbf{y}}_{m-1}^n(t-i), \tilde{\mathbf{y}}_{m+1}^n(t-i), \cdots, \tilde{\mathbf{y}}_M^n(t-i)] \in R^{\tilde{N}\times(M-1)}$ with $m = 1, 2, \cdots, M$, $n = 1, 2, \cdots, P$, and $t = 1, 2, \cdots, \tilde{N}$, $\tilde{\boldsymbol{\theta}}_m^n(i) = [\theta_1^n(i), \theta_2^n(i), \cdots, \theta_{m-1}^n(i), \theta_{m+1}^n(i), \cdots, \theta_M^n(i)]^T$ is the coefficient vector and $\tilde{\boldsymbol{e}}_m^n(t)$ is the residual vector. The model (2) can be re-written in a linear regression form as

$$\tilde{\boldsymbol{y}}_m^n(t) = \tilde{\mathbf{A}}_m^n(t) \cdot \tilde{\boldsymbol{\theta}}_m^n + \tilde{\boldsymbol{e}}_m^n(t) \tag{3}$$

where $\tilde{\mathbf{A}}_m^n(t) = [\tilde{\boldsymbol{Y}}_m^n(t), \tilde{\boldsymbol{Y}}_m^n(t-1), \cdots, \tilde{\boldsymbol{Y}}_m^n(t-q)]$ denotes the regressor matrix with the dimension of $\tilde{N} \times ((M-1) \times (q+1))$ and $\tilde{\boldsymbol{\theta}}_m^n = \left[\tilde{\boldsymbol{\theta}}_m^n(0); \tilde{\boldsymbol{\theta}}_m^n(1); \cdots; \tilde{\boldsymbol{\theta}}_m^n(q)\right]$ is a column vector with $(M-1) \times (q+1)$ elements. The weights of the model are estimated by minimizing a group constrained objective function to reduce the inter-subject variability. The topological structure of different subjects are forced to be consistent via an additional l_2-norm penalization across all subjects. The object function of the ultra-group constrained model is represented as

$$J_{(\tilde{\boldsymbol{\Theta}}_m)} = \sum_{n=1}^{P} \left(\frac{1}{2}\left\|\tilde{\boldsymbol{y}}_m^n - \tilde{\mathbf{A}}_m^n \cdot \tilde{\boldsymbol{\theta}}_m^n\right\|_2^2\right) + \lambda \left\|\tilde{\boldsymbol{\Theta}}_m\right\|_{2,1} \tag{4}$$

where $\tilde{\boldsymbol{\Theta}}_m = \left[\tilde{\boldsymbol{\theta}}_m^1, \tilde{\boldsymbol{\theta}}_m^2, \cdots, \tilde{\boldsymbol{\theta}}_m^P\right]$ and $\left\|\tilde{\boldsymbol{\Theta}}_m\right\|_{2,1}$ is the summation of l_2-norms of row vectors of $\tilde{\boldsymbol{\Theta}}_m$, λ is the regularization parameter that controls the 'sparsity' of the model. The regularization parameter λ forces certain coefficients to zero, effectively choosing a subset of features with non-zero coefficients. The weights associated with certain ultra-time series across different subjects are grouped together via the $l_{2,1}$-norms constraint. This constraint promotes a consistent connection topology among subjects and reduces inter-subject variability meanwhile allows variation of coefficient values between subjects [10]. Additionally, the coefficients of different time lagged ultra-time series of the same ROI are summed together, which are treated as an indicator on whether other ROIs have an influence on the currently considered ROI. Furthermore, the ROIs with non-zero coefficients are consistent for different subjects, and these ROIs are considered to have a connection with the target ROI. Then the associated ultra-time series of these ROIs are selected and arranged into the candidate regressor dictionary to be used for connectivity parameter estimation, while ultra-time series of other ROIs are discarded. The SLEP toolbox [24] was used to solve (4).

C. Connectivity Strength Estimation With UOFR Algorithm

The coefficients $\tilde{\boldsymbol{\Theta}}_m$ estimated via the ultra-group constrained (ultra-GC) method can simply be regarded as the effective connectivity strength (connection weights) between ROIs in an effective connectivity network. However, these estimated coefficients are unscaled and biased, and may lead to difficulty in interpreting and analyzing the effective network [25]. To eliminate the shrinking effect of the ultra-group constrained sparse learning model, we adopt an UOFR algorithm to estimate the effective connectivity strength. By employing the ultra-group constrained structure detection algorithm, a subset of ROIs with non-zero weights is considered to have a connection with the target ROI and the candidate dictionary is constructed by keeping ultra-time series of these ROIs while discarding the others.

In general, the connectivity strengths are estimated in a stepwise orthogonalized forward manner. The values of the error reduction ratio based ultra-time series (UERR.) are first calculated [22]. With $\tilde{N}$ as the length of ultra-time series, the ultra-time series of the target ROI can be expressed as $\tilde{\boldsymbol{y}}_m^n = [\tilde{y}_m^n(1), \tilde{y}_m^n(2), \cdots, \tilde{y}_m^n\left(\tilde{N}\right)]^T$ and the ultra-time series

of the candidate ROIs in the candidate dictionary can be defined as $\tilde{\boldsymbol{x}}_i = [\tilde{x}_i(1), \tilde{x}_i(2), \cdots, \tilde{x}_i\left(\tilde{N}\right)]^T$, the UERR is then calculated as

$$\begin{aligned} \mathrm{UERR}\left(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{y}}_m^n\right) &= \frac{\left((\tilde{\boldsymbol{x}}_i)^T \tilde{\boldsymbol{y}}_m^n\right)^2}{\left((\tilde{\boldsymbol{x}}_i)^T \tilde{\boldsymbol{x}}_i\right)\left((\tilde{\boldsymbol{y}}_m^n)^T \tilde{\boldsymbol{y}}_m^n\right)} \\ &= \frac{\left(\sum_{t=1}^{\tilde{N}} \tilde{x}_i(t)\tilde{y}_m^n(t)\right)^2}{\sum_{t=1}^{\tilde{N}} \left(\tilde{x}_i(t)\right)^2 \sum_{t=1}^{\tilde{N}} \left(\tilde{y}_m^n(t)\right)^2} \end{aligned} \tag{5}$$

where 'T' is transpose of the matrix or vector. Then the connectivity strength between the target ROI and the candidate ROIs are estimated using the commonly used standard orthogonal forward regression algorithm [45]. The UERR values corresponding to the same ROIs are summed and arranged into the effective connectivity matrix which contains every possible effective connectivity of ROIs pairs. The weights in the connectivity matrix can be interpreted as the influence that one ROI has upon another. The details of the effective connectivity construction process are summarized in supplementary material.

D. Feature Extraction

Feature extraction is a special dimensionality reduction approach in machine learning, which projects a high dimensional vector onto low dimensional vector to avoid the curse of dimensionality.

The local efficiency (LE) is a measure that quantifies the cliquishness of the nodes in the network, which indicates the contribution of a node to the communication of the network [26]. For a weighted and directed graph, the local efficiency of the m-th node is defined as [26]

$$E_{loc}(m) = \frac{\sum_{i,j\in G_m} (w_{mi}^{\frac{1}{3}} + w_{im}^{\frac{1}{3}})(w_{mj}^{\frac{1}{3}} + w_{jm}^{\frac{1}{3}})(l_{ji}^{-1} + l_{ji}^{-1})/2}{\left(\sum_{i\in G_m} \delta_{mi}\right)^2 - \sum_{i\in G_m} \delta_{mi}^2} \tag{6}$$

where G_m is the subgraph that contains only neighbors of the m-th node, w and δ are the connection weights and the number of non-zero connections, l is the length of the shortest directed path in G_m, respectively. The weighted local efficiency in (6) distinguishes the influence of different paths based on connection weights of the associated neighbors to the node.

Additionally, recent studies have shown that the presence and absence of the rich club organization reveals the information of the higher-order structure of a connectivity network [27]. The rich club coefficients (RCC) of the network are the ratio of the number of connections among nodes of the degree k or higher versus the total possible number of connections. For weighted networks, it is modified as the fraction of edge weights that connect nodes of degree k or higher out of the maximum edge weights that such nodes might share [28]. Formally, the RCC at level k is given by [28]

$$RCC(k) = \frac{\sum_{i=1}^{E_k} w_i(k)}{\sum_{i=1}^{E_k} w_i^r} \tag{7}$$

where $w_i(k)$ is the weight among nodes of degree k or higher, E_k indicates the number of links among these nodes, and w_i^r is the ranked connections of the whole network, respectively.

Topological attributes above are extracted from the effective connectivity network for each subject. These topological properties quantify connectivity profiles associated with individual network elements (such as nodes or links). Previous studies revealed that the disruption in interactions between brain subnetworks or regions as characterized by topology measures may lead to less efficient information processing and cognitive deficits [29]. Since the deteriorating functional connectivity and cognitive performance are associated with the changes in network topology, the usage of these topology measures as features may potentially boost the classification performance [27], [30]. The feature extraction process was performed using brain connectivity toolbox [26].

E. Classification

In this study, we adopt a MLPC for MCI classification. The MLPC is a kind of deep learning algorithm that models high-level representations in the data by using model architectures of multiple nonlinear transformations. The general idea of the MLPC algorithm is stacking up the nonlinear transformation layers. The more layers the data goes through within the network, the more complicated are the nonlinear transformations and the more abstract is the information extracted from the network. The nonlinear transformations enhance the fitting ability of the MLPC while increasing the risk of over-fitting. Thus, a kernel regularization scheme is employed to constrain the weights of the MLPC and improve its generalizability.

The l_1-norm and l_2-norm regularizations are added to the objective function of the MLPC to reduce the risk of over-fitting and improve the generalization performance of the model. The objective function of the EMLPC with a weight optimization process is defined as

$$\begin{aligned} J_{(W)} = \frac{1}{2Q}\sum_{n=1}^{Q} \left\| p^{\{n\}}(W) - t^{\{n\}} \right\|^2 + \alpha \sum_{i=1}^{N_l} \|W_i\|_1 \\ + \beta \sum_{i=1}^{N_l} \|W_i\|_2 \end{aligned} \tag{8}$$

where $p^{\{n\}}(W)$ is the predicted label of the subject n with weight W, $t^{\{n\}}$ is the true label of the subject n, Q is the number of training subjects, W_i is the weights of the i-th layer, α is the sparse parameter, β is the weight decay parameter and N_l is the total number of layers. The equation (8) consists of three parts: mean square error, sparsity punishment, and weight decay. The first term is a mean square error that measures the residual between the prediction and the true label. The second term is a sparsity penalty which is defined as the summation of l_1-norm of weights. Particularly, the sparsity punishment constraints the weights of the classifier to be sparse such that only a few non-zero weights. Significant features are automatically selected while spurious features are eliminated, thus the noise-resist ability of the classifier is enhanced. The final term represents weight decay which decreases the range of the weights via a l_2-norm regularization. The use of l_1-norm and l_2-norm regularization rules in Equation (8) decreases the

complexity of the model and increases the generalizability of the model.

The structure of EMLPC is determined through grid search in the inner LOO cross-validation. Specifically, the number of layer N_l is searched within the range of [1, 5] with a step of 1 while the number of hidden nodes N_n is searched within the range of [25, 100] with a step of 25. To simplify the model structure and reduce the number of free parameters, the number of nodes in each hidden layer of the EMLPC are set to the same. Secondly, we use l_1-norm regularization parameter α and l_2-norm regularization β to regularize the model to adjust the fitting ability and generalizability without dropout layer in our model. We set the EMLPC with different α (from 10^{-6} to 10^{-2}) and β (from 10^{-6} to 10^{-2}). Thirdly, the weights of the model are initialized close to zeros with a random uniform distribution in the range of $[-0.05, 0.05]$ and the loss function is set as the regularized mean square error. The combination of parameters that produces the best performance on the validation set will be used to construct the MCI diagnosis model.

With the optimal model structure, the hierarchical features are extracted and the MCI diagnosis models are trained. The training process includes two steps: an unsupervised pre- training step to train the network layer by layer via stacked autoencoder [31], and a supervised fine-tuning step to boost the performance by incorporating the label information. The pre-training step is used to greedily extract latent representation of the input data while reducing the feature dimensionality. Firstly, we train the first layer of the model with an autoencoder which nonlinearly combines the data from the input layer into a short code in the middle layer, and then decompress it into representation that closely matches the original data. Taking the input data as low-level features, the output of the middle layer is a compressed representation of the input data which we consider it relatively high-level features. Then, we take these features as the input of the next layer and another autoencoder is trained to further compress the data and extract higher level features. This process is executed recurrently for each hidden layer of the EMLPC and the outputs of different layers are indeed the hierarchical features from low-level features (the input layer) to high-level features (the output layer). Finally, the whole network is fine-tuned via RMSProp algorithm with a learning rate of 10^{-3} to minimize the difference between the true label and the prediction label of the EMLPC model. The fine-tuning step aims to slightly alter the weights to adjust the boundaries between different groups. The detailed training process is graphically shown in the supplementary material.

F. Parameters Optimization

In the proposed network inference and classification scheme, several free parameters, such as the model order q, sparsity level λ, l_1-norm regularization parameter α, l_2-norm regularization parameter β, the number of layers N_l and the number of nodes N_n of the neural network, should be optimized for achieving the best classification performance. In this study, the model size q is determined by minimizing the Bayesian information criteria (BIC) [33] for each sparsity level and the other five parameters are optimized via a LOO cross-validation scheme.

The model size q is the number of past samples which are needed to accurately predict the present data. Particularly, if the model size q is set to zero, the model simply calculates the partial correlation of current measurements from different brain regions. Thus, the proposed method is capable of evaluating the time-lagged relationship between different brain regions only if the model order is greater than zero. The optimal model size q can be determined based on the BIC value [33]

$$BIC(q) = \tilde{N} \log [mse(q)] + q \log\left(\tilde{N}\right) \tag{9}$$

where $\tilde{N}$ is the length of ultra-time series, mse is the mean square residuals for all subjects and ROIs. By minimizing BIC value, we search for the optimal model order q in the range of $[q_{min}, q_{max}]$. Following the previous study [33], q_{min} is set to 0 and q_{max} is set to 10.

Additionally, an inner LOO cross-validation scheme is used to optimize other hyper-parameters. It should be noted that the classification performance of outer LOO loops can be used to estimate the hyper parameters of a model and then those hyper-parameters are applied to fit a model to the whole dataset, which is likely to be optimistically biased and lead to the concern of over-fitting [34]. The main reason is that part of the model (the hyper-parameters) have been selected to optimize the final classification performance, so if the LOO statistic has a non-zero variance, there is the possibility of over-fitting the model selection criterion [34]. In order to choose the hyper- parameters and estimate the performance of the resulting model in an unbiased manner, we need to perform an inner LOO cross-validation to determine the hyper-parameters, with the outer LOO used to assess the performance of the model built based on the selected hyper-parameters on the unseen testing subject.

Specifically, suppose P subjects are involved in the study, a subject is first left out for testing and the remaining $P-1$ subjects are used to search for the optimal parameters. Among $P-1$ training subjects, one more subject is left out and the remaining $P-2$ subjects are used to construct a classifier based on a specific set of parameter values. The constructed classifier is then used to predict the label of the second left out subject. This process is repeated for $P-1$ times, each time a different subject is left out from the $P-1$ training subjects, to predict the labels of all $P-1$ training subjects and hence the inner LOO cross-validation accuracy. The process is repeated for all potential sets of parameters to determine the optimal set of parameters that gives the best classification accuracy based on $P-1$ training subjects. To fully utilize the information, an EMLPC is constructed with the identified optimal parameters using all $P-1$ training subjects and then used to predict the label of the first left out testing subject. This process is repeated P times, each time leaving out a different subject, to predict the labels of all P subjects and then obtain the overall cross-validation accuracy. Note that the nested cross-validation scheme is an unbiased

and robust evaluation method that can minimize the risk of overfitting [34]. Therefore, the classification model is reliable and competent to be applied for different datasets. We report only the results of overall cross-validation performance in this study. The classification and parameter optimization scheme is graphically illustrated in the supplementary material.

