Multi-Modal Neuroimaging Feature Learning for Multi-Class Diagnosis of Alzheimer's Disease

Siqi Liu, *Student Member, IEEE,* Sidong Liu, *Student Member, IEEE,* Weidong Cai, *Member, IEEE,* Hangyu Che, Sonia Pujol, Ron Kikinis, Dagan Feng, *Fellow, IEEE*, Michael J. Fulham, ADNI

***Abstract*—The accurate diagnosis of Alzheimers disease (AD) is essential for patient care and will be increasingly important as disease modifying agents become available, early in the course of the disease. Although studies have applied machine learning methods for the computer aided diagnosis (CAD) of AD, a bottleneck in the diagnostic performance was shown in previous methods, due to the lacking of efficient strategies for representing neuroimaging biomarkers. In this study, we designed a novel diagnostic framework with deep learning architecture to aid the diagnosis of AD. This framework uses a zero-masking strategy for data fusion to extract complementary information from multiple data modalities. Compared to the previous state-of-the-art work-flows, our method is capable of fusing multi-modal neuroimaging features in one setting and has the potential to require less labelled data. A performance gain was achieved in both binary classification and multi-class classification of AD. The advantages and limitations of the proposed framework are discussed.**

***Index Terms*—Alzheimer's Disease, Classification, Neuroimaging, MRI, PET, Deep Learning.**

I. Introduction

ALZHEIMER's disease (AD) is a degenerative brain disorder which is characterised by a progressive dementia that is charactered by the degeneration of specific nerve cells, presence of neuritic plaques and neurofibrillary tangles [1]. A decline in memory and other cognitive functions are the usual early syndromes. AD will be a global burden over the coming decades, due to the increasing age of societies. It was reported that in 2006 there were 26.6 million AD cases in the world, including about 56% of the cases that are at the early stage. In 2050, the population of the AD patients is predicted to grow fourfold to 106.8 million [2]. The precise diagnosis of AD was considered as a difficult clinical task with insufficient specificity because the evaluation of the mental status cannot be made when the consciousness is impaired. Another difficulty is caused by the confusion of other non-AD dementia syndromes. Mild Cognitive Impairment (MCI), a prodromal stage of AD, has drawn attention of researchers recently, because it is useful for clinical trials. Though MCI does not notably interfere with daily activities, it has been constantly proven that MCI patients are at a high risk of AD progression [3]. To conduct prediction of transition risk of MCI, MCI subjects can be further categorised as MCI Converters (cMCI) and MCI Non-converters (ncMCI). It is essential to detect the early stages as well as across the full spectrum of AD progression, therefore patients are allowed to control the risk factors, for example isolated systolic hypertension [4], [5], before irreversible brain damage develops. Neuroimaging techniques, such as Magnetic Resonance Imaging (MRI) [6]–[11] and Positron Emission Tomography (PET) [12]–[18], have been widely used in the assessment of AD, along with many other non-imaging biomarkers [6], [19], [20].

Machine learning methods have been proposed to aid the diagnosis of AD. Pre-computed medical descriptors were widely used to represent biomedical images. Approximate measurements, such as the volume [21] and the cerebral metabolic rate of glucose (CMRGlc) [22], were normally computed from segmented 3D brain regions of interest (ROI), and were used for AD classification with Support Vector Machine (SVM) [23], Bayesian method [24] or other methods [25], [26]. However, there are several constraints in such work-flows. The methods based on these conventional machine learners often work well in binary classification, such as categorising AD subjects from normal control (NC) subjects, but it is difficult to extend them to multi-classes [27]. As a result, although the diagnosis of AD should be naturally modelled as a multi-class classification problem, it was normally simplified as a set of binary classification tasks [23], [28] which distinguish AD or MCI subjects from NC subjects. Another constraint is the embedding of clinical prior knowledge. A method based on the graph cut algorithm was proposed recently by Liu *et al.* [10]. This work-flow adjusted the graph cut algorithm with parameters corresponding to the relationships between different stages of AD. Though such customisation tends to yield promising classification results, the work-flow can be sensitive to changes in the dataset and can be difficult to extend to a large scale. Another challenge of AD diagnosis is to represent the original biomarkers in an unsupervised approach. Some frameworks reduces the dimensionality of each type of biomarker in a supervised way and then fuses the feature modalities to form a new feature space [29]–[31]. Such work-flows depend heavily on the quantity of the

This work was supported in part by the ARC, AADRF, NA-MIC (NIH U54EB005149), and NAC (NIH P41EB015902).

S.Q. Liu, S. Liu, W. Cai, H. Che and D. Feng are with the Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Information Technologies, University of Sydney, Australia.

S.Liu, W.Cai, S. Pujol and R. Kikinis are with the Surgical Planning Laboratory, Department of Radiology, Brigham and Womens Hospital, Harvard Medical School, Boston, USA.

D. Feng is also with the Med-X Research Institute, Shanghai Jiao Tong University, China.

M.J. Fulham is with the Department of PET and Nuclear Medicine, Royal Prince Alfred Hospital, and the Sydney Medical School, University of Sydney, Australia.

labelled samples, which are difficult to achieve. Separating the dimensionality reduction and the data fusion may also result in losing complementary information.

We believe the previous work-flows can be optimised by designing a new framework to efficiently represent the multi-modal biomarkers and effectively characterise the multiple stages of AD. The conventional feature engineering work-flows with shallow structures and affine data transformation often simply result in feature repetition or dimensionality selection. As shown in many recent studies, deep data representation can be more efficient than the shallow architectures in multi-class classification by disentangling complex patterns in the inputs [32]–[36]. Deep learning architectures extract high-level features progressively via several layers of feature representations [37]. The high-level features tend to be more separable in classification problems due to the sequential transformations of the feature space.

Brosch and Tam, using MR, reported that multi-layered learning structure was effective in capturing shape variations of the brain regions that correlate with demographic and disease information, such as the ventricle size [38]. In the framework proposed by Suk *et al.* [39], one setting of stacked auto-encoders (SAE) was trained for each image modality, then the learnt high-level features were further fused with a multi-kernel support vector machine (MKSVM). In such work-flows, the single-modal high-level features were learnt regardless of the other modalities, which may ignore the synergy between different modalities in the feature learning.

In this study, we propose a novel framework of multi-class AD diagnosis with deep learning architecture embedded which benefits from the synergy between multi-modal neuroimaging features. The framework is constructed with an SAE and a softmax logistic regressor. The auto-encoders represent the data in an unsupervised way which can be extended to use unlabelled data in practice. The proposed framework is capable of data fusion when multi-modal neuroimaging image data are available. Following the concepts of de-noising auto-encoder [40], we applied the zero-mask strategy on bimodal deep learning tasks to extract the synergy between different image modalities. By randomly hiding one modality of the training set, the hidden layers of the neural network tend to be able to reconstruct the missing modality with corrupted inputs by inferring the correlations between multi-modal features. With the softmax regression embedded in the deep learning architecture, our framework is capable of classifying AD patients into four AD stages.