III. Experiment Results

A. Overview of Classification Performance

Several evaluation metrics such as the accuracy, sensitivity, specificity and balanced accuracy are adopted to evaluate the classification performance. We compare our proposed method with several other connectivity inference methods for MCI classification on the same dataset, which is shown in Table II. From Table II, the proposed method outperforms the competing state-of-the-art approaches. The classification accuracy by the proposed scheme of 80.82% is the highest among the competing methods, indicating its excellent ability to identify MCI patients from normal controls. In addition, the proposed method achieves a sensitivity of 80.56%, specificity of 81.08%, also the highest among all competing methods. Note that sensitivity and specificity are usually combined into a single measure as balanced accuracy, which is the arithmetic mean of the specification and sensitivity. The balanced accuracy is more accurate and stable for the performance evaluation particularly for imbalanced dataset. Concerning the balanced accuracy of 80.82%, the proposed method also outperforms all other methods, which indicates the excellent diagnostic power of the proposed method.

TABLE II
Comparison of Classification Performance

Method	Classifier	ACC(%)	SEN(%)	SPE(%)	BAC(%)
Pearson Correlation	KNN	54.79	53.66	56.25	54.95
	RF	60.27	58.97	61.76	60.37
	MLPC	64.38	64.71	64.10	64.40
	L_1-RMLPC	65.75	63.41	68.75	66.08
	L_2-RMLPC	64.38	62.50	66.67	64.58
	EMLPC	65.75	68.97	63.64	66.30
Lasso	KNN	60.27	60.00	60.53	60.26
	RF	65.75	62.22	71.43	66.83
	MLPC	65.75	67.74	64.29	66.01
	L_1-RMLPC	67.12	70.00	65.12	67.56
	L_2-RMLPC	65.75	70.37	63.04	66.71
	EMLPC	68.49	69.70	67.50	68.60
Group constrained method	KNN	60.27	58.54	62.50	60.52
	RF	63.01	64.52	61.90	63.21
	MLPC	67.12	68.75	65.85	67.30
	L_1-RMLPC	71.23	69.23	73.53	71.38
	L_2-RMLPC	69.86	69.44	70.27	69.86
	EMLPC	72.60	76.67	69.77	73.22
Ultra-group constrained method	KNN	65.75	65.71	65.79	65.75
	RF	64.38	64.71	64.10	64.40
	MLPC	69.86	68.42	71.43	69.92
	L_1-RMLPC	73.97	75.76	72.50	74.13
	L_2-RMLPC	69.86	69.44	70.27	69.86
	EMLPC	79.45	80.00	78.95	79.47
Proposed method	KNN	73.97	77.42	71.43	74.42
	RF	75.34	76.47	74.36	75.41
	MLPC	73.97	74.29	73.68	73.98
	L_1-RMLPC	78.08	76.32	80.00	78.16
	L_2-RMLPC	75.34	71.43	80.65	76.04
	EMLPC	**80.82**	**80.56**	**81.08**	**80.82**

where bold fonts indicate the best results.

B. Classification Performance by Ultra-Group Constraint

We evaluate the performance of different sparsity constraints used for constructing effective connectivity networks. In Fig. 2(a), the classification performance with ultra-GC method outperforms that of the classical group constrained sparse learning method using six different classifiers. Specifically, using the EMLPC as the classifier, the ultra-GC method receives an accuracy of 79.45%, which is 6.85% higher than that of the conventional group constrained learning method. Additionally, compared to the classical group constrained sparse learning based network, the ultra-GC approach improves the classification accuracy from 60.27% to 65.75% in KNN classifier, from 63.01% to 63.38% in RF classifier, from 71.23% to 73.97% in MLPC. These results indicate that the incorporation of the dependent information of datum points into connectivity network construction can indeed improve the performance of clinical identification of MCI.

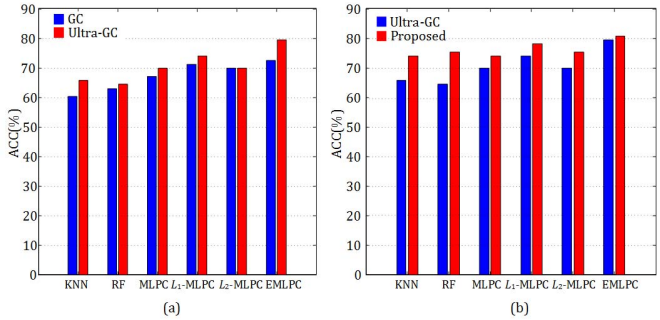

Fig. 2. The comparison of classification performance by (a) the GC and ultra-GC method; (b) the Ultra-GC and the proposed method.

C. Classification Performance Depending on the UOFR Algorithm

The UOFR algorithm is used to eliminate the shrinking effect of the conventional group constrained (GC) method and thus may lead to considerable improvement for MCI classification. To evaluate the effectiveness of the UOFR algorithm, a comparison of the proposed method and the ultra-GC method is graphically shown in Fig. 2(b). In detail, using traditional classification methods such as KNN and RF, an accuracy increase of 8.22% and 10.96% is achieved, respectively. With the deep architecture, the improvement of the performance classification is relatively moderate. For example, when the naive MLPC, L_1-RMLPCN, L_2-RMLPC and EMLPC are used to classify MCI, the classification performance gains of 4.11%, 4.11%, 5.48%, 1.37% are obtained respectively. The improvement of classification accuracy indicates that the step forward regression process of the UOFR algorithm indeed helps to overcome the limitation of the shrinking effect in the group constrained sparse learning method.

D. The Influence of Regularizations

To demonstrate the advantage of regularizations, we compare the classification performance of the MLPCs using varies

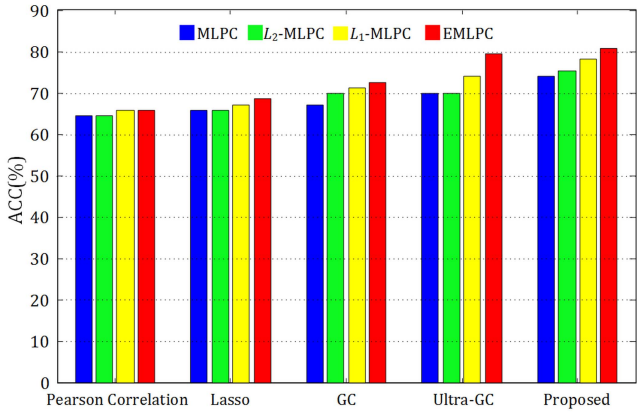

Fig. 3. The comparison performance by different regularizations.

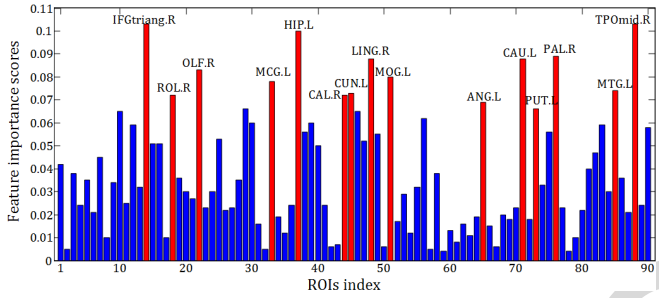

Fig. 4. The importance scores of different brain regions.

regularization strategies. Fig. 3 provides the classification performance of every connectivity network with the adoption of l_1-norm and l_2-norm regularizations. Specifically, with the employment of l_2-norm regularization, the accuracy improves from 67.12% to 69.86% by group constrained sparse network. With the proposed network, the classification accuracy increases from 73.97% to 75.34%. Moreover, the usage of l_1-norm regularization can further improve the classification performance. In detail, the L_1-RMLPC achieves an accuracy increase of at least 1.37% by using the Lasso and Pearson correlation based networks, and 4.11% by other networks. The combination of l_1-norm and l_2-norm regularization leads to a further improvement of classification accuracy and a maximum performance gain of 6.85% is obtained by the ultra-GC method based network. These results indicate that the l_1-norm and l_2-norm regularization potentially improves the generalizability of the MLPC.

E. Discriminative Features

To evaluate the contribution of different features to the classification, we calculate the feature importance scores with the trained weights of the EMLPC. In each layer of EMLPC, the weights in this layer indicates the contribution of different nodes of the previous layer. Thus a linear projection of the weights from the input layer to the last layer represents the importance of the input features for the classification. The feature importance scores are calculated as the inner product of weights matrixes across all the layers. For example, for a simple three layer perceptron with the number of nodes

TABLE III
TOP FIFTEEN ROIS SELECTED BY EMLPC

No.	Abbreviations	Full Names	Scores
14	IFGtriang.R	Right inferior frontal gyrus (triangular)	0.1034
88	TPOmid.R	Right temporal pole (middle)	0.1026
37	HIP.L	Left hippocampus	0.1003
76	PAL.R	Right pallidum	0.0893
71	CAU.L	Left caudate	0.0885
48	LING.R	Right lingual gyrus	0.0884
22	OLF.R	Right olfactory	0.0831
51	MOG.L	Left middle occipital gyrus	0.0802
33	MCG.L	Left middle cingulate gyrus	0.0779
85	MTG.L	Left middle temporal gyrus	0.0740
45	CUN.L	Left cuneus	0.0733
44	CAL.R	Right calcarine cortex	0.0723
18	ROL.R	Right rolandic operculum	0.0720
65	ANG.L	Left angular gyrus	0.0694
73	PUT.L	Left Putamen	0.0685

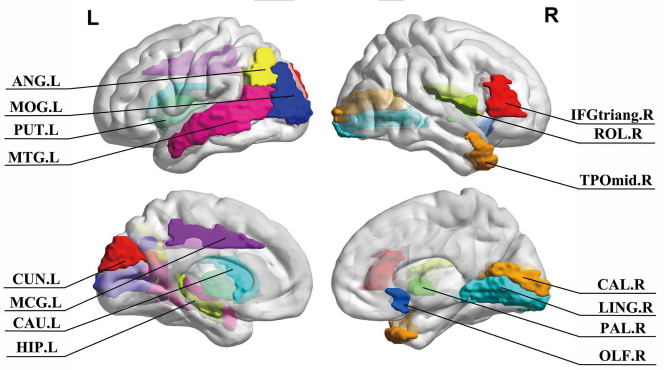

Fig. 5. The most discriminative regions selected via the EMLPC.

set as 270-50-1, the weight matrix of the middle layer is $W_2 \in R^{270\times50}$ and the weight matrix of the output layer is $W_3 \in R^{50\times1}$. The inner product of weights matrixes $S = W_2 \times W_3 \in R^{270\times1}$ represents the feature importance scores of the 270 input features. Features with higher scores indicate a higher capacity to discriminate between two groups. In addition, the final feature sets are selected by using the average feature importance scores. Namely, we calculate the feature importance scores in each outer cross-validation and then average them to get the final feature importance scores. All features are then sorted and selected based on the final feature importance scores.

As local efficiency coefficients and RCC are two different types of features, they are evaluated separately. Each local efficiency coefficient corresponds to a brain region, and the most significant fifteen brain regions are listed in Table III with their locations on the brain space are shown in Fig. 5. For rich club coefficients, each feature corresponds to a cutoff point on rich club coefficient curve. This point represents the density of connections among a group of brain regions with a degree greater than k. Noted that for a directional weighted connectivity, the degree is computed as the summation of its in-degree and out-degree values. It is clearly observed in Fig. 6(a) that the rich club coefficient generally decreases with an increase of k in both NCs and MCI groups. Compared with NCs, MCI patients show a relatively lower RCC which

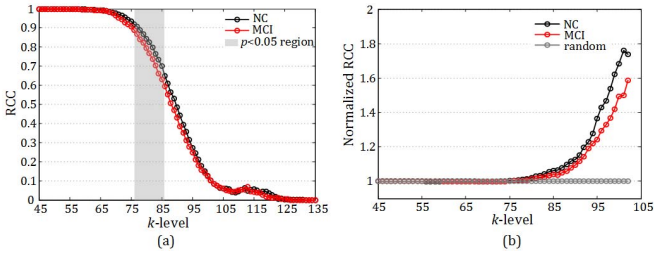

Fig. 6. Rich club coefficient curves (a) and normalized rich club coefficient curves (b) between MCI and NC.

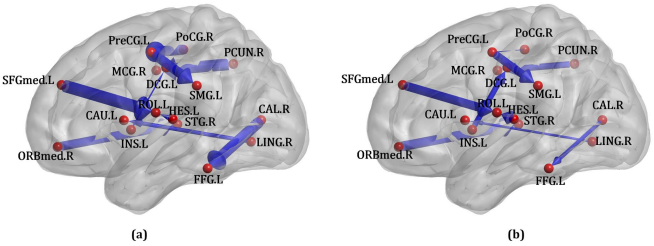

Fig. 7. Comparison of connectivity differences between NC (a) and MCI (b).

indicates the decreased density of connections. With the standard two sample t-test, a discriminative region is obtained ($p < 0.05$) as is shown in Fig 6(a). As a real-world network, the rich club coefficients of brain effective connectivity should be greater than the random networks. Thus, to further evaluate the rich club effect in brain connectivity network, by normalizing its rich club coefficients with the rich club coefficients computed from the randomized networks of equal size and similar connectivity distribution [35]. The normalized rich club coefficients of both MCI patients and NC subjects, as shown in Fig. 6(b), are larger than one, indicating the existence of rich club effect in the brain network. Importantly, the normalized rich club coefficients of MCI patients are lower than NC when the network degree threshold is larger than 75, suggesting that MCI patients suffer from a loss of rich club effect which indicates a less efficiency in the brain organizations and functions particularly the high-order organization.

F. Connectivity Analysis

To further evaluate the significant differences of the effective connectivity strength between ROIs and to visually show the differences in connectivity networks for MCI patients and NCs, a standard two-sample t-test is performed using the whole dataset. As topological properties are extracted as features for classification, besides at a connectivity level, we also evaluate the differences at a nodal level. Connections with $p < 0.005$ are listed in Table IV. There are totally sixteen ROIs involved in these connections and half of them are found to be consistent with the ROIs found by the EMLPC. Fig. 7 graphically illustrates the differences of the discriminative connections between MCI and NC (the thickness of edges indicates the strength of connections).

TABLE IV
THE MOST DISCRIMINATIVE CONNECTIONS

Selected ROIs	Direction of connectivity	Neighbors of selected ROIs	p-values
ROL.L	←	HES.L	0.0002
PCUN.R	→	MCG.R	0.0008
SMG.L	←	PreCG.L	0.0028
FFG.L	←	CAL.R	0.0029
INS.L	→	MCG.L	0.0036
PoCG.R	→	PreCG.L	0.0036
CAU.L	←	LING.R	0.0041
ORBmed.R	→	STG.R	0.0043
SFGmed.L	→	ROL.L	0.0046

Note: STG=Superior temporal gyrus; PreCG=Precentral gyrus; HES= Heschl gyrus; PCUN= Precuneus; MCG=Middle cingulate gyrus; SMG= Supramarginal gyrus; CAL= Calcarine; FFG=Fusiform gyrus; ROL=Rolandic operculum; ORBmed=Orbitofrontal cortex (medial); INS=Insula; PoCG=Postcentral gyrus; CAU= Caudate; LING=Lingual gyrus; SFGmed= Superior frontal gyrus (medial); L=Left; R=Right.

G. The Impact of Parameters Optimization

In this section, we explore the influences of different parameters of the proposed method including the model order q, the sparsity level λ, the l_1-norm regularization parameter α, the l_2-norm regularization parameter β, the number of hidden layers N_l, and the number of nodes N_n on the final classification performance.

Considering that q and λ are the parameters used for determining the characteristics of the connectivity network and are independent from the other four parameters, we design two independent experiments to evaluate the impact of these two parameters. Firstly, we calculate the BIC values of different model size q in the range of [0, 10]. The BIC curve is graphically illustrated in Fig. 8(a) where the minimum BIC value is achieved at $q = 1$. Low and consistent BIC values are obtained over a relatively wide range of q ($0 \leq q \leq 6$). This indicates that the proposed method is relatively robust to the model order q. Secondly, we evaluate the effect of sparsity level parameter λ on the classification performance while keeping other parameters constant (i.e., $q = 1$, $\alpha = 10^{-3}$, $\beta = 10^{-5}$, $N_l = 1$, $N_n = 50$) and the classification results are illustrated in Fig. 8(b). The classification accuracy changes smoothly with λ, implying the robustness of our proposed method with respect to the parameter λ. With the current dataset, the best sparsity level parameter is 0.031.

On the other hand, we perform two sets of experiments to investigate how the other four parameters jointly affect the prediction performance. As N_n and N_l determine the structure of EMLPC, and α and β determine the contribution of regularization terms, we evaluate these two pairs of parameters separately. Firstly, we evaluate the classification performance with varied model structures. Fig. 9(a) provides the accuracy with respect to the numbers of layers and nodes, while keeping the other parameters constant (i.e., $q = 1$, $\alpha = 10^{-3}$, $\beta = 10^{-5}$, $\lambda = 0.03$). As shown in Fig. 9(a), the optimal classification performance is achieved at $N_l = 1$ and $N_n = 50$. With a fixed N_l, the classification accuracy is

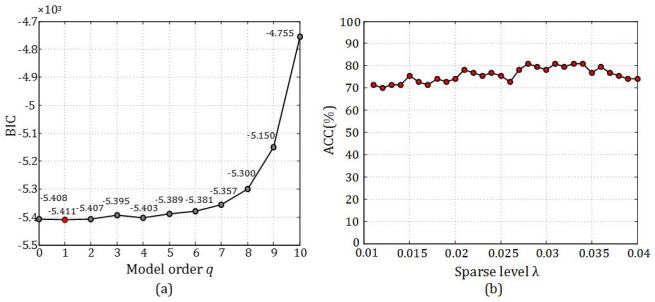

Fig. 8. Influence of parameters. (a) The BIC value for different model orders; (b) The impact of sparse level on final classification accuracy.

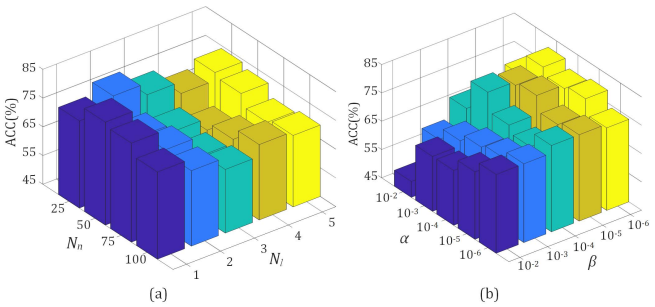

Fig. 9. (a) The classification performance with different the number of nodes and layers; (b) The classification performance with different l_1-norm and l_2-norm regularization parameters.

relatively stable with N_n, implying that our proposed method is not very sensitive to N_n. While the classification performance is predominantly affected by N_l, it is thus important to select the optimal N_l for EMLPC construction. It can be noticed that the classification accuracy decreased with the increase of the number of layers N_l, indicating that the complexity of the EMLPC with one hidden layer is capable of representing the model distribution of the current dataset. Secondly, two regularization parameters α and β balance the relative contributions of l_1-norm regularization and l_2-norm regularization terms. To evaluate the effects of regularization parameters, we explore different combinations of α and β values and their performances are provided in Fig. 9(b). The optimal performance is achieved at $\alpha = 10^{-3}$ and $\beta = 10^{-6}$. Moreover, with a fixed α, the classification accuracy of the proposed method varies smoothly with β. While with a fixed β, the classification performance varies greatly with α. These results imply that the proposed method is robust with respect to l_2-norm regularization parameter while the selection of l_1-norm regularization parameter significantly affect the classification performance. This is reasonable since the l_1-norm regularization parameter works as a sparsity control parameter and determines the scale of the selected feature subset for classification.

IV. Discussion

A. Significance of Results

In this paper, we proposed a novel ultra-group constrained sparse linear regression model for an effective connectivity inference, and a novel EMLPC was employed as the classifier for MCI identification. The classification performance was systematically evaluated under various conditions, including: 1) the usage of the ultra-group constrained structure detection algorithm; 2) the presence or absence of the UOFR algorithm to estimate the connectivity strengths; and 3) the presence or absence of l_1-norm regularization and l_2-norm regularization to constraint the weights of the classifier.

The key findings of our experiment results are summarized as follows: 1) the ultra-group constrained structure detection algorithm improves the discriminative power of the effective connectivity network; 2) the UOFR algorithm for connectivity strength estimation further enhances the discriminative power; 3) the l_1-norm and l_2-norm regularization increases the generalizability of the MLPC; 4) hierarchical features are learned from the weights of the EMLPC. The maximum classification accuracy (80.82%) was obtained using the proposed effective connectivity inference method with the EMLPC. The proposed framework outperforms the other network construction methods and conventional classification approaches on the same data set, indicating its superiority for MCI identification.

B. Efficacy of the Ultra-group Constrained Structure Detection Algorithm

The ultra-group constrained structure detection algorithm incorporates the dependent information of weak derivatives of original signals into linear regression model and thus enhances the noise-resistibility and robustness of the method [11]. Incorporating the weak derivatives information into group constrained sparse learning may reflect the useful information in the data that ultimately helps to detect the model structure more accurately. The ultra-group constrained structure detection algorithm involves two steps. First, the ultra-time series are built by adding the weak derivatives of original time series to the dataset. Then these ultra-time series data are feed into the group constrained sparse learning algorithm to detect the topological structure of effective network. As the ultra-time series includes not only the information of the datum points but also the dependent relationship information, the ultra-group constrained topology structure detection algorithm can characterize the between-ROI interactions more accurately. Also, the group constraint is employed to minimize the effect of inter-subject variability in network representation. By employing a group constraint across different subjects, connectivity networks of different subjects shares a common structure which is consistent with the knowledge that different subjects share a default mode connectivity pattern [36].

C. Efficacy of the UOFR Algorithm

The UOFR algorithm for connectivity strength estimation contributes to interpreting the effective connectivity and further improves the classification performance. With ultra-group constrained structure detection algorithm, the estimated coefficients are unscaled with some coefficients being negative, which leads to difficulty in interpreting and analyzing the effective network. Moreover, though the novel ultra-group constrained structure detection algorithm is efficient in picking up the most significant regressors, the shrinkage may produce

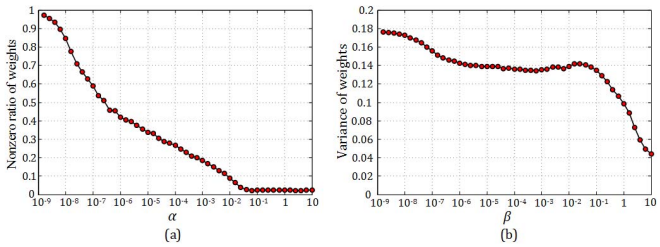

Fig. 10. The influence of regularizations. (a) The sparsity effect of l_1-norm regularization; (b) The variance reduction effect of l_2-norm regularization.

biased estimates in the risk of achieving a suboptimal result. To alleviate these limitations, with the effective connectivity structure identified by the non-zero coefficients estimated via a novel ultra-group constrained structure detection algorithm, we utilize an UOFR algorithm to estimated weights of the effective connectivity network. With the orthogonal process and a step forward regression procedure, connectivity parameters are scaled to the range of (0,1) by using the UERR criterion as the measurement of connectivity strength. As is demonstrated in Section III, with the UOFR algorithm for connectivity strength estimation, the discriminative connectivity power is greatly improved, and hence a classification accuracy increase is obtained.

D. Efficacy of Regularizations

Note that the control of the weight sparsity via the application of the l_1-norm and l_2-norm regularization can obviously improve the classification performance of the MLPC. The l_1-norm regularization and l_2-norm regularization behave differently in constraining the weights of the MLPC [37].

The l_1-norm regularization increases the sparsity of weights in MLPC, leading to a decrease of the nonzero ratio as shown in Fig. 10(a). The greater the l_1-norm regularization parameter, the fewer features will be selected for the MCI classification [38]. Different from l_1-norm regularization, the l_2-norm regularization leads to more average weights in MLPC shown in Fig. 10(b), which means the deviation of weights is decreased. With the l_2-norm regularization, the weights corresponding to different features are more averaged thus the classifier pays attention to more features [37]. For example, when two features are highly correlated, the classifier with l_1-norm regularization prefers to pick one of the two features by setting coefficient of the other to zero. In contrast, the l_2-norm regularization will keep both of them and jointly shrink the corresponding coefficients a little bit. By adjusting the parameters of the l_1-norm regularization and l_2-norm regularization, a trade-off between fitting ability to the training data and generalizability to unseen data is obtained. The best combination of parameters is automatically determined via grid search mentioned in Section II.

E. Interpretation of Discriminative Features

The discriminative features we found by linear projection across all layers are in line with previous studies [26], [27]. Based on the quantitative analysis in Section III, 15 brain regions are identified to be associated with MCI diseases, and a loss of rich club effect is observed in MCI groups. Local efficiency coefficients and rich club coefficients are extracted as original features to feed into the classifier. The brain regions identified with local efficiency coefficients include hippocampus, middle temporal gyrus, olfactory, and other brain regions which were founded to be associated with MCI/AD in previous researches [39], further justify the efficiency of the proposed method. Additionally, the un-normalized and normalized rich club coefficients are greater in healthy controls than in MCI patients, which is consistent with the hypothesis that the functional organization of the normal brain may be impaired by the MCI disease.

In addition, multi-scale features are fused by the EMLPC. Specifically, the local efficiency coefficient is a node level feature that measures the characteristic of brain regions separately, while rich club coefficient is a whole brain level feature which measures the characteristic of all brain regions. By the combination of nodal level features and whole brain level features, the proposed classifier can acquire a more discriminative power which leads to the final increase of classification accuracy.

F. The Most Discriminative Connections

The brain regions associated with the most discriminative connections that are identified based on group-based analysis are mostly overlapped with the brain regions that are selected by the EMLPC in the classification process, further validating the effectiveness of the proposed method in identifying disease-related brain regions. Furthermore, the identified regions mainly locates in frontal lobes (e.g., superior frontal gyrus (medial), Orbitofrontal cortex (medial) [40]), temporal lobes (e.g., heschl gyrus, superior temporal gyrus [40]), and limbic lobe (e.g., middle cingulate gyrus [41]), parietal lobe (e.g., supramarginal gyrus, precuneus [42]), in line with biological findings that MCI causes atrophy in the frontal and temporal lobes [43].

Noted that the connections in the current study are directed and weighted, which thus provides a more comprehensive characterization of the interactions within the whole-brain network. It is obvious from the connectivity analysis in section II that the effective connectivity for most of the optimal connections including left superior frontal gyrus (medial) to left rolandic operculum, from right calcarine to left fusiform gyrus, from right postcentral gyrus to left precentral gyrus, from left heschl gyrus to left rolandic operculum, are much weaker in MCI patients than that of NCs, implying that the neural interactions between these brain regions are deteriorated and the communication pathways between them are interrupted. In addition, the rolandic operculum always involves in the selected connections, indicating that it may play a significant role in information transformation for cognitive function [44]. However, more attention is required in future to better understand the roles of rolandic operculum in human cognitive function.

G. Limitation and Future Directions

Though achieving good MCI classification performance, the proposed method possesses two major limitations.

First, the selection of brain atlas may potentially influence the generalization performance of the classifier. The computed effective connectivity may be quite different by using different atlases with different levels. Another limitation is the size and type of the dataset. The dataset used in this study is a bit small compared with other datasets commonly used for machine learning method. Nevertheless, the obtained results do provide evidence on the efficacy of the proposed framework for identifying MCI patients from normal controls. In the future, we will validate the effectiveness of the proposed framework on the dataset with larger sample sizes.

V. Conclusion

In summary, we proposed a new effective connectivity network based classification framework for MCI identification, and achieve superior classification performance compared to the state-of-the-art methods. Experimental results illustrated that by imposing weak derivative of original time series into the group constrained sparse learning algorithm, the linear regression model can construct effective connectivity networks with greater discriminative power. In addition, the classification performance successfully validated the feasibility of the EMLPC classifier toward the automated diagnosis of MCI patients using the local efficiency coefficients and rich club coefficients as input patterns. The results of the proposed EMLPC method are consistent with the connectivity analysis results, providing additional evidence of disrupted network organization in MCI at a nodal level and whole brain level. The combination of the novel effective connectivity network inference method and the EMLPC classifier can be useful for providing a complementary objective opinion to the clinical diagnosis of cognitive impairment. Ultimately, the proposed method can benefit the development of computer-aided diagnosis tools in the clinical and pre-clinical settings.

References

[1] S. Gauthier *et al.*, "Mild cognitive impairment," *Lancet*, vol. 367, no. 9518, pp. 1262–1270, 2006.

[2] C. Misra, Y. Fan, and C. Davatzikos, "Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI," *NeuroImage*, vol. 44, no. 4, pp. 1415–1422, 2009.

[3] J. Bischkopf, A. Busse, and M. C. Angermeyer, "Mild cognitive impairment—A review of prevalence, incidence and outcome according to current approaches," *Acta Psychiatrica Scandinavia*, vol. 106, no. 6, pp. 403–414, 2002.

[4] B. Singh *et al.*, "A prospective study of chronic obstructive pulmonary disease and the risk for mild cognitive impairment," *JAMA Neurol*, vol. 71, no. 5, pp. 581–588, 2014.

[5] J. J. Hsiao and E. Teng, "Depressive symptoms in clinical and incipient Alzheimer's disease," *Neurodegenerative Disease Manage.*, vol. 3, no. 2, pp. 147–155, 2013.

[6] M. Bola and B. A. Sabel, "Dynamic reorganization of brain functional networks during cognition," *Neuroimage*, vol. 114, pp. 398–413, Jul. 2015.

[7] "Multimodal hyper-connectivity of functional networks using functionally-weighted LASSO for MCI classification," *Med. Image Anal.*, vol. 52, pp. 80–96, 2019.

[8] H. Lee, D. S. Lee, H. Kang, B.-N. Kim, and M. K. Chung, "Sparse brain network recovery under compressed sensing," *IEEE Trans. Med. Imag.*, vol. 30, no. 5, pp. 1154–1165, May 2011.

[9] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc., B (Statist. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.

[10] C.-Y. Wee, P.-T. Yap, D. Zhang, L. Wang, and D. Shen, "Group-constrained sparse fMRI connectivity modeling for mild cognitive impairment identification," *Brain Struct. Function*, vol. 219, no. 2, pp. 641–656, 2014.

[11] Y. Guo, L. Z. Guo, S. A. Billings, and H.-L. Wei, "Ultra-orthogonal forward regression algorithms for the identification of non-linear dynamic systems," *Neurocomputing*, vol. 173, pp. 715–723, Jan. 2016.

[12] B. Jie, C.-Y. Wee, D. Shen, and D. Zhang, "Hyper-connectivity of functional networks for brain disease diagnosis," *Med. Image Anal.*, vol. 32, pp. 84–100, Aug. 2016.

[13] R. A. Khan *et al.*, "fNIRS-based neurorobotic interface for gait rehabilitation," *J. Neuroeng. Rehabil.*, vol. 15, p. 7, Feb. 2018.

[14] J. L. Lancaster *et al.*, "Automated Talairach atlas labels for functional brain mapping," *Hum. Brain Mapping*, vol. 10, no. 3, pp. 120–131, Jul. 2000.