The rest of this paper is organised as follows. We introduce the proposed learning framework and the training strategies in Section II. The experiments and results of this study are presented in Section III. We discuss the proposed framework and conclusions of the paper in sections IV and V.

II. Methodology

The pipeline of the proposed framework is illustrated in Fig. 1. In this study, MR and PET data are used as two input neuroimaging modalities. All collected brain images are firstly pre-processed and segmented into 83 functional ROI, and a set of descriptors are computed from each ROI. The dataset is divided into a training set and a testing set. We perform Elastic Net [9], [41], [42] only on the training samples to select the discriminative subset of the feature parameters. A multi-layered neural network consisting of several auto-encoders is then trained using the selected feature subset in the training dataset. Each layer of the network obtains a higher level of abstraction of the previous layer with non-linear transformation [43]–[45]. The softmax layer is added on the top of the stacked auto-encoders for classification. The trained network is then evaluated with the labelled testing samples.

A. Data Acquisition and Feature Extraction

The neuroimaging data used in this study were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database[1] [46]. This database was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organisations as a 5-year public partnership. The primary purpose of ADNI project was to study the effects of combining multiple biomarkers, such as MRI, PET and CSF data accompanied with neuropsychological assessments, to predict the progression of MCI and early AD. Around 200 normal instances and 400 MCI instances were followed for 3 years, 200 AD patients were followed within 2 years. Determining the sensitive biomarkers to the progression of AD might also aid the clinicians to discover new treatments, as well as other possible biomedical exploration.

We obtained two datasets from ADNI. For the dataset with only MR images, 816 age and sex matched subjects were recruited from the ADNI repository and a T1-weighted MR image was acquired from each subject. We excluded 20 subjects with multiple conversions or reversions as well as 21 MCI subjects whose data were incomplete. We labelled the MCI subjects that converted to AD from 0.5 to 3 years from the first scan as MCI converters (cMCI), otherwise the MCI subjects were labelled as MCI non-converters (ncMCI). The normal subjects and the AD patients were labelled as Normal Control (NC) and Alzheimers Disease (AD) [10]. All raw MR images were corrected following the ADNI MR image protocol, and were non-linearly registered to the ICBM_152 template [47] using the Image Registration Toolkit (IRTK) [48]. Only 17 images were excluded because of the intolerable distortion. Finally, 758 MR subjects were reserved for the experiments conducted in this study, including 180 AD subjects, 160 cMCI subjects, 214 ncMCI subjects and 204 normal ageing control subjects.

For the dataset with multi-modal data fusion, 331 age and sex matched subjects were selected from the baseline cohort, including 77 NC-, 102 ncMCI-, 67 cMCI-, 85 AD-subjects with both MR and PET data available. Each instance was associated with T1-weighted volumes and FDG-PET images. All the 3D images were pre-processed with the similar work-flow described earlier for MR images. The PET images were

[1]The database is available at adni.loni.ucla.edu

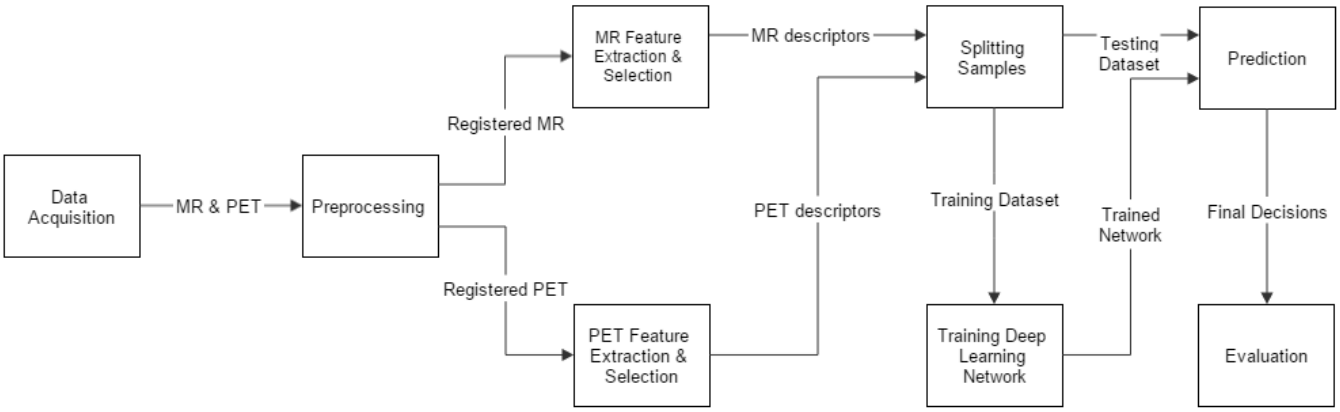

Fig. 1. The proposed diagnostic framework of Alzheimer's Disease with deep learning architecture embedded.

aligned to the corresponding MR image using FSL FLIRT [49].

For each registered 3D image, 83 brain regions were mapped in the template space using multi-atlas propagation with enhanced registration (MAPER) approach [50]. The grey matter volumes were extracted from MR images, same as in [9], [10]. For PET images, we extracted the regional average CMRGlc feature same as in [22], [51]. We then normalised features to be between 0 and 1 to support the sigmoidal decoders by shifting the negative values and rescaling.

B. Learning Framework

1) Pre-Training Stacked Auto-Encoders: We applied stacked auto-encoders (SAE) [45], [52], [53] to learn the high-level features in an unsupervised way as shown in Fig. 2. Each auto-encoder framework encodes an input vector x into a hidden representation y with an affine mapping followed by non-linear sigmoidal distortion,

$$y = \sigma(Wx + b) \tag{1}$$

where σ is set as a sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$; W is a weight matrix and b is a vector of bias terms. y is the encodings that represent the original input x. The ideal case is that we can maximally reconstruct x with only knowing y. The decoder reconstructs the input vector from the hidden representation by

$$x^* = \tau(W^{'} y + b^{'}) \tag{2}$$

where τ is another sigmoidal filter; $W^{'}$ is the decoding weights. The number of the hidden neurons determines the dimensionality of the encodings at each layer. By controlling the number of hidden units, we can either perform dimensionality reduction or learn over-complete features. The decoding results in a reconstruction of input vector x with high probability of $P(x^*|x)$. Therefore the reconstruction loss can be minimised by optimising the log likelihood,

$$L(x, x^*) \propto -logP(x^*|x) \tag{3}$$

Since the features extracted from MR images were real valued and were normalised to a domain $x \in [0, 1]$, we used the mean squared error to measure the reconstruction loss $L(x, x^*)$. To prevent the auto-encoder from learning merely an identity function, the objective function is regularised by adding a weight decay, e.g.

$$L(W, b, x, x^*) = \min_{W,b} L(x, x^*) + \lambda \|W\|_2^2 \tag{4}$$

where $\|W\|_2^2$ is the weight decay that controls over-fitting.

Though the objective function is not convex, the gradients of the objective function in Eq. (4) can be exactly computed by error back-propagation algorithm. In this study we applied the Non-Linear Conjugate Gradient algorithm to optimise Eq. (4) [52].