[15] S. R. Kesler *et al.*, "Predicting long-term cognitive outcome following breast cancer with pre-treatment resting state fMRI and random forest machine learning," *Front Hum. Neurosci.*, vol. 11, pp. 256–265, Nov. 2017.

[16] S. Farhan, M. A. Fahiem, and H. Tauseef, "An ensemble-of-classifiers based approach for early diagnosis of Alzheimer's disease: Classification using structural features of brain images," *Comput. Math. Method Med.*, vol. 2014, Aug. 2014, Art. no. 862307.

[17] C. Wu *et al.*, "Improving interpretability and regularization in deep learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 2, pp. 256–265, Feb. 2018.

[18] J. Ashburner, "SPM: A history," *Neuroimage*, vol. 62, no. 2, pp. 791–800, 2012.

[19] Y. Chao-Gan and Z. Yu-Feng, "DPARSF: A MATLAB toolbox for 'pipeline' data analysis of resting-state fMRI," *Front Syst. Neurosci.*, vol. 4, p. 13, May 2010.

[20] D. Shen and C. Davatzikos, "HAMMER: Hierarchical attribute matching mechanism for elastic registration," *IEEE Trans. Med. Imag.*, vol. 21, no. 11, pp. 1421–1439, Nov. 2002.

[21] L.-Z. Liu, F.-X. Wu, and W.-J. Zhang, "A group LASSO-based method for robustly inferring gene regulatory networks from multiple time-course datasets," *BMC Syst. Biol.*, vol. 8, no. 3, p. S1, 2014.

[22] Y. Li *et al.*, "Time-varying system identification using an ultra-orthogonal forward regression and multiwavelet basis functions with applications to EEG," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2960–2972, Jul. 2018.

[23] G. Deshpande and X. Hu, "Investigating effective brain connectivity from fMRI data: Past findings and current issues with reference to Granger causality analysis," *Brain Connectivity*, vol. 2, no. 5, pp. 235–245, 2012.

[24] J. Liu, S. Ji, and J. Ye, "SLEP: Sparse learning with efficient projections," Dept. Comput. Sci. Eng., Arizona State Univ., Tempe, AZ, USA, Tech. Rep., 2009, pp. 1–61, vol. 6, no. 491.

[25] L. Wang, Y. You, and H. Lian, "Convergence and sparsity of Lasso and group Lasso in high-dimensional generalized linear models," *Stat. Papers*, vol. 56, no. 3, pp. 819–828, 2015.

[26] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[27] T. Yan, W. Wang, L. Yang, K. Chen, R. Chen, and Y. Han, "Rich club disturbances of the human connectome from subjective cognitive decline to Alzheimer's disease," *Theranostics*, vol. 8, no. 12, pp. 3237–3255, 2018.

[28] T. Opsahl, V. Colizza, P. Panzarasa, and J. J. Ramasco, "Prominence and control: The weighted rich-club effect," *Phys. Rev. Lett.*, vol. 101, no. 16, p. 168702, 2008.

[29] Z. Wang, Y. Yuan, F. Bai, J. You, and Z. Zhang, "Altered topological patterns of brain networks in remitted late-onset depression: A resting-state fMRI study," *J. Clin. Psychiatry*, vol. 77, no. 1, pp. 123–130, 2016.

[30] K. T. E. O. Dubbelink *et al.*, "Disrupted brain network topology in Parkinson's disease: A longitudinal magnetoencephalography study," *Brain*, vol. 137, pp. 197–207, Jan. 2014.

[31] N. Jaitly and G. E. Hinton, "Using an autoencoder with deformable templates to discover features for automated speech recognition," in *Proc. INTERSPEECH*, vols. 1–5, 2013, pp. 1737–1740.

[32] M. E. Lynall *et al.*, "Functional connectivity and brain networks in schizophrenia," *J. Neurosci.*, vol. 30, no. 28, pp. 9477–9487, 2010.

[33] Y. Li, C.-Y. Wee, B. Jie, Z. Peng, and D. Shen, "Sparse multivariate autoregressive modeling for mild cognitive impairment classification," *Neuroinformatics*, vol. 12, no. 3, pp. 455–469, 2014.

[34] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, Jul. 2010.

[35] M. P. van den Heuvel and O. Sporns, "Rich-club organization of the human connectome," *J. Neurosci.*, vol. 31, no. 44, pp. 15775–15786, 2011.

[36] W. Li, X. Mai, and C. Liu, "The default mode network and social understanding of others: What do brain connectivity studies tell us," *Front Hum. Neurosci.*, vol. 8, p. 74, Feb. 2014.

[37] E. A. Smirnov, D. M. Timoshenko, and S. N. Andrianov, "Comparison of regularization methods for ImageNet classification with deep convolutional neural networks," *AASRI Proc.*, vol. 6, no. 1, pp. 89–94, 2014.

[38] Z. Wang, S. Gao, and L.-T. Chia, "Learning class-to-image distance via large margin and L1-norm regularization," in *Proc. ECCV*, vol. 7573, ~~2012,~~ pp. 230–244.

[39] M. M. Vasavada *et al.*, "Olfactory cortex degeneration in Alzheimer's disease and mild cognitive impairment," *J. Alzheimers Disease*, vol. 45, no. 3, pp. 947–958, 2015.

[40] A. Convit, J. Asis, M. J. de Leon, C. Y. Tarshish, S. De Santi, and H. Rusinek, "Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease," *Neurobiol. Aging*, vol. 21, no. 1, pp. 19–26, 2000.

[41] E. Guedj *et al.*, "Effects of medial temporal lobe degeneration on brain perfusion in amnestic MCI of AD type: Deafferentation and functional compensation?" *Eur. J. Nucl. Med. Mol. Imag.*, vol. 36, no. 7, pp. 1101–1112, 2009.

[42] M. Bailly *et al.*, "Precuneus and cingulate cortex atrophy and hypometabolism in patients with Alzheimer's disease and mild cognitive impairment: MRI and F-FDG PET quantitative analysis using freesurfer," *Biomed. Res. Int.*, ~~vol. 2015, May 2015,~~ Art. no. 583931.

[43] X. He *et al.*, "Changes in theta activities in the left posterior temporal region, left occipital region and right frontal region related to mild cognitive impairment in Parkinson's disease patients," *Int. J. Neurosci.*, vol. 127, no. 1, pp. 66–72, 2017.

[44] M. Cao *et al.*, "Early development of functional network segregation revealed by connectomic analysis of the preterm human brain," *Cerebral Cortex*, vol. 27, no. 3, pp. 1949–1963, 2017.

[45] Y. Li *et al.*, "Epileptic seizure detection in EEG signals using sparse multiscale radial basis function networks and the Fisher vector approach," *Knowl.-Based Syst.*, to be published, doi: 10.1016/j.knosys.2018.10.029.

Novel Effective Connectivity Inference Using Ultra-Group Constrained Orthogonal Forward Regression and Elastic Multilayer Perceptron Classifier for MCI Identification

Yang Li, Hao Yang, Baiying Lei, Jingyu Liu, and Chong-Yaw Wee

***Abstract*— Mild cognitive impairment (MCI) detection is important, such that appropriate interventions can be imposed to delay or prevent its progression to severe stages, including Alzheimer's disease (AD). Brain connectivity network inferred from the functional magnetic resonance imaging data has been prevalently used to identify the individuals with MCI/AD from the normal controls. The capability to detect the causal or effective connectivity is highly desirable for understanding directed functional interactions between brain regions and further helping the detection of MCI. In this paper, we proposed a novel sparse constrained effective connectivity inference method and an elastic multilayer perceptron classifier for MCI identification. Specifically, a ultra-group constrained structure detection algorithm is first designed to identify the parsimonious topology of the effective connectivity network, in which the weak derivatives of the observable data are considered. Second, based on the identified topology structure, an effective connectivity network is then constructed by using an ultra-orthogonal forward regression algorithm to minimize the shrinking effect of the group constraint-based method. Finally, the effective connectivity network is validated in MCI identification using an elastic multilayer perceptron classifier, which extracts lower to higher level information from initial input features and hence improves the classification performance. Relatively high classification accuracy is achieved by the proposed method when compared with the state-of-the-art classification methods. Furthermore, the network analysis results demonstrate that MCI patients suffer a rich club effect loss and have decreased connectivity among several brain regions. These findings suggest that the proposed method not only improves the classification performance but also successfully discovers critical disease-related neuroimaging biomarkers.**

***Index Terms*— Functional imaging, brain, computer-aided detection and diagnosis, connectivity analysis, machine learning.**

I. Introduction

MILD cognitive impairment (MCI) is identified as the clinical stage between normal forgetfulness and dementia, which commonly suffers from a cognitive decline that does not interfere notably with activities of daily living [1]. However, MCI is associated with increased risk of developing Alzheimer's disease (AD) at a rate of approximately 10%∼15% per year [2] compared with healthy controls who develop dementia at a rate of 1%∼2% per year [3]. By providing appropriate pharmacological treatments and behavioral interventions to MCI, it is possible to delay or prevent the progression of MCI to moderate and severe stages [4]. Thus, there is an urgent need for the accurate MCI detection and diagnosis. However, MCI is difficult to diagnose due to its mild symptoms, especially in high functioning individuals who maintain a positive public profile without showing apparent cognitive impairment [5]. Some anatomical and physiological results suggest that the cognitive process is substantially associated with interactions among brain regions [6].

Constructing brain connectivity from neuroimaging data holds great promise for diagnosing brain diseases and understanding the brain activity interactions. Many functional and effective connectivity modeling approaches based on functional magnetic resonance imaging (fMRI) have been proposed to distinguish diseases, e.g., AD and MCI, from normal controls (NCs) [7]. Among functional connectivity modeling methods, the correlation-based methods are commonly used for MCI identification and can obtain relatively high sensitivity for detecting potentially affected network connections. However, the spurious or insignificant connections in fully

Manuscript received August 28, 2018; revised October 27, 2018; accepted November 12, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61671042 and Grant 61403016 and in part by the Beijing Natural Science Foundation under Grant 4172037. *(Corresponding authors: Hao ~~Yao~~; Chong-Yaw Wee.)*

Y. Li is with the Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China, also with the Beijing Advanced Innovation Center for Big Date-based Precision Medicine, Beihang University, Beijing 100191, China, and also with the School of Automation Sciences and Electrical Engineering, Beihang University, Beijing 100191, China (e-mail: liyang@buaa.edu.cn).

H. Yang and J. Liu are with the School of Automation Sciences and Electrical Engineering, Beihang University, Beijing 100191, China (e-mail: hansyang@buaa.edu.cn).

B. Lei is with the Health Science Center, Shenzhen University, Shenzhen 518060, China (e-mail: leiby@szu.edu.cn).

C. Wee is with the Department of Biomedical Engineering, National University of Singapore, Singapore 117583 (e-mail: cywee2000@gmail.com).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2018.2882189

connected correlation-based network lead to difficult interpretation of the network structure. In order to address this problem, the least absolute shrinkage and selection operator (Lasso) method has been proposed for constructing the sparse connectivity network with only a few of significant connections [8]. The strength of insignificant or spurious connections is forced to be zero and hence make the interpretation of constructed sparse connectivity relatively easier. Meanwhile, since the Lasso method applies l_1-norm penalization at an individual level, the network topological structures vary between subjects [9]. This inter-subject variability inevitably cause comparison difficult and thus possibly degrades the classification performance. The group constrained sparse learning method thus enforces sparsity at the group level via $l_{2,1}$-norm penalization rather than at the level of the individual covariates. Through multi-task learning approach, it minimizes inter-subject variability and keeps the network topological structures identical among subjects [10]. However, the group constrained method above has its own weakness. For example, it heavily shrinks large regression coefficients which lead to the bias problem and it treats the datum points individually, omitting the interconnections among them [11]. These interconnections determine many critical characteristics of a constructed model and the absence of this information may lead to an over-fitting problem [11]. Moreover, the group constrained method focus only on constructing functional connectivity networks, generally ignoring the time-lagged relationships between different brain regions.

Recently, some machine learning approaches have been successfully deployed in the automated classification of connectivity networks related to MCI/AD [12]. These algorithms learn a relationship between the input data attributes and the target attribute by minimizing a loss function defined on the pairs of input data and the associated target attribute. Typically, a k nearest neighbors (KNN) stores all available cases and classifies new cases based on a similarity measure [13]. The random forest is an ensemble classification method that operates by constructing a host of decision trees at training time and outputting the class of the new cases as the mode of the classes of the individual trees [14]. However, these conventional methods only extract the low-level information from the fMRI data and thus cannot reveal the inherent and essential information of the connectivity network [15]. With the recent rejuvenation of neural networks, multilayer perceptron classifiers (MLPCs) have become a robust feature extraction method for MCI classification [16]. Different from the traditional classifiers, MLPCs extract the high-level information from the data by using model architectures of multiple nonlinear transformations. Though the stacked nonlinear transformation layers bring the MLPC excellent power to fit the training data, it also raises the concern of overfitting problem that the MLPC may perform excellently on the training data while works badly on test data, especially for small dataset [17].

To address these problems, in this paper, we propose a novel effective connectivity network modeling method and employed an elastic multilayer perceptron classifier (EMLPC) for MCI identification. Specifically, the effective network topological structure is first detected via an ultra-group constrained structure detection algorithm, which considers not only the discrepancy between observed original data and the model prediction but also the discrepancy between the associated weak derivatives. Thus it increases the noise- resistibility and robustness of the modeling and this helps to characterize the model more accurately. Second, based on the detected structure, an ultra-orthogonal forward regression (UOFR) algorithm is employed to estimate the strength of the effective connectivity networks. The UOFR algorithm alleviates the limitation of shrinking effect via a step forward regression process. Then several topological attributes are further extracted to characterize the effective connectivity networks. Following the feature extraction, an EMLPC is used as the classifier to identify MCI, where the EMLPC combines both the l_1-norm and l_2-norm regularization to obtain a trade-off between fitting ability to the training data and generalization ability on unknown data. The proposed framework is validated via a leave–one-out (LOO) cross-validation strategy for MCI identification. Experiment results illustrate that the proposed approach obtains a competitive performance with a relatively high classification accuracy compared to state-of-the-art methods, hence allows a more accurate detection of brain abnormalities which can useful for better interpretation of the pathological underpinnings of MCI.

II. Methodology

The proposed framework for diagnosing MCI essentially involves data acquisition, the effective connectivity network construction, feature extraction and classification, which is graphically shown in Fig. 1. Procedures of the proposed framework are summarized as follows: 1) Generate ultra-time series using the original regional mean time series and the weak derivatives of the fMRI data; 2) Detect the topology of effective connectivity networks via ultra-group constrained structure detection algorithm; 3) Apply an UOFR algorithm to estimate the effective connectivity strength; 4) Extract topological features from effective connectivity networks; 5) Split the dataset into train data and test data with a LOO cross-validation strategy, and optimize the parameters of the classifier via grid search with the train data; 6) With the optimal parameters, train a EMLPC and test it with the test data. The performance of the proposed framework is evaluated by the average classification accuracy of all the LOO loops.

A. Data Acquisition and Preprocessing

The present study involved 73 participants (36 MCI patients and 37 socio-demographically matched healthy controls) selected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (http://adni.loni.usc.edu/), and the subject IDs are provided in the supplementary material. Data acquisition was performed using a 3 Tesla Siemens scanner with the following parameters: flip angle $= 90°$; matrix size $= 64 \times 64$; voxel thickness $= 3.4$ mm; 197 volumes; 48 slices; TR $=$ 3000 ms and TE $= 30$ ms. Demographic and clinical information of the participants is provided in Table I. There is no difference between healthy subjects and MCI patients in terms of age, Mini-Mental State Examination (MMSE) score and head-motion measured with mean framewise displacement.

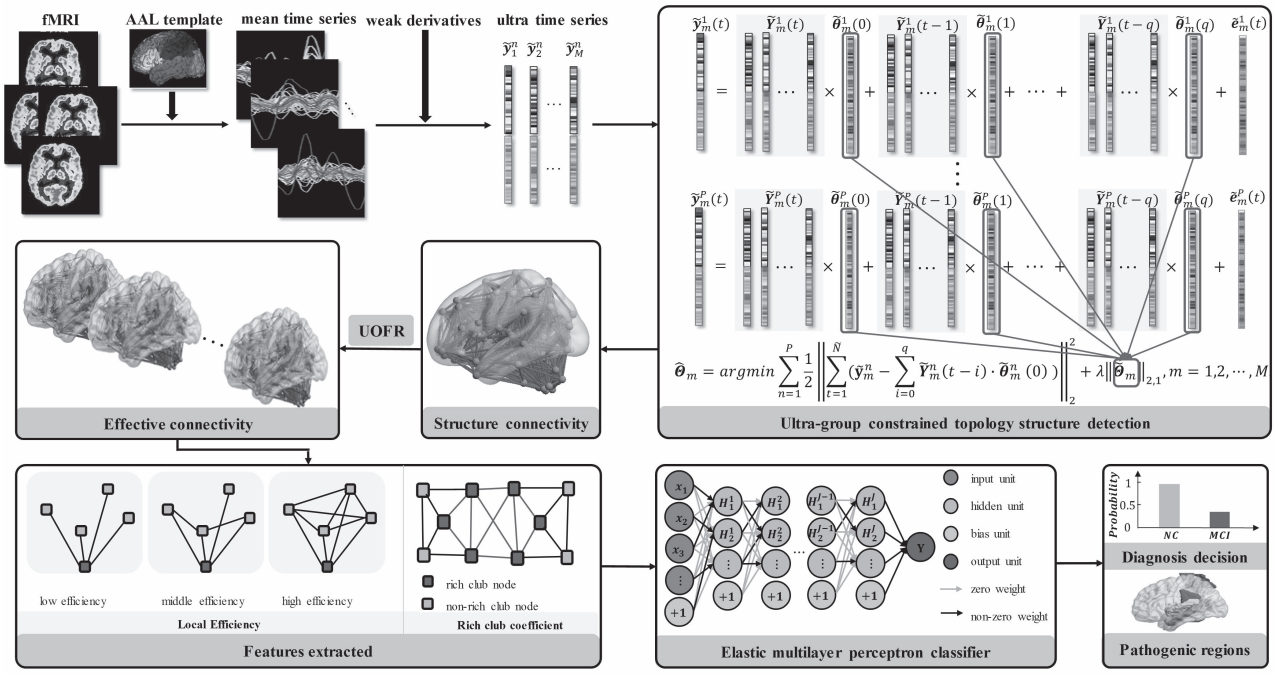

Fig. 1. The flowchart of the proposed method. Specifically, the ultra-mean time series are first computed. Then, a connectivity network topological structure detecting process and connectivity strength estimation process are performed. Finally, connectivity features are extraction and classifiers are trained and tested.

TABLE I
DEMOGRAPHIC AND CLINICAL INFORMATION OF THE PARTICIPANT

Group	MCI	NC	p-value
Male/Female	23/13	14/23	-
Age (mean±SD)	76.08±8.30	75.00±7.26	0.5560
MMSE (mean±SD)	27.31±3.65	28.52±2.83	0.1202
MFD (mean±SD)	0.14±0.08	0.14±0.09	0.9356

MMSE: Mini-mental state examination.
MFD: Mean framewise displacement.

Standard data preprocessing was performed using the Statistical Parametric Mapping 12 (SPM12) [18] software package and the Data Processing Assistant for Resting-State fMRI (DPARSF) toolbox [19]. Specifically, we used only the last 180 volumes of the acquired R-fMRI series for the following preprocessing steps to ensure magnetization equilibrium. The preprocessing of R-fMRI images included four main steps: 1) Slice timing correction: correct the remaining 180 volumes for the staggered order of slice acquisition during echo-planar scanning by matching all time points to intermediate time points; 2) Head motion correction: reduce the head-motion artifacts in the R-fMRI time-series by registering all volumes using to the first volume of the remaining R-fMRI time-series of in each subject as a reference to register all the following volume; 3) Spatial normalization: normalize the volumes to the Montreal Neurological Institute (MNI) atlas space and then resample the images to the voxel size of $3 \times 3 \times 3$ mm^3; 4) Spatial smoothing: smooth the dataset with a 4 mm full width half maximum Gaussian kernel. Afterward, we used the cerebrospinal fluid template and the white matter template to extract the values of the associated voxels, *i.e.*, the signal of whole brain is regressed against the average cerebrospinal fluid and white-matter signals as well as the six parameters from motion correction. The brain space was parcellated into 90 regions of interesting (ROIs) by warping the Automated Anatomical Labeling (AAL) template [14] to the subject space using the deformation fields estimated via a deformable registration method called HAMMER [20]. Finally, we obtained the mean time series of each ROI of each subject via averaging time series over gray matter voxels in that particular ROI followed by a band-pass filter of frequency interval (0.01 Hz $\leq f \leq$ 0.08 Hz) to minimize the effects of low-frequency drift and high-frequency noise.

B. Ultra-Group Constrained Topology Structure Detection

Inspired by associated works [8], [9], in our study, a sparse linear regression model is used to detect the structure of connectivity network. Sparse network construction methods, such as the Lasso and group constrained sparse learning method, have been employed to construct functional connectivity networks [21]. These methods use the information of individual datum point and discarding the relationship between them. However, the datum points of a dynamic system are time-dependent and connected with each other through the derivatives of the time continuous function. These interconnections determine many essential characteristics of a system and the absence of the interconnection information may lead to the overfitting problem of the identified model [22]. Moreover, the previous methods only measure the correlation relationship between the brain regions while ignoring their hysteresis effects [23]. Namely, the current brain region

activities are commonly influenced not only by the current activities of other regions but also by the previous activities [23]. To investigate the directional interactions among brain regions and to better understand the interaction mechanism within the brain networks, an ultra-group constrained structure detection algorithm is used to identify the topological structure of the effective connectivity networks.

The weak derivative information of the original fMRI time series in each ROI is first incorporated into the original time series to build the ultra-time series to extract dependent relation information of the associated fMRI time series. Since the original time series are discrete datum points and are not always differentiable, the signal is first smoothed with a test function and then the weak derivatives of original time series are calculated with the smoothed signals. For detailed derivation procedure, please refer to [22]. Suppose there are M ROIs for each subject and P subjects in total, with N being the number of time points, the mean time series of the m-th ROI for n-th subject is represented as $\boldsymbol{y}_m^n(t)$ $(t = 1, 2, \cdots, N)$. Given a normalized test function $\varphi(t)$, the l-th order weak derivative of $\boldsymbol{y}_m^n(\mathrm{t})$ can be defined as

$$\boldsymbol{y}_m^{n(l)}(t) = \sum\nolimits_{\tau=t}^{t+N_0} y_m^n(\tau)\varphi^{(l)}(\tau - t) \tag{1}$$

where N_0 is the support of the discrete test function with $t = 1, 2, \cdots, N - N_0$. The cubic B-spline basis function is employed as the test function to measure the agreement of data at a local level, and the first and second order derivatives of the smoothed signals are considered as in [11]. Therefore, the ultra-time series generated from y_m^n is defined as $\tilde{\boldsymbol{y}}_m^n = [y_m^n(1), \cdots, y_m^n(N), y_m^{n(1)}(1), \cdots, y_m^{n(1)}(N - N_0), y_m^{n(2)}(1), \ldots, y_m^{n(2)}(N - N_0)]^T$. The length of ultra-time series is $\tilde{N} = 3N - 2N_0$. Note that the ultra-time series $\tilde{\boldsymbol{y}}_m^n$ consist of two parts: the first part $[y_m^n(1), \cdots, y_m^n(N)]^T$ is the original time series that emphases the original datum points, while the second part $[y_m^{n(1)}(1), \cdots, y_m^{n(1)}(N-N_0), y_m^{n(2)}(1), \cdots, y_m^{n(2)}(N-N_0)]^T$ is the weak derivatives that essentially focuses on the dependent relationship of the associated original time series.

With the ultra-time series, the topological structure of the effective connectivity network is then identified via an ultra-group constrained structure detection algorithm which reveals the linear relationship between the current and previous ultra-time series and explores the underlying neuronal interactions. For the n-th subject, the m-th ROI activity $\tilde{\boldsymbol{y}}_m^n(t)$ is represented by a linear combination of current and previous activities of other ROIs, which is defined as

$$\tilde{\boldsymbol{y}}_m^n(t) = \sum\nolimits_{i=0}^{q} \tilde{\boldsymbol{Y}}_m^n(t-i) \cdot \tilde{\boldsymbol{\theta}}_m^n(i) + \tilde{\boldsymbol{e}}_m^n(t) \tag{2}$$

where q is the model order with $i = 0, 1, \cdots, q$, $\tilde{\boldsymbol{Y}}_m^n(t-i) = [\tilde{\mathbf{y}}_1^n(t-i), \tilde{\mathbf{y}}_2^n(t-i), \cdots, \tilde{\mathbf{y}}_{m-1}^n(t-i), \tilde{\mathbf{y}}_{m+1}^n(t-i), \cdots, \tilde{\mathbf{y}}_M^n(t-i)] \in R^{\tilde{N}\times(M-1)}$ with $m = 1, 2, \cdots, M$, $n = 1, 2, \cdots, P$, and $t = 1, 2, \cdots, \tilde{N}$, $\tilde{\boldsymbol{\theta}}_m^n(i) = [\theta_1^n(i), \theta_2^n(i), \cdots, \theta_{m-1}^n(i), \theta_{m+1}^n(i), \cdots, \theta_M^n(i)]^T$ is the coefficient vector and $\tilde{\boldsymbol{e}}_m^n(t)$ is the residual vector. The model (2) can be re-written in a linear regression form as

$$\tilde{\boldsymbol{y}}_m^n(t) = \tilde{\mathbf{A}}_m^n(t) \cdot \tilde{\boldsymbol{\theta}}_m^n + \tilde{\boldsymbol{e}}_m^n(t) \tag{3}$$

where $\tilde{\mathbf{A}}_m^n(t) = [\tilde{\boldsymbol{Y}}_m^n(t), \tilde{\boldsymbol{Y}}_m^n(t-1), \cdots, \tilde{\boldsymbol{Y}}_m^n(t-q)]$ denotes the regressor matrix with the dimension of $\tilde{N} \times ((M-1) \times (q+1))$ and $\tilde{\boldsymbol{\theta}}_m^n = \left[\tilde{\boldsymbol{\theta}}_m^n(0); \tilde{\boldsymbol{\theta}}_m^n(1); \cdots; \tilde{\boldsymbol{\theta}}_m^n(q)\right]$ is a column vector with $(M-1) \times (q+1)$ elements. The weights of the model are estimated by minimizing a group constrained objective function to reduce the inter-subject variability. The topological structure of different subjects are forced to be consistent via an additional l_2-norm penalization across all subjects. The object function of the ultra-group constrained model is represented as

$$J_{(\tilde{\boldsymbol{\Theta}}_m)} = \sum\nolimits_{n=1}^{P} \left(\frac{1}{2}\left\|\tilde{\boldsymbol{y}}_m^n - \tilde{\mathbf{A}}_m^n \cdot \tilde{\boldsymbol{\theta}}_m^n\right\|_2^2\right) + \lambda \left\|\tilde{\boldsymbol{\Theta}}_m\right\|_{2,1} \tag{4}$$

where $\tilde{\boldsymbol{\Theta}}_m = \left[\tilde{\boldsymbol{\theta}}_m^1, \tilde{\boldsymbol{\theta}}_m^2, \cdots, \tilde{\boldsymbol{\theta}}_m^P\right]$ and $\left\|\tilde{\boldsymbol{\Theta}}_m\right\|_{2,1}$ is the summation of l_2-norms of row vectors of $\tilde{\boldsymbol{\Theta}}_m$, λ is the regularization parameter that controls the 'sparsity' of the model. The regularization parameter λ forces certain coefficients to zero, effectively choosing a subset of features with non-zero coefficients. The weights associated with certain ultra-time series across different subjects are grouped together via the $l_{2,1}$-norms constraint. This constraint promotes a consistent connection topology among subjects and reduces inter-subject variability meanwhile allows variation of coefficient values between subjects [10]. Additionally, the coefficients of different time lagged ultra-time series of the same ROI are summed together, which are treated as an indicator on whether other ROIs have an influence on the currently considered ROI. Furthermore, the ROIs with non-zero coefficients are consistent for different subjects, and these ROIs are considered to have a connection with the target ROI. Then the associated ultra-time series of these ROIs are selected and arranged into the candidate regressor dictionary to be used for connectivity parameter estimation, while ultra-time series of other ROIs are discarded. The SLEP toolbox [24] was used to solve (4).

C. Connectivity Strength Estimation With UOFR Algorithm

The coefficients $\tilde{\boldsymbol{\Theta}}_m$ estimated via the ultra-group constrained (ultra-GC) method can simply be regarded as the effective connectivity strength (connection weights) between ROIs in an effective connectivity network. However, these estimated coefficients are unscaled and biased, and may lead to difficulty in interpreting and analyzing the effective network [25]. To eliminate the shrinking effect of the ultra-group constrained sparse learning model, we adopt an UOFR algorithm to estimate the effective connectivity strength. By employing the ultra-group constrained structure detection algorithm, a subset of ROIs with non-zero weights is considered to have a connection with the target ROI and the candidate dictionary is constructed by keeping ultra-time series of these ROIs while discarding the others.

In general, the connectivity strengths are estimated in a stepwise orthogonalized forward manner. The values of the error reduction ratio based ultra-time series (UERR.) are first calculated [22]. With $\tilde{N}$ as the length of ultra-time series, the ultra-time series of the target ROI can be expressed as $\tilde{\boldsymbol{y}}_m^n = [\tilde{y}_m^n(1), \tilde{y}_m^n(2), \cdots, \tilde{y}_m^n\left(\tilde{N}\right)]^T$ and the ultra-time series

of the candidate ROIs in the candidate dictionary can be defined as $\tilde{\boldsymbol{x}}_i = [\tilde{x}_i(1), \tilde{x}_i(2), \cdots, \tilde{x}_i\left(\tilde{N}\right)]^T$, the UERR is then calculated as

$$\begin{aligned} \mathrm{UERR}\left(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{y}}_m^n\right) &= \frac{\left((\tilde{\boldsymbol{x}}_i)^T \tilde{\boldsymbol{y}}_m^n\right)^2}{\left((\tilde{\boldsymbol{x}}_i)^T \tilde{\boldsymbol{x}}_i\right)\left((\tilde{\boldsymbol{y}}_m^n)^T \tilde{\boldsymbol{y}}_m^n\right)} \\ &= \frac{\left(\sum_{t=1}^{\tilde{N}} \tilde{x}_i(t)\tilde{y}_m^n(t)\right)^2}{\sum_{t=1}^{\tilde{N}} \left(\tilde{x}_i(t)\right)^2 \sum_{t=1}^{\tilde{N}} \left(\tilde{y}_m^n(t)\right)^2} \end{aligned} \tag{5}$$

where 'T' is transpose of the matrix or vector. Then the connectivity strength between the target ROI and the candidate ROIs are estimated using the commonly used standard orthogonal forward regression algorithm [45]. The UERR values corresponding to the same ROIs are summed and arranged into the effective connectivity matrix which contains every possible effective connectivity of ROIs pairs. The weights in the connectivity matrix can be interpreted as the influence that one ROI has upon another. The details of the effective connectivity construction process are summarized in supplementary material.

D. Feature Extraction

Feature extraction is a special dimensionality reduction approach in machine learning, which projects a high dimensional vector onto low dimensional vector to avoid the curse of dimensionality.