Following the greedy layer-wised training strategy, rather than training all the hidden layers of the unsupervised network altogether, we train one auto-encoder with a single hidden layer at a time [43]. When an auto-encoder is trained with the features obtained from the previously trained hidden layers, the hidden layer of the current auto-encoder is then stacked on the trained network. After training all the auto-encoders, the final high-level features are obtained by feed-forwarding the activation signals through the stacked sigmoidal filters. When unlabelled subjects are available, the unsupervised feature learning can be performed with a mixture of the labelled and the unlabelled samples.

2) Multi-Modal Data Fusion: When more than one image modality are used for model training, modality fusion methods are required to discover the synergy between different modalities. Shared representation can be obtained by jointly training the auto-encoders with the concatenated MR and PET inputs. The first shared hidden layer is used to model the correlations between different data modalities. However the simple feature concatenation strategy often results in hidden neurons that are only activated by one single modality, because the correlations of MR and PET are highly non-linear. Inspired by Ngiam *et al.* [54], we applied the pre-training method with a proportion of corrupted inputs which had only one modality presented, following the de-noising concepts of training deep architecture. One of the modalities is randomly hidden by

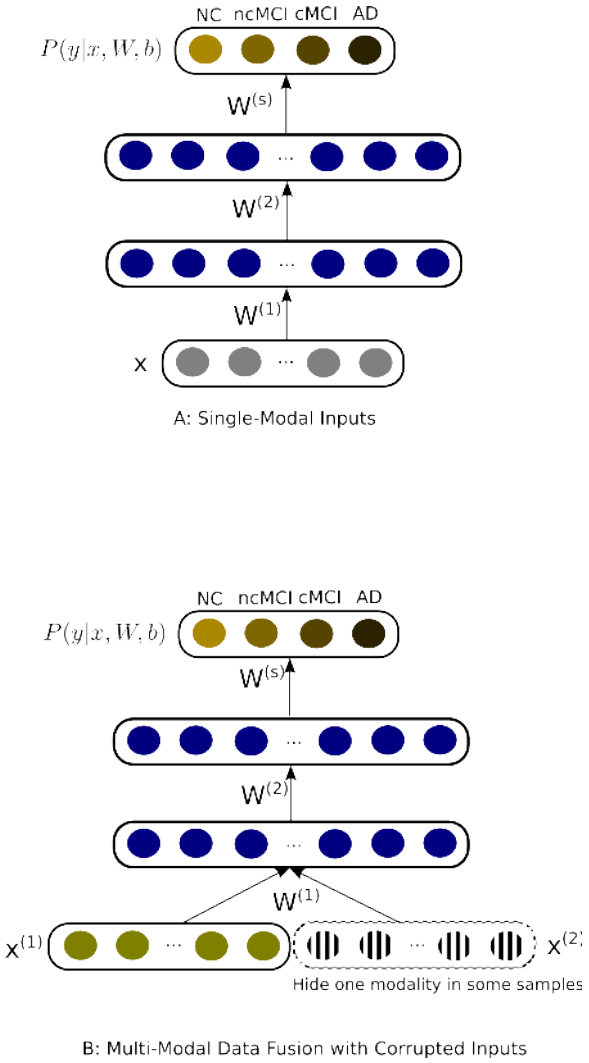

Fig. 2. Illustration of the single-modal & multi-modal architectures of the proposed framework.

replacing these inputs with 0; the rest of the training samples are presented with both modalities. The hidden layer of the first auto-encoder is trained to reconstruct all of the original inputs from the inputs that are mixed with hidden modalities. The original inputs and the corrupted inputs are propagated to the higher layers of the neural network independently to obtain both the clean representation and the noisy representation using the same neural network. Each higher layer is then trained progressively to reconstruct the clean high-level representation from the propagated noisy representation. Thus some of the hidden neurons are expected to infer the correlations between different neuroimaging modalities.

3) Fine-Tuning for AD Classification: For the the AD diagnosis, we modelled the task as a four-class classification problem containing four pre-defined labels: NC, cMCI, ncMCI and AD). Although the features learnt by the unsupervised network can also be transferred to a conventional classifier, such as SVM, softmax logistic regression enables us to jointly optimise the entire network via fine-tuning.

The features extracted by the unsupervised network are fed to an output layer with softmax regression [55]. The softmax layer uses a different activation function, which might have non-linearity different from the one applied in previous layers. The softmax filter is defined as

$$P(Y=i|x) = \frac{e^{W_i^{(s)}a+b_i^{(s)}}}{\sum_i e^{W_i^{(s)}a+b_i^{(s)}}} \tag{5}$$

where Y is the possible stages of AD progression; a is the feature representation obtained from the last hidden layer of the pre-trained network; $W_i^{(s)}$ and $b_i^{(s)}$ are the weight and bias for the i-th possible decision. For example $P(Y=$ 'cMCI' $|x^{(l)})$, is the probability that the patient is diagnosed as a MCI converter. The label with the highest probability is selected as the final diagnosis. Optimising softmax layer is similar to the unsupervised network. The objective function of fine-tuning the network with softmax layer is defined as

$$L(W,b,X,Y) = \min_{W,b} J(X,Y) + \lambda^{(s)}\|W^{(s)}\|_2^2 \tag{6}$$

where W and b are the weights and bias of the entire network, including the pre-trained SAE and the softmax regression layer; $J(X,Y)$ is the logistic regression cost between the diagnosis generated with on the input features X and the pre-labelled results Y; $\lambda^{(s)}$ is the relative weight of the weight decay on softmax layer, which can be tuned to control the over-fitting problem. To fine-tune the pre-trained structure, the softmax layer is then connected to the last hidden layer of the unsupervised network. We then propagate the activation signals through the entire neural network and optimise all the parameters according to the classification loss as a supervised neural network [43], [56] as shown in Fig. 2-A. When more than one modality are used in training the supervised network, a small proportion of the single modal inputs are dropped out in a similar way described in Section II-B-2. The hidden neurons are trained to make diagnosis even when one modality is absent. This strategy is supposed to make some of the hidden neurons at the first hidden layer easy to be activated by the incoming weights from both modalities [54].

C. Feature Examination

Hidden neurons at the first layer of our network are trained to catch different patterns of input subjects. In deep learning tasks with general images the hidden neurons can be visualised as

$$x_{ij}^{'} = \frac{W_{ij}^{(l)}}{\|W\|_2} \tag{7}$$

where $x_{ij}^{'}$ is the input pattern that maximally activates the i-th hidden neuron.

Unlike pixels, the patterns in biomarkers are brain ROI measurements that may be non-trivial to be visualised. We examined the representation quality by mapping the input patterns produced by Eq. (7) back to a masked 3D MR image, with 83 segmented ROIs. Each input $x_{ij}^{'}$ corresponds to the brain ROI where it was extracted. By splitting the pattern x into m features (volumes, CMLGLc, etc.), we compute the variance $D_j^{(m)}$ of all of the same ROI, measuring how the ROI activate different hidden neurons. When $D_j^{(m)}$ is low, the

biomarkers from region j are more effective for AD diagnosis than the other regions. The overall feature stability S_j of the j-th ROI can be computed as

$$S_j = \sum_m \frac{\sum_j D_j^{(m)}}{D_j^{(m)}} \tag{8}$$

S can be convolved with a Gaussian filter to enlarge the distinctions between different ROIs. The brain ROIs with relatively high stability score are considered as more effective to the AD progression. These mappings on the MR image can be examined with the clinical prior-knowledge to monitor the performance of the feature learning network in the context of AD diagnosis.

III. Experiments and Results

A. Visualisation of High-level Biomarkers

With the feature examination method described in Section II-C, we calculated the stability score of each brain ROI and mapped the stability score to a masked 3D MR image (83 ROIs) of a Normal Control subject as shown in Fig. 3. The distinctions between various ROIs were clearly visualised. The darker regions tend to be more sensitive to the progression of AD and MCI than the lighter ROIs, since features extracted from these ROIs tend to benefit all the hidden neurons equally. The light regions are not denoted to be totally trivial, but carrying less predictive information.

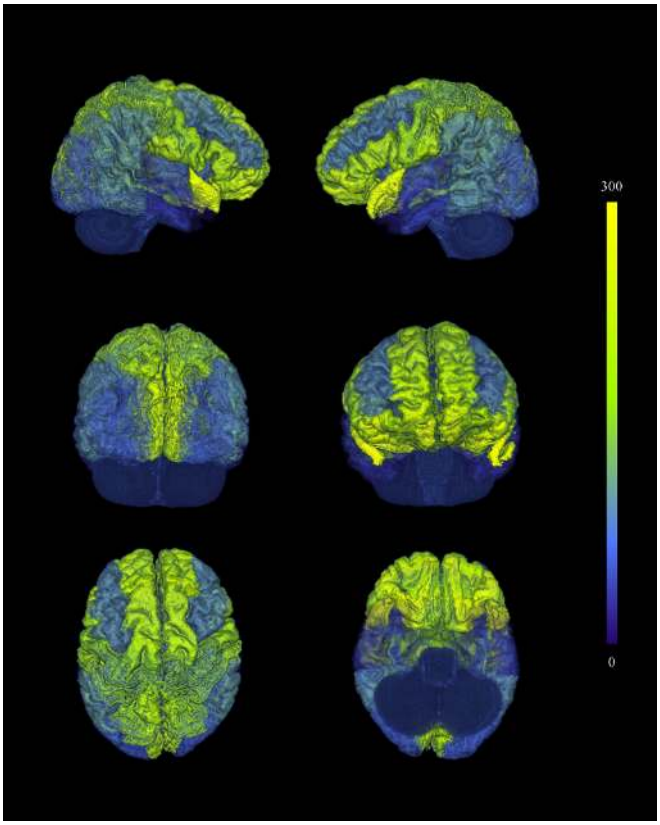

Fig. 3. The image was generated using 3D Slicer (V4.3) [57]. The stability map denotes that the darker ROIs tend to be more affected by AD, hence more sensitive to the AD progression than the brighter regions. The features extracted from the darker areas are also considered as more stable predictors for our deep learning model and tend to be beneficial to all of the hidden neurons in the neural network.

B. Performance Evaluation

We compared the proposed framework with the widely applied methods using the single-kernel SVM and the multi-kernel SVM (MKSVM) [23], [28]. To evaluate the proposed data fusion method, we compared the zero-mask method to the architecture proposed in [39] that trains two stacked auto-encoders independently and then fuses the high-level features with MKSVM after each SAE is fine-tuned. All of the experiments were evaluated with the same features extracted from MR images and PET images as described in Section II-A.

The proposed framework was implemented on Matlab 2013a. The SVM-based experiments were performed using LIBSVM [58]. The multi-kernel SVM was implemented by using precomputed kernels and fusing the multiple kernels with relative weights.

The evaluation was conducted by using 10-fold cross-validation. In experiments including multiple modalities, we compared the performance with only single modal data, MR or PET, and the data fusion methods with both modalities. To avoid the 'lucky trials', we randomly sampled the training and testing instances from each class to ensure they have similar distributions as the original dataset. The entire network was trained and fine-tuned with the 90% of data and then tested with the rest of samples in each validation trial. The hyper-parameters of all compared methods were chosen in each validation trial using the approx search in log-domain to obtain the best performed model [59]. Two hidden layers were used in all neural network based experiments because adding additional hidden layers did not show further improvements on AD classification. It is reasonable to assume that 2 non-linear transformations could be ideal to represent the neuroimaging features for AD classification. The number of neurons at hidden layers were chosen between 30 to 200 according to the classification performance in each fold. In each neural network, hidden layers shared the same number of hidden neurons [60]. The MKSVM was trained with the training samples. Following the work-flow in [23], the relative weights of each kernel in MKSVM were chosen through a coarse-grid search with a step size of 0.1. In the experiments that used MKSVM for fusing two SAE networks, each SAE was firstly pre-trained and fine-tuned with the training data, and then the high-level features obtained from each network were fused with MKSVM with the procedure stated before.

1) Experiments on MR (758 Subjects): We firstly evaluated the proposed framework with 758 3D MR images. Since only one modality is presented, no modality fusion strategy was used in both SVM and the proposed method.

The performance of binary classification (NC vs. AD and NC vs. MCI) is displayed in Table I. The first two columns are precisions on individual classes. The following three columns/ are the overall performance including accuracy, sensitivity, specificity. The proposed method (SAE) outperformed SVM in classifying AD subjects from the NC subjects by leading the

TABLE I
THE PERFORMANCE (%) OF BINARY AD CLASSIFICATION WITH MR-ONLY SAMPLES.

NC vs. AD

Methods	NC	AD	ACC	SEN	SPE
SVM	$\mathbf{82.95 \pm 8.57}$	80.16 ± 6.46	81.04 ± 6.28	82.83 ± 6.02	$\mathbf{78.89 \pm 14.66}$
SAE	82.23 ± 6.54	$\mathbf{84.31 \pm 7.36}$	$\mathbf{82.59 \pm 5.33}$	$\mathbf{86.83 \pm 6.83}$	77.78 ± 10.83

NC vs. MCI

Methods	NC	MCI	ACC	SEN	SPE
SVM	62.39 ± 4.87	74.81 ± 4.34	71.27 ± 3.26	47.02 ± 12.37	$\mathbf{84.50 \pm 4.25}$
SAE	$\mathbf{67.44 \pm 13.14}$	$\mathbf{75.80 \pm 3.58}$	$\mathbf{71.98 \pm 5.48}$	$\mathbf{49.52 \pm 13.68}$	84.31 ± 13.15

TABLE II
THE PERFORMANCE (%) OF MULTI-CLASS AD CLASSIFICATION WITH MR-ONLY SAMPLES.

Methods	NC	ncMCI	cMCI	AD	ACC	SEN	SPE
SVM	46.96 ± 3.95	$\mathbf{42.95 \pm 10.80}$	37.88 ± 10.18	44.62 ± 8.04	44.45 ± 3.07	$\mathbf{75.00 \pm 9.04}$	68.59 ± 5.13
SAE	$\mathbf{52.40 \pm 8.43}$	41.25 ± 7.16	$\mathbf{38.71 \pm 23.18}$	$\mathbf{46.89 \pm 4.40}$	$\mathbf{46.30 \pm 4.24}$	66.14 ± 10.57	$\mathbf{77.78 \pm 4.48}$

TABLE III
THE PERFORMANCE (%) OF THE BINARY AD CLASSIFICATION WITH MR AND PET.