The local efficiency (LE) is a measure that quantifies the cliquishness of the nodes in the network, which indicates the contribution of a node to the communication of the network [26]. For a weighted and directed graph, the local efficiency of the m-th node is defined as [26]

$$\begin{aligned} &E_{loc}(m) \\ &= \frac{\sum_{i,j\in G_m} (w_{mi}^{\frac{1}{3}} + w_{im}^{\frac{1}{3}})(w_{mj}^{\frac{1}{3}} + w_{jm}^{\frac{1}{3}})(l_{ji}^{-1} + l_{ji}^{-1})/2}{\left(\sum_{i\in G_m} \delta_{mi}\right)^2 - \sum_{i\in G_m} \delta_{mi}^2} \end{aligned} \tag{6}$$

where G_m is the subgraph that contains only neighbors of the m-th node, w and δ are the connection weights and the number of non-zero connections, l is the length of the shortest directed path in G_m, respectively. The weighted local efficiency in (6) distinguishes the influence of different paths based on connection weights of the associated neighbors to the node.

Additionally, recent studies have shown that the presence and absence of the rich club organization reveals the information of the higher-order structure of a connectivity network [27]. The rich club coefficients (RCC) of the network are the ratio of the number of connections among nodes of the degree k or higher versus the total possible number of connections. For weighted networks, it is modified as the fraction of edge weights that connect nodes of degree k or higher out of the maximum edge weights that such nodes might share [28]. Formally, the RCC at level k is given by [28]

$$RCC(k) = \frac{\sum_{i=1}^{E_k} w_i(k)}{\sum_{i=1}^{E_k} w_i^r} \tag{7}$$

where $w_i(k)$ is the weight among nodes of degree k or higher, E_k indicates the number of links among these nodes, and w_i^r is the ranked connections of the whole network, respectively.

Topological attributes above are extracted from the effective connectivity network for each subject. These topological properties quantify connectivity profiles associated with individual network elements (such as nodes or links). Previous studies revealed that the disruption in interactions between brain subnetworks or regions as characterized by topology measures may lead to less efficient information processing and cognitive deficits [29]. Since the deteriorating functional connectivity and cognitive performance are associated with the changes in network topology, the usage of these topology measures as features may potentially boost the classification performance [27], [30]. The feature extraction process was performed using brain connectivity toolbox [26].

E. Classification

In this study, we adopt a MLPC for MCI classification. The MLPC is a kind of deep learning algorithm that models high-level representations in the data by using model architectures of multiple nonlinear transformations. The general idea of the MLPC algorithm is stacking up the nonlinear transformation layers. The more layers the data goes through within the network, the more complicated are the nonlinear transformations and the more abstract is the information extracted from the network. The nonlinear transformations enhance the fitting ability of the MLPC while increasing the risk of over-fitting. Thus, a kernel regularization scheme is employed to constrain the weights of the MLPC and improve its generalizability.

The l_1-norm and l_2-norm regularizations are added to the objective function of the MLPC to reduce the risk of over-fitting and improve the generalization performance of the model. The objective function of the EMLPC with a weight optimization process is defined as

$$\begin{aligned} J_{(W)} = \frac{1}{2Q}\sum_{n=1}^{Q} \left\| p^{\{n\}}(W) - t^{\{n\}} \right\|^2 + \alpha \sum_{i=1}^{N_l} \|W_i\|_1 \\ + \beta \sum_{i=1}^{N_l} \|W_i\|_2 \end{aligned} \tag{8}$$

where $p^{\{n\}}(W)$ is the predicted label of the subject n with weight W, $t^{\{n\}}$ is the true label of the subject n, Q is the number of training subjects, W_i is the weights of the i-th layer, α is the sparse parameter, β is the weight decay parameter and N_l is the total number of layers. The equation (8) consists of three parts: mean square error, sparsity punishment, and weight decay. The first term is a mean square error that measures the residual between the prediction and the true label. The second term is a sparsity penalty which is defined as the summation of l_1-norm of weights. Particularly, the sparsity punishment constraints the weights of the classifier to be sparse such that only a few non-zero weights. Significant features are automatically selected while spurious features are eliminated, thus the noise-resist ability of the classifier is enhanced. The final term represents weight decay which decreases the range of the weights via a l_2-norm regularization. The use of l_1-norm and l_2-norm regularization rules in Equation (8) decreases the

complexity of the model and increases the generalizability of the model.

The structure of EMLPC is determined through grid search in the inner LOO cross-validation. Specifically, the number of layer N_l is searched within the range of [1, 5] with a step of 1 while the number of hidden nodes N_n is searched within the range of [25, 100] with a step of 25. To simplify the model structure and reduce the number of free parameters, the number of nodes in each hidden layer of the EMLPC are set to the same. Secondly, we use l_1-norm regularization parameter α and l_2-norm regularization β to regularize the model to adjust the fitting ability and generalizability without dropout layer in our model. We set the EMLPC with different α (from 10^{-6} to 10^{-2}) and β (from 10^{-6} to 10^{-2}). Thirdly, the weights of the model are initialized close to zeros with a random uniform distribution in the range of $[-0.05, 0.05]$ and the loss function is set as the regularized mean square error. The combination of parameters that produces the best performance on the validation set will be used to construct the MCI diagnosis model.

With the optimal model structure, the hierarchical features are extracted and the MCI diagnosis models are trained. The training process includes two steps: an unsupervised pre- training step to train the network layer by layer via stacked autoencoder [31], and a supervised fine-tuning step to boost the performance by incorporating the label information. The pre-training step is used to greedily extract latent representation of the input data while reducing the feature dimensionality. Firstly, we train the first layer of the model with an autoencoder which nonlinearly combines the data from the input layer into a short code in the middle layer, and then decompress it into representation that closely matches the original data. Taking the input data as low-level features, the output of the middle layer is a compressed representation of the input data which we consider it relatively high-level features. Then, we take these features as the input of the next layer and another autoencoder is trained to further compress the data and extract higher level features. This process is executed recurrently for each hidden layer of the EMLPC and the outputs of different layers are indeed the hierarchical features from low-level features (the input layer) to high-level features (the output layer). Finally, the whole network is fine-tuned via RMSProp algorithm with a learning rate of 10^{-3} to minimize the difference between the true label and the prediction label of the EMLPC model. The fine-tuning step aims to slightly alter the weights to adjust the boundaries between different groups. The detailed training process is graphically shown in the supplementary material.

F. Parameters Optimization

In the proposed network inference and classification scheme, several free parameters, such as the model order q, sparsity level λ, l_1-norm regularization parameter α, l_2-norm regularization parameter β, the number of layers N_l and the number of nodes N_n of the neural network, should be optimized for achieving the best classification performance. In this study, the model size q is determined by minimizing the Bayesian information criteria (BIC) [33] for each sparsity level and the other five parameters are optimized via a LOO cross-validation scheme.

The model size q is the number of past samples which are needed to accurately predict the present data. Particularly, if the model size q is set to zero, the model simply calculates the partial correlation of current measurements from different brain regions. Thus, the proposed method is capable of evaluating the time-lagged relationship between different brain regions only if the model order is greater than zero. The optimal model size q can be determined based on the BIC value [33]

$$BIC\,(q) = \tilde{N} \log\left[mse\,(q)\right] + q \log\left(\tilde{N}\right) \tag{9}$$

where $\tilde{N}$ is the length of ultra-time series, mse is the mean square residuals for all subjects and ROIs. By minimizing BIC value, we search for the optimal model order q in the range of $[q_{min}, q_{max}]$. Following the previous study [33], q_{min} is set to 0 and q_{max} is set to 10.

Additionally, an inner LOO cross-validation scheme is used to optimize other hyper-parameters. It should be noted that the classification performance of outer LOO loops can be used to estimate the hyper parameters of a model and then those hyper-parameters are applied to fit a model to the whole dataset, which is likely to be optimistically biased and lead to the concern of over-fitting [34]. The main reason is that part of the model (the hyper-parameters) have been selected to optimize the final classification performance, so if the LOO statistic has a non-zero variance, there is the possibility of over-fitting the model selection criterion [34]. In order to choose the hyper- parameters and estimate the performance of the resulting model in an unbiased manner, we need to perform an inner LOO cross-validation to determine the hyper-parameters, with the outer LOO used to assess the performance of the model built based on the selected hyper-parameters on the unseen testing subject.

Specifically, suppose P subjects are involved in the study, a subject is first left out for testing and the remaining $P-1$ subjects are used to search for the optimal parameters. Among $P-1$ training subjects, one more subject is left out and the remaining $P-2$ subjects are used to construct a classifier based on a specific set of parameter values. The constructed classifier is then used to predict the label of the second left out subject. This process is repeated for $P-1$ times, each time a different subject is left out from the $P-1$ training subjects, to predict the labels of all $P-1$ training subjects and hence the inner LOO cross-validation accuracy. The process is repeated for all potential sets of parameters to determine the optimal set of parameters that gives the best classification accuracy based on $P-1$ training subjects. To fully utilize the information, an EMLPC is constructed with the identified optimal parameters using all $P-1$ training subjects and then used to predict the label of the first left out testing subject. This process is repeated P times, each time leaving out a different subject, to predict the labels of all P subjects and then obtain the overall cross-validation accuracy. Note that the nested cross-validation scheme is an unbiased

and robust evaluation method that can minimize the risk of overfitting [34]. Therefore, the classification model is reliable and competent to be applied for different datasets. We report only the results of overall cross-validation performance in this study. The classification and parameter optimization scheme is graphically illustrated in the supplementary material.

III. Experiment Results

A. Overview of Classification Performance

Several evaluation metrics such as the accuracy, sensitivity, specificity and balanced accuracy are adopted to evaluate the classification performance. We compare our proposed method with several other connectivity inference methods for MCI classification on the same dataset, which is shown in Table II. From Table II, the proposed method outperforms the competing state-of-the-art approaches. The classification accuracy by the proposed scheme of 80.82% is the highest among the competing methods, indicating its excellent ability to identify MCI patients from normal controls. In addition, the proposed method achieves a sensitivity of 80.56%, specificity of 81.08%, also the highest among all competing methods. Note that sensitivity and specificity are usually combined into a single measure as balanced accuracy, which is the arithmetic mean of the specification and sensitivity. The balanced accuracy is more accurate and stable for the performance evaluation particularly for imbalanced dataset. Concerning the balanced accuracy of 80.82%, the proposed method also outperforms all other methods, which indicates the excellent diagnostic power of the proposed method.

TABLE II
Comparison of Classification Performance

Method	Classifier	ACC(%)	SEN(%)	SPE(%)	BAC(%)
Pearson Correlation	KNN	54.79	53.66	56.25	54.95
	RF	60.27	58.97	61.76	60.37
	MLPC	64.38	64.71	64.10	64.40
	L_1-RMLPC	65.75	63.41	68.75	66.08
	L_2-RMLPC	64.38	62.50	66.67	64.58
	EMLPC	65.75	68.97	63.64	66.30
Lasso	KNN	60.27	60.00	60.53	60.26
	RF	65.75	62.22	71.43	66.83
	MLPC	65.75	67.74	64.29	66.01
	L_1-RMLPC	67.12	70.00	65.12	67.56
	L_2-RMLPC	65.75	70.37	63.04	66.71
	EMLPC	68.49	69.70	67.50	68.60
Group constrained method	KNN	60.27	58.54	62.50	60.52
	RF	63.01	64.52	61.90	63.21
	MLPC	67.12	68.75	65.85	67.30
	L_1-RMLPC	71.23	69.23	73.53	71.38
	L_2-RMLPC	69.86	69.44	70.27	69.86
	EMLPC	72.60	76.67	69.77	73.22
Ultra-group constrained method	KNN	65.75	65.71	65.79	65.75
	RF	64.38	64.71	64.10	64.40
	MLPC	69.86	68.42	71.43	69.92
	L_1-RMLPC	73.97	75.76	72.50	74.13
	L_2-RMLPC	69.86	69.44	70.27	69.86
	EMLPC	79.45	80.00	78.95	79.47
Proposed method	KNN	73.97	77.42	71.43	74.42
	RF	75.34	76.47	74.36	75.41
	MLPC	73.97	74.29	73.68	73.98
	L_1-RMLPC	78.08	76.32	80.00	78.16
	L_2-RMLPC	75.34	71.43	80.65	76.04
	EMLPC	**80.82**	**80.56**	**81.08**	**80.82**

where bold fonts indicate the best results.

B. Classification Performance by Ultra-Group Constraint

We evaluate the performance of different sparsity constraints used for constructing effective connectivity networks. In Fig. 2(a), the classification performance with ultra-GC method outperforms that of the classical group constrained sparse learning method using six different classifiers. Specifically, using the EMLPC as the classifier, the ultra-GC method receives an accuracy of 79.45%, which is 6.85% higher than that of the conventional group constrained learning method. Additionally, compared to the classical group constrained sparse learning based network, the ultra-GC approach improves the classification accuracy from 60.27% to 65.75% in KNN classifier, from 63.01% to 63.38% in RF classifier, from 71.23% to 73.97% in MLPC. These results indicate that the incorporation of the dependent information of datum points into connectivity network construction can indeed improve the performance of clinical identification of MCI.

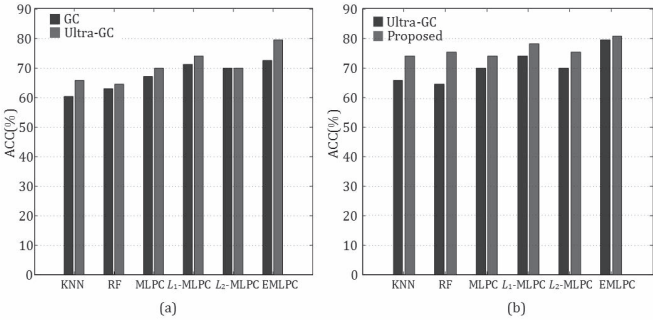

Fig. 2. The comparison of classification performance by (a) the GC and ultra-GC method; (b) the Ultra-GC and the proposed method.

C. Classification Performance Depending on the UOFR Algorithm

The UOFR algorithm is used to eliminate the shrinking effect of the conventional group constrained (GC) method and thus may lead to considerable improvement for MCI classification. To evaluate the effectiveness of the UOFR algorithm, a comparison of the proposed method and the ultra-GC method is graphically shown in Fig. 2(b). In detail, using traditional classification methods such as KNN and RF, an accuracy increase of 8.22% and 10.96% is achieved, respectively. With the deep architecture, the improvement of the performance classification is relatively moderate. For example, when the naive MLPC, L_1-RMLPCN, L_2-RMLPC and EMLPC are used to classify MCI, the classification performance gains of 4.11%, 4.11%, 5.48%, 1.37% are obtained respectively. The improvement of classification accuracy indicates that the step forward regression process of the UOFR algorithm indeed helps to overcome the limitation of the shrinking effect in the group constrained sparse learning method.

D. The Influence of Regularizations

To demonstrate the advantage of regularizations, we compare the classification performance of the MLPCs using varies

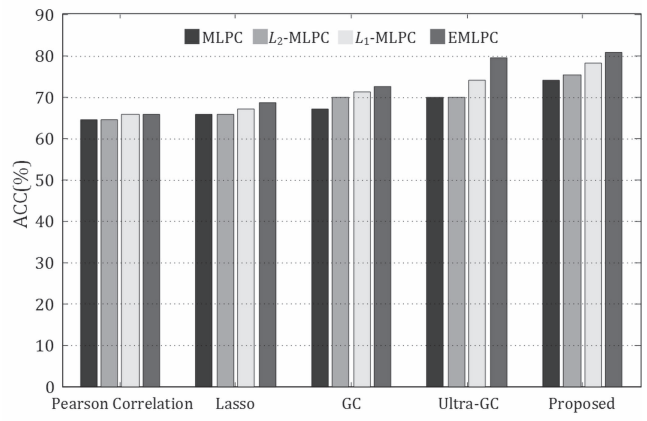

Fig. 3. The comparison performance by different regularizations.