NC vs. AD

Methods	NC	AD	ACC	SEN	SPE
SVM-MR	87.83 ± 11.81	84.06 ± 10.42	84.67 ± 8.45	80.54 ± 11.59	88.33 ± 12.19
SVM-PET	84.58 ± 7.02	87.57 ± 10.12	84.60 ± 4.05	84.11 ± 12.86	84.58 ± 9.07
MKSVM	89.68 ± 5.67	91.50 ± 9.37	90.11 ± 5.57	89.64 ± 11.43	90.56 ± 4.76
SAE-MR	88.28 ± 11.68	88.74 ± 10.82	87.79 ± 9.12	87.32 ± 11.19	88.47 ± 12.19
SAE-PET	83.27 ± 12.44	85.91 ± 10.73	83.53 ± 9.80	82.86 ± 15.59	83.75 ± 12.41
SAE-CONCAT	88.67 ± 12.84	92.56 ± 8.35	90.15 ± 9.54	92.14 ± 9.08	88.19 ± 12.72
2SAE-MKSVM	$\mathbf{91.35 \pm 8.15}$	92.42 ± 8.81	$\mathbf{91.40 \pm 6.82}$	90.89 ± 10.40	$\mathbf{91.67 \pm 7.40}$
SAE-ZEROMASK	90.38 ± 7.36	$\mathbf{92.89 \pm 6.17}$	$\mathbf{91.40 \pm 5.56}$	$\mathbf{92.32 \pm 6.29}$	90.42 ± 6.93

NC vs. MCI

Methods	NC	MCI	ACC	SEN	SPE
SVM-MR	67.52 ± 14.15	$\mathbf{83.92 \pm 7.56}$	77.70 ± 5.27	$\mathbf{62.50 \pm 20.11}$	84.60 ± 6.60
SVM-PET	62.66 ± 22.67	77.17 ± 4.78	72.35 ± 8.67	45.54 ± 10.23	84.60 ± 10.94
MKSVM	70.85 ± 17.69	80.16 ± 3.40	76.88 ± 5.83	52.14 ± 8.56	88.16 ± 8.73
SAE-MR	65.56 ± 24.61	76.86 ± 5.59	74.02 ± 7.58	40.36 ± 15.63	89.26 ± 7.61
SAE-PET	50.69 ± 22.53	77.12 ± 7.51	70.00 ± 9.33	46.96 ± 19.12	80.44 ± 8.42
SAE-CONCAT	73.56 ± 16.55	80.00 ± 4.94	77.65 ± 5.18	49.46 ± 14.35	90.51 ± 7.09
2SAE-MKSVM	$\mathbf{90.42 \pm 11.46}$	73.85 ± 10.01	77.90 ± 5.18	61.43 ± 18.99	$\mathbf{92.92 \pm 7.64}$
SAE-ZEROMASK	81.95 ± 14.99	83.88 ± 4.99	$\mathbf{82.10 \pm 4.91}$	60.00 ± 13.93	92.32 ± 8.74

TABLE IV
THE PERFORMANCE (%) OF MULTI-CLASS AD CLASSIFICATION WITH MR AND PET.

Methods	NC	ncMCI	cMCI	AD	ACC	SEN	SPE
SVM-MR	49.74 ± 8.79	44.58 ± 14.91	46.45 ± 31.63	53.74 ± 10.20	47.74 ± 1.82	$\mathbf{66.43 \pm 14.46}$	78.78 ± 8.13
SVM-PET	30.30 ± 20.15	36.90 ± 11.63	45.79 ± 27.08	50.30 ± 7.00	42.60 ± 2.90	35.36 ± 23.00	79.95 ± 8.33
MKSVM	47.71 ± 12.73	$\mathbf{52.76 \pm 19.33}$	38.17 ± 31.94	53.81 ± 6.81	48.65 ± 4.29	61.07 ± 18.95	79.86 ± 6.43
SAE-MR	47.80 ± 17.97	40.39 ± 9.46	45.08 ± 24.95	56.33 ± 14.03	45.61 ± 8.31	48.04 ± 14.97	82.69 ± 7.88
SAE-PET	41.79 ± 11.76	35.17 ± 10.10	41.06 ± 10.06	54.25 ± 11.79	42.91 ± 6.63	43.04 ± 17.45	82.26 ± 5.36
SAE-CONCAT	49.21 ± 14.74	43.54 ± 9.43	49.62 ± 9.66	56.35 ± 14.21	48.96 ± 5.32	46.61 ± 22.04	84.63 ± 8.51
2SAE-MKSVM	53.86 ± 11.47	52.08 ± 18.65	$\mathbf{53.17 \pm 26.63}$	55.58 ± 13.06	51.39 ± 5.64	66.25 ± 18.34	82.66 ± 6.16
SAE-ZEROMASK	$\mathbf{59.07 \pm 19.74}$	52.21 ± 11.84	40.17 ± 14.42	$\mathbf{64.07 \pm 15.24}$	$\mathbf{53.79 \pm 4.76}$	52.14 ± 11.81	$\mathbf{86.98 \pm 9.62}$

overall accuracy (82.59%) and overall sensitivity (86.83%). The overall specificities between these two methods were very closed (78.89% and 77.78%). The proposed method outperformed SVM in all overall performance measurements of classifying NC from MCI. The proposed method achieved 5% higher precision on classifying the normal control subjects.

The performance of multi-class classification is displayed in Table II. The first four columns are the precisions of the individual classes and the following three are the overall performance. The proposed method performed better precisions than SVM in three classes (52.40% on NC, 38.71% on cMCI and 46.89% on AD). The proposed method leads in the overall accuracy (46.30%) and overall specificity (77.78%). SVM achieved higher sensitivity (75.00%). In summary, our proposed method outperformed the state-of-the-art SVM-based methods in most of the performance measurements in both binary and multi-class AD classification problems when only MR data are presented.

2) Experiments on MR and PET (331 Subjects): There are totally 331 subjects with both MR and PET data available. We firstly evaluated the performance of SVM and the proposed SAE based method with only MR images (SVM-MR, SAE-MR) or PET images (SVM-PET, SAE-PET). The performance of fusing modalities with multi-kernel SVM is shown as MKSVM. For deep learning methods, we compared the proposed zero-masking training strategy (SAE-ZEROMASK) to the simple feature concatenation (SAE-CONCAT).