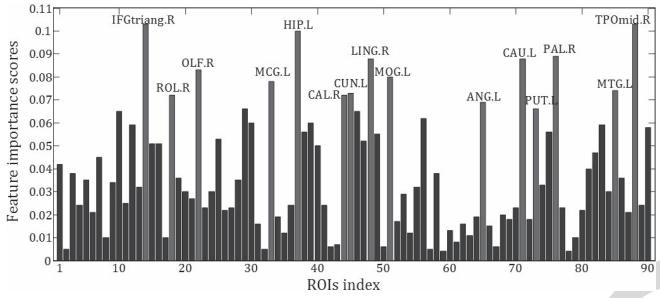

Fig. 4. The importance scores of different brain regions.

regularization strategies. Fig. 3 provides the classification performance of every connectivity network with the adoption of l_1-norm and l_2-norm regularizations. Specifically, with the employment of l_2-norm regularization, the accuracy improves from 67.12% to 69.86% by group constrained sparse network. With the proposed network, the classification accuracy increases from 73.97% to 75.34%. Moreover, the usage of l_1-norm regularization can further improve the classification performance. In detail, the L_1-RMLPC achieves an accuracy increase of at least 1.37% by using the Lasso and Pearson correlation based networks, and 4.11% by other networks. The combination of l_1-norm and l_2-norm regularization leads to a further improvement of classification accuracy and a maximum performance gain of 6.85% is obtained by the ultra-GC method based network. These results indicate that the l_1-norm and l_2-norm regularization potentially improves the generalizability of the MLPC.

E. Discriminative Features

To evaluate the contribution of different features to the classification, we calculate the feature importance scores with the trained weights of the EMLPC. In each layer of EMLPC, the weights in this layer indicates the contribution of different nodes of the previous layer. Thus a linear projection of the weights from the input layer to the last layer represents the importance of the input features for the classification. The feature importance scores are calculated as the inner product of weights matrixes across all the layers. For example, for a simple three layer perceptron with the number of nodes

TABLE III
TOP FIFTEEN ROIS SELECTED BY EMLPC

No.	Abbreviations	Full Names	Scores
14	IFGtriang.R	Right inferior frontal gyrus (triangular)	0.1034
88	TPOmid.R	Right temporal pole (middle)	0.1026
37	HIP.L	Left hippocampus	0.1003
76	PAL.R	Right pallidum	0.0893
71	CAU.L	Left caudate	0.0885
48	LING.R	Right lingual gyrus	0.0884
22	OLF.R	Right olfactory	0.0831
51	MOG.L	Left middle occipital gyrus	0.0802
33	MCG.L	Left middle cingulate gyrus	0.0779
85	MTG.L	Left middle temporal gyrus	0.0740
45	CUN.L	Left cuneus	0.0733
44	CAL.R	Right calcarine cortex	0.0723
18	ROL.R	Right rolandic operculum	0.0720
65	ANG.L	Left angular gyrus	0.0694
73	PUT.L	Left Putamen	0.0685

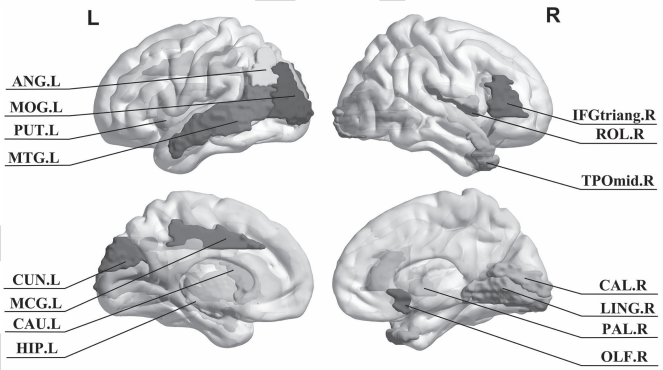

Fig. 5. The most discriminative regions selected via the EMLPC.

set as 270-50-1, the weight matrix of the middle layer is $W_2 \in R^{270\times50}$ and the weight matrix of the output layer is $W_3 \in R^{50\times1}$. The inner product of weights matrixes $S = W_2 \times W_3 \in R^{270\times1}$ represents the feature importance scores of the 270 input features. Features with higher scores indicate a higher capacity to discriminate between two groups. In addition, the final feature sets are selected by using the average feature importance scores. Namely, we calculate the feature importance scores in each outer cross-validation and then average them to get the final feature importance scores. All features are then sorted and selected based on the final feature importance scores.

As local efficiency coefficients and RCC are two different types of features, they are evaluated separately. Each local efficiency coefficient corresponds to a brain region, and the most significant fifteen brain regions are listed in Table III with their locations on the brain space are shown in Fig. 5. For rich club coefficients, each feature corresponds to a cutoff point on rich club coefficient curve. This point represents the density of connections among a group of brain regions with a degree greater than k. Noted that for a directional weighted connectivity, the degree is computed as the summation of its in-degree and out-degree values. It is clearly observed in Fig. 6(a) that the rich club coefficient generally decreases with an increase of k in both NCs and MCI groups. Compared with NCs, MCI patients show a relatively lower RCC which

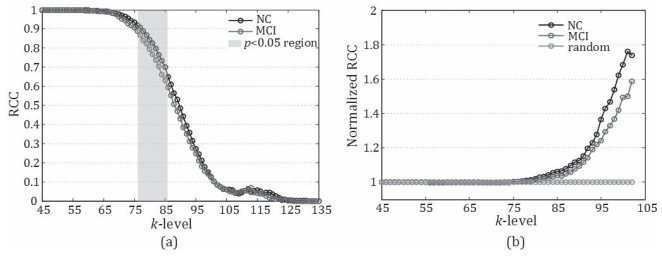

Fig. 6. Rich club coefficient curves (a) and normalized rich club coefficient curves (b) between MCI and NC.

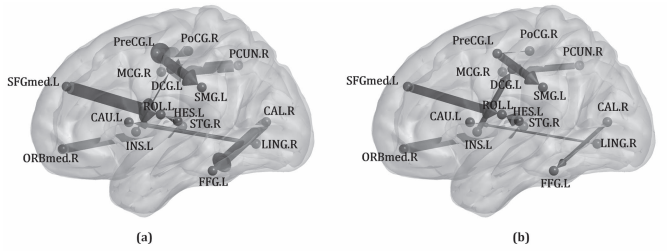

Fig. 7. Comparison of connectivity differences between NC (a) and MCI (b).

indicates the decreased density of connections. With the standard two sample t-test, a discriminative region is obtained ($p < 0.05$) as is shown in Fig 6(a). As a real-world network, the rich club coefficients of brain effective connectivity should be greater than the random networks. Thus, to further evaluate the rich club effect in brain connectivity network, by normalizing its rich club coefficients with the rich club coefficients computed from the randomized networks of equal size and similar connectivity distribution [35]. The normalized rich club coefficients of both MCI patients and NC subjects, as shown in Fig. 6(b), are larger than one, indicating the existence of rich club effect in the brain network. Importantly, the normalized rich club coefficients of MCI patients are lower than NC when the network degree threshold is larger than 75, suggesting that MCI patients suffer from a loss of rich club effect which indicates a less efficiency in the brain organizations and functions particularly the high-order organization.

F. Connectivity Analysis

To further evaluate the significant differences of the effective connectivity strength between ROIs and to visually show the differences in connectivity networks for MCI patients and NCs, a standard two-sample t-test is performed using the whole dataset. As topological properties are extracted as features for classification, besides at a connectivity level, we also evaluate the differences at a nodal level. Connections with $p < 0.005$ are listed in Table IV. There are totally sixteen ROIs involved in these connections and half of them are found to be consistent with the ROIs found by the EMLPC. Fig. 7 graphically illustrates the differences of the discriminative connections between MCI and NC (the thickness of edges indicates the strength of connections).

TABLE IV
THE MOST DISCRIMINATIVE CONNECTIONS

Selected ROIs	Direction of connectivity	Neighbors of selected ROIs	p-values
ROL.L	←	HES.L	0.0002
PCUN.R	→	MCG.R	0.0008
SMG.L	←	PreCG.L	0.0028
FFG.L	←	CAL.R	0.0029
INS.L	→	MCG.L	0.0036
PoCG.R	→	PreCG.L	0.0036
CAU.L	←	LING.R	0.0041
ORBmed.R	→	STG.R	0.0043
SFGmed.L	→	ROL.L	0.0046

Note: STG=Superior temporal gyrus; PreCG=Precentral gyrus; HES= Heschl gyrus; PCUN= Precuneus; MCG=Middle cingulate gyrus; SMG= Supramarginal gyrus; CAL= Calcarine; FFG=Fusiform gyrus; ROL=Rolandic operculum; ORBmed=Orbitofrontal cortex (medial); INS=Insula; PoCG=Postcentral gyrus; CAU= Caudate; LING=Lingual gyrus; SFGmed= Superior frontal gyrus (medial); L=Left; R=Right.

G. The Impact of Parameters Optimization

In this section, we explore the influences of different parameters of the proposed method including the model order q, the sparsity level λ, the l_1-norm regularization parameter α, the l_2-norm regularization parameter β, the number of hidden layers N_l, and the number of nodes N_n on the final classification performance.

Considering that q and λ are the parameters used for determining the characteristics of the connectivity network and are independent from the other four parameters, we design two independent experiments to evaluate the impact of these two parameters. Firstly, we calculate the BIC values of different model size q in the range of [0, 10]. The BIC curve is graphically illustrated in Fig. 8(a) where the minimum BIC value is achieved at $q = 1$. Low and consistent BIC values are obtained over a relatively wide range of q ($0 \leq q \leq 6$). This indicates that the proposed method is relatively robust to the model order q. Secondly, we evaluate the effect of sparsity level parameter λ on the classification performance while keeping other parameters constant (i.e., $q = 1$, $\alpha = 10^{-3}$, $\beta = 10^{-5}$, $N_l = 1$, $N_n = 50$) and the classification results are illustrated in Fig. 8(b). The classification accuracy changes smoothly with λ, implying the robustness of our proposed method with respect to the parameter λ. With the current dataset, the best sparsity level parameter is 0.031.

On the other hand, we perform two sets of experiments to investigate how the other four parameters jointly affect the prediction performance. As N_n and N_l determine the structure of EMLPC, and α and β determine the contribution of regularization terms, we evaluate these two pairs of parameters separately. Firstly, we evaluate the classification performance with varied model structures. Fig. 9(a) provides the accuracy with respect to the numbers of layers and nodes, while keeping the other parameters constant (i.e., $q = 1$, $\alpha = 10^{-3}$, $\beta = 10^{-5}$, $\lambda = 0.03$). As shown in Fig. 9(a), the optimal classification performance is achieved at $N_l = 1$ and $N_n = 50$. With a fixed N_l, the classification accuracy is

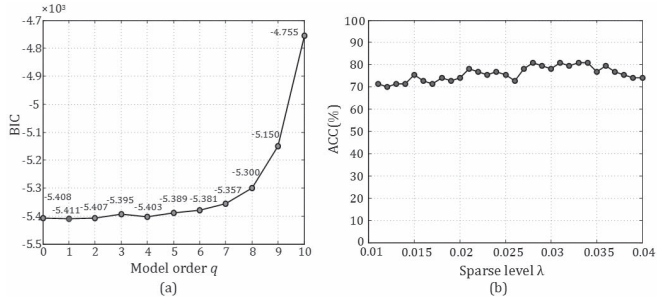

Fig. 8. Influence of parameters. (a) The BIC value for different model orders; (b) The impact of sparse level on final classification accuracy.

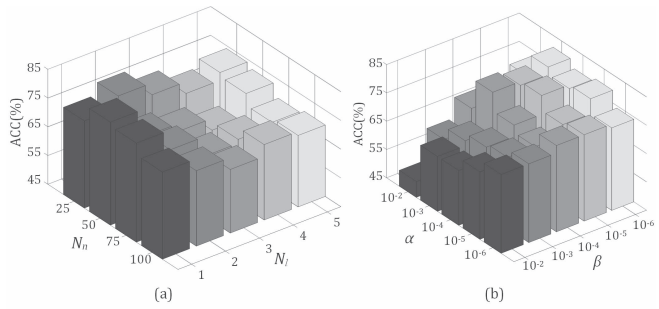

Fig. 9. (a) The classification performance with different the number of nodes and layers; (b) The classification performance with different l_1-norm and l_2-norm regularization parameters.

relatively stable with N_n, implying that our proposed method is not very sensitive to N_n. While the classification performance is predominantly affected by N_l, it is thus important to select the optimal N_l for EMLPC construction. It can be noticed that the classification accuracy decreased with the increase of the number of layers N_l, indicating that the complexity of the EMLPC with one hidden layer is capable of representing the model distribution of the current dataset. Secondly, two regularization parameters α and β balance the relative contributions of l_1-norm regularization and l_2-norm regularization terms. To evaluate the effects of regularization parameters, we explore different combinations of α and β values and their performances are provided in Fig. 9(b). The optimal performance is achieved at $\alpha = 10^{-3}$ and $\beta = 10^{-6}$. Moreover, with a fixed α, the classification accuracy of the proposed method varies smoothly with β. While with a fixed β, the classification performance varies greatly with α. These results imply that the proposed method is robust with respect to l_2-norm regularization parameter while the selection of l_1-norm regularization parameter significantly affect the classification performance. This is reasonable since the l_1-norm regularization parameter works as a sparsity control parameter and determines the scale of the selected feature subset for classification.

IV. Discussion

A. Significance of Results

In this paper, we proposed a novel ultra-group constrained sparse linear regression model for an effective connectivity inference, and a novel EMLPC was employed as the classifier for MCI identification. The classification performance was systematically evaluated under various conditions, including: 1) the usage of the ultra-group constrained structure detection algorithm; 2) the presence or absence of the UOFR algorithm to estimate the connectivity strengths; and 3) the presence or absence of l_1-norm regularization and l_2-norm regularization to constraint the weights of the classifier.

The key findings of our experiment results are summarized as follows: 1) the ultra-group constrained structure detection algorithm improves the discriminative power of the effective connectivity network; 2) the UOFR algorithm for connectivity strength estimation further enhances the discriminative power; 3) the l_1-norm and l_2-norm regularization increases the generalizability of the MLPC; 4) hierarchical features are learned from the weights of the EMLPC. The maximum classification accuracy (80.82%) was obtained using the proposed effective connectivity inference method with the EMLPC. The proposed framework outperforms the other network construction methods and conventional classification approaches on the same data set, indicating its superiority for MCI identification.

B. Efficacy of the Ultra-group Constrained Structure Detection Algorithm

The ultra-group constrained structure detection algorithm incorporates the dependent information of weak derivatives of original signals into linear regression model and thus enhances the noise-resistibility and robustness of the method [11]. Incorporating the weak derivatives information into group constrained sparse learning may reflect the useful information in the data that ultimately helps to detect the model structure more accurately. The ultra-group constrained structure detection algorithm involves two steps. First, the ultra-time series are built by adding the weak derivatives of original time series to the dataset. Then these ultra-time series data are feed into the group constrained sparse learning algorithm to detect the topological structure of effective network. As the ultra-time series includes not only the information of the datum points but also the dependent relationship information, the ultra-group constrained topology structure detection algorithm can characterize the between-ROI interactions more accurately. Also, the group constraint is employed to minimize the effect of inter-subject variability in network representation. By employing a group constraint across different subjects, connectivity networks of different subjects shares a common structure which is consistent with the knowledge that different subjects share a default mode connectivity pattern [36].