The binary classification performance is displayed in Table III. It can be observed that the experiments with both modalities (MKSVM, SAE-CONCAT and SAE-ZEROMASK) yielded better performance than those with only single modality in both binary classification tasks. SAE-CONCAT outperformed MKSVM slightly in the overall accuracy (90.15% - 90.11% and 77.65% - 76.88%). It can be observed that when the proposed SAE-ZEROMASK method is used, the performance is enhanced in all measurements comparing to SAE-CONCAT. MKSVM performed slightly higher specificity in classifying NC and AD comparing to ZERO-MASK. Though SVM-MR achieved slightly higher precision (83.92%) on MCI, it is reasonable to assume this performance may be due to an unbalanced decision making (only 67% on NC). Among all the methods, SAE-ZEROMASK achieved the most balanced performance in the classification between NC and MCI (81.95% on NC and 83.88% on AD), which is relatively difficult when MCI occupies a big proportion of the dataset (169 out of 246). The proposed data fusion method SAE-ZEROMASK with only one neural network achieved comparable performance with 2SAE-MKSVM, which fuses two high-level feature matrices from two independently trained networks. The accuracy of 2SAE-MKSVM was not obviously higher than that of simple feature concatenation (77.90% to 77.65%), because it was observed in the experiments the MKSVM added for feature fusion only preserved the higher accuracy achieved by a single modal network in some of the validation trials.

The performance of the multi-class classification is shown in Table IV. The proposed framework with the corrupted inputs (SAE-ZEROMASK) leads the overall accuracy and specificity (53.79% and 86.98%). Deep learning based methods (SAE-CONCAT and SAE-ZEROMASK) lead the precision on NC, cMCI and AD. The precision of cMCI was constrained by the quantity of cMCI instances (67 out of 331) and was effected by its sibling class ncMCI with 102 instances. For ncMCI the precision achieved by SAE-ZEROMASK and MKSVM were very closed. Comparing to the simple feature concatenation (SAE-CONCAT), SAE-ZEROMASK increased the overall accuracy by about 5%. SAE-ZEROMASK also outperformed the other data fusion option 2SAE-MKSVM in the overall accuracy and specificity. SVM-based methods tend to have better sensitivity.

IV. Discussion

A. Model Designing and Training

Studies have shown that learning architecture with multi-layered non-linear representations of the original data would yield meaningful features for classification [56], [61]–[63]. For accurate diagnosis in AD subjects, we investigated the use of multi-layered representations of neuroimaging biomarkers on AD classification. Our results showed that multi-layered structure can be used to distinguish MR and PET subjects along the spectrum of AD progression with higher accuracy than conventional shallow architectures. The performance of classification primarily benefited from the depth (a notion derived from complexity theory) of the learning architecture, which can be illustrated as a sequence of non-linear transformations of the feature space. During fine-tuning, the neuroimaging feature space is distorted and folded to minimise the classification loss on the training data. Thus, after several layers of transformations, the inseparable samples would become separable in the learnt high-level feature space. Compared to traditional methods, the proposed framework is more powerful in extracting the complex correlations between neuroimaging ROI based biomarkers as well as different feature modalities. Another motivation of using multi-layered structure for AD diagnosis is to reuse the high-level features for semi-supervised learning [64]. Besides the supervised data fusion or dimensionality reduction [29], the proposed workflow can be easily extended to use unlabelled neuroimaging data.

We combined different data modalities with the proposed zero-mask fusion strategy by propagating noisy signals with one modality randomly hidden. The auto-encoders were trained to reconstruct the original incoming signals with the corrupted incoming signals. We also tried to avoid training separate neural networks on different data modalities, because this may ignore the complementary information during feature learning. The training subjects with one hidden modality tend to force some neurons to be sensitive to MR and PET inputs, which makes the zero-mask fusion network different as it has two independent feature learning networks. It was noticeable that 2SAE+MKSVM also achieved an overall classification accuracy of 91.4% and a higher specificity of 91.67% in the binary classification of NC and AD. It may indicate that when relatively larger margins exist between different feature clusters, the binary decision boundaries might be similar between both feature fusion methods. Observing the

experimental results with non-convertible and convertible MCI subjects involved, we assume that the proposed zero-mask method may have more advantages when subtler differences and more outliers are included in a noisy training set.

Instead of using raw image patches for the medical feature learning, we applied the feature engineering pipeline to extract the initial ROI measurements of MR and PET images as inputs. The differences between 3D medical images of AD related patients tend to be subtle and the variance tends to be large. From this perspective, the hidden neurons of the network decision system can also be interpreted as automatically encoded inferences of diagnostic rules [65]. Our experiments showed that when using ROI precomputed features, the unsupervised network achieved the best performance with two hidden layers in pre-training. This means that relatively shallower architectures are practically required when using the approximately measured imaging features, compared with the learning tasks which use raw images as inputs [38]. The networks with the same number of neurons in all hidden layers often performed better in our experiments. We found that both over-completed manifolds or low-dimensional manifolds yielded effective features for AD classification. The number of hidden neurons was chosen according to different training sets.

The feature selection, using Elastic Net, enhanced the performance of all examined methods. It helped control the over-fitting caused by the noisy and redundant feature parameters. Notably, the majority of the selected feature parameters were consistently chosen by Elastic Net. The validation trials, with fewer chosen feature parameters, tended to have higher generalisation errors, which might be due to the biased outliers that were included in the training set.

Although the extracted features can be used by some other conventional classifiers, such as SVM, we connected an output layer with softmax regression to the unsupervised network. With a different non-linearity from the one used in other layers, softmax regression corresponded to multinomial log-output-variables. As a result, it is capable of classifying samples among several AD stages; it also simplified the fine-tuning phase of training because the softmax layer can be jointly optimised with the hidden layers. We also investigated the framework designs of transferring the fine-tuned features to popular classifiers other than the embedded softmax regressor. It was interesting to see that, taking as input the same high-level features learnt by our deep learning network, all of the investigated classifiers tended to make highly consistent decisions.

B. Limitations and Future Work

Considering the limited quantity of the available neuroimaging data, we assume that the synergy between different biomarkers can be further extracted with more training samples which may have smaller variance. The proposed data fusion strategy follows the de-noising fashion of training auto-encoders, which theoretically increased the difficulty of feature learning, but controlled the over-fitting. Although the predicted probability distribution of the 4-class AD classification may be of more practical use in a decision-making system, the performance that we achieved with the available dataset should be improved before multi-class classification frameworks are applied to clinical use. All the methods that we compared our methodology to, tended to over-fit, but had high accuracy on the training set and low accuracy on the testing set. Since the multi-modal learning architectures with neural networks (2SAE-MKSVM and SAE-ZEROMASK) are parametric models, we assume that they may have the potential to achieve better diagnostic accuracy on multi-class AD diagnosis when larger datasets are available. This will allow better extraction of subject-independent features with lower variance.

V. Conclusion

We propose a novel framework for the diagnosis of AD with deep learning embedded. The framework can distinguish four stages of AD progression with less clinical prior knowledge required. Since the unsupervised feature representation is embedded in this work-flow, it has potential to be extended to more unlabelled data for feature engineering in practice. In the unsupervised pre-training stage, we used stacked auto-encoders to obtain high-level features. When more than one neuroimaging modality was used, we applied a zero-masking strategy to extract the synergy between different modalities following a de-noising fashion. After the unsupervised feature engineering, a softmax regression was used. We used a novel method of visualising high-level brain biomarkers to analyse the high-level features that were extracted.