C. Efficacy of the UOFR Algorithm

The UOFR algorithm for connectivity strength estimation contributes to interpreting the effective connectivity and further improves the classification performance. With ultra-group constrained structure detection algorithm, the estimated coefficients are unscaled with some coefficients being negative, which leads to difficulty in interpreting and analyzing the effective network. Moreover, though the novel ultra-group constrained structure detection algorithm is efficient in picking up the most significant regressors, the shrinkage may produce

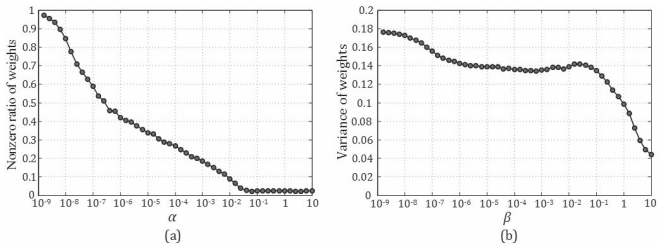

Fig. 10. The influence of regularizations. (a) The sparsity effect of l_1-norm regularization; (b) The variance reduction effect of l_2-norm regularization.

biased estimates in the risk of achieving a suboptimal result. To alleviate these limitations, with the effective connectivity structure identified by the non-zero coefficients estimated via a novel ultra-group constrained structure detection algorithm, we utilize an UOFR algorithm to estimated weights of the effective connectivity network. With the orthogonal process and a step forward regression procedure, connectivity parameters are scaled to the range of (0,1) by using the UERR criterion as the measurement of connectivity strength. As is demonstrated in Section III, with the UOFR algorithm for connectivity strength estimation, the discriminative connectivity power is greatly improved, and hence a classification accuracy increase is obtained.

D. Efficacy of Regularizations

Note that the control of the weight sparsity via the application of the l_1-norm and l_2-norm regularization can obviously improve the classification performance of the MLPC. The l_1-norm regularization and l_2-norm regularization behave differently in constraining the weights of the MLPC [37].

The l_1-norm regularization increases the sparsity of weights in MLPC, leading to a decrease of the nonzero ratio as shown in Fig. 10(a). The greater the l_1-norm regularization parameter, the fewer features will be selected for the MCI classification [38]. Different from l_1-norm regularization, the l_2-norm regularization leads to more average weights in MLPC shown in Fig. 10(b), which means the deviation of weights is decreased. With the l_2-norm regularization, the weights corresponding to different features are more averaged thus the classifier pays attention to more features [37]. For example, when two features are highly correlated, the classifier with l_1-norm regularization prefers to pick one of the two features by setting coefficient of the other to zero. In contrast, the l_2-norm regularization will keep both of them and jointly shrink the corresponding coefficients a little bit. By adjusting the parameters of the l_1-norm regularization and l_2-norm regularization, a trade-off between fitting ability to the training data and generalizability to unseen data is obtained. The best combination of parameters is automatically determined via grid search mentioned in Section II.

E. Interpretation of Discriminative Features

The discriminative features we found by linear projection across all layers are in line with previous studies [26], [27]. Based on the quantitative analysis in Section III, 15 brain regions are identified to be associated with MCI diseases, and a loss of rich club effect is observed in MCI groups. Local efficiency coefficients and rich club coefficients are extracted as original features to feed into the classifier. The brain regions identified with local efficiency coefficients include hippocampus, middle temporal gyrus, olfactory, and other brain regions which were founded to be associated with MCI/AD in previous researches [39], further justify the efficiency of the proposed method. Additionally, the un-normalized and normalized rich club coefficients are greater in healthy controls than in MCI patients, which is consistent with the hypothesis that the functional organization of the normal brain may be impaired by the MCI disease.

In addition, multi-scale features are fused by the EMLPC. Specifically, the local efficiency coefficient is a node level feature that measures the characteristic of brain regions separately, while rich club coefficient is a whole brain level feature which measures the characteristic of all brain regions. By the combination of nodal level features and whole brain level features, the proposed classifier can acquire a more discriminative power which leads to the final increase of classification accuracy.

F. The Most Discriminative Connections

The brain regions associated with the most discriminative connections that are identified based on group-based analysis are mostly overlapped with the brain regions that are selected by the EMLPC in the classification process, further validating the effectiveness of the proposed method in identifying disease-related brain regions. Furthermore, the identified regions mainly locates in frontal lobes (e.g., superior frontal gyrus (medial), Orbitofrontal cortex (medial) [40]), temporal lobes (e.g., heschl gyrus, superior temporal gyrus [40]), and limbic lobe (e.g., middle cingulate gyrus [41]), parietal lobe (e.g., supramarginal gyrus, precuneus [42]), in line with biological findings that MCI causes atrophy in the frontal and temporal lobes [43].

Noted that the connections in the current study are directed and weighted, which thus provides a more comprehensive characterization of the interactions within the whole-brain network. It is obvious from the connectivity analysis in section II that the effective connectivity for most of the optimal connections including left superior frontal gyrus (medial) to left rolandic operculum, from right calcarine to left fusiform gyrus, from right postcentral gyrus to left precentral gyrus, from left heschl gyrus to left rolandic operculum, are much weaker in MCI patients than that of NCs, implying that the neural interactions between these brain regions are deteriorated and the communication pathways between them are interrupted. In addition, the rolandic operculum always involves in the selected connections, indicating that it may play a significant role in information transformation for cognitive function [44]. However, more attention is required in future to better understand the roles of rolandic operculum in human cognitive function.

G. Limitation and Future Directions

Though achieving good MCI classification performance, the proposed method possesses two major limitations.

First, the selection of brain atlas may potentially influence the generalization performance of the classifier. The computed effective connectivity may be quite different by using different atlases with different levels. Another limitation is the size and type of the dataset. The dataset used in this study is a bit small compared with other datasets commonly used for machine learning method. Nevertheless, the obtained results do provide evidence on the efficacy of the proposed framework for identifying MCI patients from normal controls. In the future, we will validate the effectiveness of the proposed framework on the dataset with larger sample sizes.

V. Conclusion

In summary, we proposed a new effective connectivity network based classification framework for MCI identification, and achieve superior classification performance compared to the state-of-the-art methods. Experimental results illustrated that by imposing weak derivative of original time series into the group constrained sparse learning algorithm, the linear regression model can construct effective connectivity networks with greater discriminative power. In addition, the classification performance successfully validated the feasibility of the EMLPC classifier toward the automated diagnosis of MCI patients using the local efficiency coefficients and rich club coefficients as input patterns. The results of the proposed EMLPC method are consistent with the connectivity analysis results, providing additional evidence of disrupted network organization in MCI at a nodal level and whole brain level. The combination of the novel effective connectivity network inference method and the EMLPC classifier can be useful for providing a complementary objective opinion to the clinical diagnosis of cognitive impairment. Ultimately, the proposed method can benefit the development of computer-aided diagnosis tools in the clinical and pre-clinical settings.

References

[1] S. Gauthier *et al.*, "Mild cognitive impairment," *Lancet*, vol. 367, no. 9518, pp. 1262–1270, 2006.

[2] C. Misra, Y. Fan, and C. Davatzikos, "Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI," *NeuroImage*, vol. 44, no. 4, pp. 1415–1422, 2009.

[3] J. Bischkopf, A. Busse, and M. C. Angermeyer, "Mild cognitive impairment—A review of prevalence, incidence and outcome according to current approaches," *Acta Psychiatrica Scandinavia*, vol. 106, no. 6, pp. 403–414, 2002.

[4] B. Singh *et al.*, "A prospective study of chronic obstructive pulmonary disease and the risk for mild cognitive impairment," *JAMA Neurol*, vol. 71, no. 5, pp. 581–588, 2014.

[5] J. J. Hsiao and E. Teng, "Depressive symptoms in clinical and incipient Alzheimer's disease," *Neurodegenerative Disease Manage.*, vol. 3, no. 2, pp. 147–155, 2013.

[6] M. Bola and B. A. Sabel, "Dynamic reorganization of brain functional networks during cognition," *Neuroimage*, vol. 114, pp. 398–413, Jul. 2015.

[7] "Multimodal hyper-connectivity of functional networks using functionally-weighted LASSO for MCI classification," *Med. Image Anal.*, vol. 52, pp. 80–96, 2019.

[8] H. Lee, D. S. Lee, H. Kang, B.-N. Kim, and M. K. Chung, "Sparse brain network recovery under compressed sensing," *IEEE Trans. Med. Imag.*, vol. 30, no. 5, pp. 1154–1165, May 2011.

[9] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc., B (Statist. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.

[10] C.-Y. Wee, P.-T. Yap, D. Zhang, L. Wang, and D. Shen, "Group-constrained sparse fMRI connectivity modeling for mild cognitive impairment identification," *Brain Struct. Function*, vol. 219, no. 2, pp. 641–656, 2014.

[11] Y. Guo, L. Z. Guo, S. A. Billings, and H.-L. Wei, "Ultra-orthogonal forward regression algorithms for the identification of non-linear dynamic systems," *Neurocomputing*, vol. 173, pp. 715–723, Jan. 2016.

[12] B. Jie, C.-Y. Wee, D. Shen, and D. Zhang, "Hyper-connectivity of functional networks for brain disease diagnosis," *Med. Image Anal.*, vol. 32, pp. 84–100, Aug. 2016.

[13] R. A. Khan *et al.*, "fNIRS-based neurorobotic interface for gait rehabilitation," *J. Neuroeng. Rehabil.*, vol. 15, p. 7, Feb. 2018.

[14] J. L. Lancaster *et al.*, "Automated Talairach atlas labels for functional brain mapping," *Hum. Brain Mapping*, vol. 10, no. 3, pp. 120–131, Jul. 2000.

[15] S. R. Kesler *et al.*, "Predicting long-term cognitive outcome following breast cancer with pre-treatment resting state fMRI and random forest machine learning," *Front Hum. Neurosci.*, vol. 11, pp. 256–265, Nov. 2017.

[16] S. Farhan, M. A. Fahiem, and H. Tauseef, "An ensemble-of-classifiers based approach for early diagnosis of Alzheimer's disease: Classification using structural features of brain images," *Comput. Math. Method Med.*, vol. 2014, Aug. 2014, Art. no. 862307.

[17] C. Wu *et al.*, "Improving interpretability and regularization in deep learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 2, pp. 256–265, Feb. 2018.

[18] J. Ashburner, "SPM: A history," *Neuroimage*, vol. 62, no. 2, pp. 791–800, 2012.

[19] Y. Chao-Gan and Z. Yu-Feng, "DPARSF: A MATLAB toolbox for 'pipeline' data analysis of resting-state fMRI," *Front Syst. Neurosci.*, vol. 4, p. 13, May 2010.

[20] D. Shen and C. Davatzikos, "HAMMER: Hierarchical attribute matching mechanism for elastic registration," *IEEE Trans. Med. Imag.*, vol. 21, no. 11, pp. 1421–1439, Nov. 2002.

[21] L.-Z. Liu, F.-X. Wu, and W.-J. Zhang, "A group LASSO-based method for robustly inferring gene regulatory networks from multiple time-course datasets," *BMC Syst. Biol.*, vol. 8, no. 3, p. S1, 2014.

[22] Y. Li *et al.*, "Time-varying system identification using an ultra-orthogonal forward regression and multiwavelet basis functions with applications to EEG," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2960–2972, Jul. 2018.

[23] G. Deshpande and X. Hu, "Investigating effective brain connectivity from fMRI data: Past findings and current issues with reference to Granger causality analysis," *Brain Connectivity*, vol. 2, no. 5, pp. 235–245, 2012.

[24] J. Liu, S. Ji, and J. Ye, "SLEP: Sparse learning with efficient projections," Dept. Comput. Sci. Eng., Arizona State Univ., Tempe, AZ, USA, Tech. Rep., 2009, pp. 1–61, vol. 6, no. 491.

[25] L. Wang, Y. You, and H. Lian, "Convergence and sparsity of Lasso and group Lasso in high-dimensional generalized linear models," *Stat. Papers*, vol. 56, no. 3, pp. 819–828, 2015.

[26] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[27] T. Yan, W. Wang, L. Yang, K. Chen, R. Chen, and Y. Han, "Rich club disturbances of the human connectome from subjective cognitive decline to Alzheimer's disease," *Theranostics*, vol. 8, no. 12, pp. 3237–3255, 2018.

[28] T. Opsahl, V. Colizza, P. Panzarasa, and J. J. Ramasco, "Prominence and control: The weighted rich-club effect," *Phys. Rev. Lett.*, vol. 101, no. 16, p. 168702, 2008.

[29] Z. Wang, Y. Yuan, F. Bai, J. You, and Z. Zhang, "Altered topological patterns of brain networks in remitted late-onset depression: A resting-state fMRI study," *J. Clin. Psychiatry*, vol. 77, no. 1, pp. 123–130, 2016.

[30] K. T. E. O. Dubbelink *et al.*, "Disrupted brain network topology in Parkinson's disease: A longitudinal magnetoencephalography study," *Brain*, vol. 137, pp. 197–207, Jan. 2014.

[31] N. Jaitly and G. E. Hinton, "Using an autoencoder with deformable templates to discover features for automated speech recognition," in *Proc. INTERSPEECH*, vols. 1–5, 2013, pp. 1737–1740.

[32] M. E. Lynall *et al.*, "Functional connectivity and brain networks in schizophrenia," *J. Neurosci.*, vol. 30, no. 28, pp. 9477–9487, 2010.

[33] Y. Li, C.-Y. Wee, B. Jie, Z. Peng, and D. Shen, "Sparse multivariate autoregressive modeling for mild cognitive impairment classification," *Neuroinformatics*, vol. 12, no. 3, pp. 455–469, 2014.

[34] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, Jul. 2010.

[35] M. P. van den Heuvel and O. Sporns, "Rich-club organization of the human connectome," *J. Neurosci.*, vol. 31, no. 44, pp. 15775–15786, 2011.

[36] W. Li, X. Mai, and C. Liu, "The default mode network and social understanding of others: What do brain connectivity studies tell us," *Front Hum. Neurosci.*, vol. 8, p. 74, Feb. 2014.

[37] E. A. Smirnov, D. M. Timoshenko, and S. N. Andrianov, "Comparison of regularization methods for ImageNet classification with deep convolutional neural networks," *AASRI Proc.*, vol. 6, no. 1, pp. 89–94, 2014.

[38] Z. Wang, S. Gao, and L.-T. Chia, "Learning class-to-image distance via large margin and L1-norm regularization," in *Proc. ECCV*, vol. 7573, ~~2012,~~ pp. 230–244.

[39] M. M. Vasavada *et al.*, "Olfactory cortex degeneration in Alzheimer's disease and mild cognitive impairment," *J. Alzheimers Disease*, vol. 45, no. 3, pp. 947–958, 2015.

[40] A. Convit, J. Asis, M. J. de Leon, C. Y. Tarshish, S. De Santi, and H. Rusinek, "Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease," *Neurobiol. Aging*, vol. 21, no. 1, pp. 19–26, 2000.

[41] E. Guedj *et al.*, "Effects of medial temporal lobe degeneration on brain perfusion in amnestic MCI of AD type: Deafferentation and functional compensation?" *Eur. J. Nucl. Med. Mol. Imag.*, vol. 36, no. 7, pp. 1101–1112, 2009.

[42] M. Bailly *et al.*, "Precuneus and cingulate cortex atrophy and hypometabolism in patients with Alzheimer's disease and mild cognitive impairment: MRI and F-FDG PET quantitative analysis using freesurfer," *Biomed. Res. Int.*, ~~vol. 2015, May 2015,~~ Art. no. 583931.

[43] X. He *et al.*, "Changes in theta activities in the left posterior temporal region, left occipital region and right frontal region related to mild cognitive impairment in Parkinson's disease patients," *Int. J. Neurosci.*, vol. 127, no. 1, pp. 66–72, 2017.

[44] M. Cao *et al.*, "Early development of functional network segregation revealed by connectomic analysis of the preterm human brain," *Cerebral Cortex*, vol. 27, no. 3, pp. 1949–1963, 2017.

[45] Y. Li *et al.*, "Epileptic seizure detection in EEG signals using sparse multiscale radial basis function networks and the Fisher vector approach," *Knowl.-Based Syst.*, to be published, doi: 10.1016/j.knosys.2018.10.029.