The proposed framework was evaluated with AD classification between stage two and four. Based on MR and PET ADNI data repository, our framework outperformed the state-of-the-art SVM based method and other deep learning frameworks. We argue that, therefore, the proposed method can be a powerful means to represent multi-modal neuroimaging biomarkers.

References

[1] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "Clinical Diagnosis of Alzheimer's Disease Report of the NINCDSADRDA Work Group Under the Auspices of Department of Health and Human Services Task Force on Alzheimer's Disease," *Neurology*, vol. 34, no. 7, pp. 939–939, 1984.

[2] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi, "Forecasting the Global Burden of Alzheimers Disease," *Alzheimer's & Dementia*, vol. 3, no. 3, pp. 186–191, 2007.

[3] B. Dubois, H. H. Feldman, C. Jacova, S. T. DeKosky, P. Barberger-Gateau, J. Cummings *et al.*, "Research Criteria for the Diagnosis of Alzheimer's Disease: Revising the NINCDSADRDA Criteria," *The Lancet Neurology*, vol. 6, no. 8, pp. 734–746, 2007.

[4] S. Gauthier, B. Reisberg, M. Zaudig, R. C. Petersen, K. Ritchie, K. Broich *et al.*, "Mild Cognitive Impairment," *The Lancet*, vol. 367, no. 9518, pp. 1262–1270, 2006.

[5] C. DeCarli, "Mild Cognitive Impairment: Prevalence, Prognosis, Aetiology, and Treatment," *The Lancet Neurology*, vol. 2, no. 1, pp. 15–21, 2003.

[6] C. Davatzikos, P. Bhatt, L. M. Shaw, K. N. Batmanghelich, and J. Q. Trojanowski, "Prediction of MCI to AD Conversion, via MRI, CSF Biomarkers, and Pattern Classification," *Neurobiology of Aging*, vol. 32, no. 12, pp. 2322.e19 – e27, 2011.

[7] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehricy, M. O. Habert *et al.*, "Automatic Classification of Patients with Alzheimer's Disease from Structural MRI: A Comparison of Ten Methods Using the ADNI Database," *Neuroimage*, vol. 56, no. 2, pp. 766–781, 2011.

[8] Y. Fan, S. M. Resnick, X. Wu, and C. Davatzikos, "Structural and Functional Biomarkers of Prodromal Alzheimer's Disease: A High-Dimensional Pattern Classification Study," *Neuroimage*, vol. 41, no. 2, pp. 277–285, 2008.

[9] S. Liu, W. Cai, L. Wen, and D. Feng, "Multi-Channel Brain Atrophy Pattern Analysis in Neuroimaging Retrieval," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2013)*. IEEE, 2013, pp. 206–209.

[10] S. Liu, W. Cai, L. Wen, and D. D. Feng, "Neuroimaging Biomarker based Prediction of Alzheimer's Disease Severity with Optimized Graph Construction," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2013)*. IEEE, 2013, pp. 1324–1327.

[11] W. Cai, S. Liu, Y. Song, S. Pujol, R. Kikinis, and D. D. Feng, "A 3D Difference of Gaussian based Lesion Detector for Brain PET," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2014)*. IEEE, 2014, pp. 677–680.

[12] G. Chetelat, B. Desgranges, V. De La Sayette, F. Viader, F. Eustache, and J. Baron, "Mild Cognitive Impairment Can FDG-PET Predict Who Is to Rapidly Convert to Alzheimers Disease?" *Neurology*, vol. 60, no. 8, pp. 1374–1377, 2003.

[13] R. Higdon, N. L. Foster, R. A. Koeppe, C. S. DeCarli, W. J. Jagust, C. M. Clark *et al.*, "A Comparison of Classification Methods for Differentiating Frontotemporal Dementia from Alzheimer's Disease Using FDG-PET Imaging," *Statistics in Medicine*, vol. 23, no. 2, pp. 315–326, 2004.

[14] N. L. Foster, J. L. Heidebrink, C. M. Clark, W. J. Jagust, S. E. Arnold *et al.*, "FDG-PET Improves Accuracy in Distinguishing Frontotemporal Dementia and Alzheimer's Disease," *Brain*, vol. 130, no. 10, pp. 2616–2635, 2007.

[15] S. Liu, W. Cai, L. Wen, S. Eberl, M. J. Fulham, and D. D. Feng, "A Robust Volumetric Feature Extraction Approach for 3D Neuroimaging Retrieval," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2010)*. IEEE, 2010, pp. 5657–5660.

[16] S. Liu, W. Cai, L. Wen, S. Eberl, M. J. Fulham, and D. Feng, "Localized Functional Neuroimaging Retrieval using 3D Discrete Curvelet Transform," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2011)*. IEEE, 2011, pp. 1877–1880.

[17] S. Liu, W. Cai, L. Wen, and D. D. Feng, "Multiscale and Multiorientation Feature Extraction with Degenerative Patterns for 3D Neuroimaging Retrieval," in *The 19th IEEE International Conference on Image Processing (ICIP 2012)*. IEEE, 2012, pp. 1249–1252.

[18] S. Liu, W. Cai, L. Wen, D. D. Feng, S. Pujol *et al.*, "Multi-Channel Neurodegenerative Pattern Analysis and Its Application in Alzheimer's Disease Characterization," *Computerized Medical Imaging and Graphics*, vol. 38, no. 4, pp. 436–444, 2014.

[19] F. H. Bouwman, S. N. M. Schoonenboom, W. M. van der Flier, E. J. van Elk, A. Kok, F. Barkhof *et al.*, "CSF Biomarkers and Medial Temporal Lobe Atrophy Predict Dementia in Mild Cognitive Impairment," *Neurobiology of Aging*, vol. 28, no. 7, pp. 1070–1074, 2007.

[20] S. Q. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. D. Feng, "Multi-Phase Feature Representation Learning for Neurodegenerative Disease Diagnosis," in *Australian Conference on Artificial Life and Computational Intelligence (ACALCI 2015)*, ser. Lecture Notes in Artificial Intelligence. Springer, 2015.

[21] S. L. Risacher, A. J. Saykin, J. D. West, L. Shen, H. A. Firpi, B. C. McDonald, and ADNI, "Baseline MRI Predictors of Conversion from MCI to Probable AD in the ADNI Cohort," *Current Alzheimer's Research*, vol. 6, no. 4, pp. 347–361, 2009.

[22] W. Cai, S. Liu, L. Wen, S. Eberl, M. J. Fulham, and D. D. Feng, "3D Neurological Image Retrieval with Localized Pathology-Centric CMRGlc Patterns," in *The 17th IEEE International Conference on Image Processing (ICIP 2010)*, 2010.

[23] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal Classification of Alzheimer's Disease and Mild Cognitive Impairment," *NeuroImage*, vol. 55, no. 3, pp. 856 – 867, 2011.

[24] S. Liu, Y. Song, W. Cai, S. Pujol, R. Kikinis, X. Wang, and D. D. Feng, "Multifold Bayesian Kernelization in Alzheimers Diagnosis," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2013)*, ser. Lecture Notes in Computer Science. Springer, 2013, pp. 303–310.

[25] S. Liu, W. Cai, L. Wen, and D. D. Feng, "Semantic-Word-Based Image Retrieval for Neurodegenerative Disorders," *Journal of Nuclear Medicine*, vol. 53, no. Supplement 1, p. 2309, 2012.

[26] F. Zhang, Y. Song, S. Liu, S. Pujol, R. Kikinis, M. J. Fulham, D. D. Feng, and W. Cai, "Semantic Association for Neuroimaging Classification of PET Images," *Journal of Nuclear Medicine*, vol. 55, no. Supplement 1, p. 2029, 2014.

[27] J. Liu, B. Shi, and Z. Wang, "Distance-Informed Metric Learning for Alzheimer's Disease Staging," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2014)*. IEEE, 2014.

[28] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer *et al.*, "Automatic Classification of MR Scans in Alzheimer's Disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008.

[29] N. Singh, A. Y. Wang, P. Sankaranarayanan, P. T. Fletcher, and S. Joshi, "Genetic, Structural and Functional Imaging Biomarkers for Early Detection of Conversion from MCI to AD," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2012)*, ser. Lecture Notes in Computer Science. Springer, 2012, pp. 132–140.

[30] S. Liu, L. Zhang, W. Cai, Y. Song, Z. Wang, L. Wen, and D. D. Feng, "A Supervised Multiview Spectral Embedding Method for Neuroimaging Classification," in *20th IEEE International Conference on Image Processing (ICIP 2013)*. IEEE, 2013, pp. 601–605.

[31] H. Che, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. D. Feng, "Co-neighbor Multi-View Spectral Embedding for Medical Content-based Retrieval," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2014)*. IEEE, 2014, pp. 911–914.

[32] J. Hastad, "Almost Optimal Lower Bounds for Small Depth Circuits," in *The 18th Annual ACM Symposium on Theory of Computing*. ACM, 1986, pp. 6–20.

[33] J. Hastad and M. Goldmann, "On the Power of Small-Depth Threshold Circuits," *Computational Complexity*, vol. 1, no. 2, pp. 113–129, 1991.

[34] S. Q. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. D. Feng, "Early Diagnosis of Alzheimer's Disease with Deep Learning," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2014)*, 2014, pp. 1015–1018.

[35] S. Liu, W. Cai, Y. Song, S. Pujol, R. Kikinis, L. Wen, and D. D. Feng, "Sparse Auto-Encoded Hypo-Metabolism Patterns in Alzheimer's Disease and Mild Cognitive Impairment," *Journal of Nuclear Medicine*, vol. 54, no. Supplement 2, p. 1807, 2013.

[36] S. Q. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, M. J. Fulham, and D. D. Feng, "High-level Feature based PET Image Retrieval with Deep Learning Architecture," *Journal of Nuclear Medicine*, vol. 55, no. Supplement 1, p. 2018, 2014.

[37] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[38] T. Brosch and R. Tam, "Manifold Learning of Brain MRIs by Deep Learning," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2013)*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013.

[39] H.-I. Suk, S.-W. Lee, and D. Shen, "Latent Feature Representation with Stacked Auto-Encoder for AD/MCI Diagnosis," *Brain Structure and Function*, pp. 1–19, 2013.

[40] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in A Deep Network with A Local Denoising Criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[41] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[42] L. Shen, S. Kim, Y. Qi, M. Inlow, S. Swaminathan, K. Nho *et al.*, "Identifying Neuroimaging and Proteomic Biomarkers for MCI and AD via the Elastic Net," in *Multimodal Brain Image Analysis (MBIA 2011)*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, vol. 7012, pp. 27–34.

[43] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy Layer-Wise Training of Deep Networks," *Advances in Neural Information Processing Systems*, vol. 19, pp. 153–160, 2007.

[44] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[45] C. Poultney, S. Chopra, and Y. L. Cun, "Efficient Learning of Sparse Representations with An Energy-Based Model," in *Advances in Neural Information Processing Systems*, 2006, pp. 1137–1144.

[46] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey *et al.*, "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI Methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[47] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster *et al.*, "A Probabilistic Atlas and Reference System for the Human Brain: International Consortium for Brain Mapping (ICBM)," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 356, no. 1412, pp. 1293–1322, 2001.

[48] J. A. Schnabel, D. Rueckert, M. Quist, J. M. Blackall, A. D. Castellano-Smith *et al.*, "A Generic Framework for Non-Rigid Registration Based on Non-Uniform Multi-Level Free-Form Deformations," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2001)*, ser. Lecture Notes in Computer Science. Springer, pp. 573–581.
[49] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images," *Neuroimage*, vol. 17, no. 2, pp. 825–841, 2002.
[50] R. A. Heckemann, S. Keihaninejad, P. Aljabar, K. R. Gray *et al.*, "Automatic Morphometry in Alzheimer's Disease and Mild Cognitive Impairment," *Neuroimage*, vol. 56, no. 4, pp. 2024–2037, 2011.
[51] S. Liu, W. Cai, L. Wen, S. Eberl, M. J. Fulham, and D. D. Feng, "Generalized Regional Disorder-Sensitive-Weighting Scheme for 3D Neuroimaging Retrieval," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2011)*. IEEE, 2011, pp. 7009–7012.
[52] J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, A. Ng, and Q. V. Le, "On Optimization Methods for Deep Learning," in *The 28th International Conference on Machine Learning (ICML 2011)*, 2011, pp. 265–272.
[53] W. Y. Zou, A. Y. Ng, and K. Yu, "Unsupervised Learning of Visual Invariance with Temporal Coherence," in *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*.
[54] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal Deep Learning," in *The 28th International Conference on Machine Learning (ICML 2011)*, 2011, pp. 689–696.
[55] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.
[56] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[57] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin *et al.*, "3D Slicer as An Image Computing Platform for The Quantitative Imaging Network," *Magnetic Resonance Imaging*, 2012.
[58] C. C. Chang and C. J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.
[59] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *The Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.
[60] Y. Bengio, *Practical Recommendations for Gradient-Based Training of Deep Architectures*. Springer, 2012, pp. 437–478.
[61] Y. Bengio and Y. LeCun, "Scaling Learning Algorithms Towards AI," *Large-Scale Kernel Machines*, vol. 34, 2007.
[62] Y. Bengio, O. Delalleau, and N. L. Roux, "The Curse of Highly Variable Functions for Local Kernel Machines," in *Advances in Neural Information Processing Systems*, 2005, pp. 107–114.
[63] Y. Bengio and O. Delalleau, "On the Expressive Power of Deep Architectures," in *Algorithmic Learning Theory*, ser. Lecture Notes in Computer Science, vol. 6925. Springer, 2011, pp. 18–36.
[64] J. Weston, F. Ratle, H. Mobahi, and R. Collobert, *Deep Learning via Semi-Supervised Embedding*. Springer, 2012, pp. 639–655.
[65] S. I. Gallant, *Neural Network Learning and Expert Systems*. MIT press, 1993.