Deep Ordinal Ranking for Multi-Category Diagnosis of Alzheimer's Disease using Hippocampal MRI data

Hongming Li, Mohamad Habes, Yong Fan

Section for Biomedical Image Analysis (SBIA), Center for Biomedical Image Computing and Analytics (CBICA), Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA

Abstraction

Increasing effort in brain image analysis has been dedicated to early diagnosis of Alzheimer's disease (AD) based on neuroimaging data. Most existing studies have been focusing on binary classification problems, e.g., distinguishing AD patients from normal control (NC) elderly or mild cognitive impairment (MCI) individuals from NC elderly. However, identifying individuals with AD and MCI, especially MCI individuals who will convert to AD (progressive MCI, pMCI), in a single setting, is needed to achieve the goal of early diagnosis of AD. In this paper, we propose a deep ordinal ranking model for distinguishing NC, stable MCI (sMCI), pMCI, and AD at an individual subject level, taking into account the inherent ordinal severity of brain degeneration caused by normal aging, MCI, and AD, rather than formulating the classification as a multi-category classification problem. The proposed deep ordinal ranking model focuses on the hippocampal morphology of individuals and learns informative and discriminative features automatically. Experiment results based on a large cohort of individuals from the Alzheimer's Disease Neuroimaging Initiative (ADNI) indicate that the proposed method can achieve better performance than traditional multi-category classification techniques using shape and radiomics features from structural magnetic resonance imaging (MRI) data.

Introduction

Alzheimer's disease (AD) is the most prevalent neurodegenerative disorder. As a major public health issue, this disease results in tremendous neurologic disability, emotional suffering, and financial difficulty for patients, their families, and the society at large. AD is characterized by tau pathology spreading from the medial temporal lobe and neocortical widespread amyloid beta deposition. Mild cognitive impairment (MCI) as a prodromal stage to AD, characterized by gradual neurodegeneration, is considered at a significantly higher risk to develop AD, with a conversion rate of 10-15% per year [1]. Although clinical criteria for MCI and early AD have been developed to formalize assessment of the gradual progression of cognitive and other symptoms in early AD, currently it is difficult to predict which individuals who meet criteria for MCI will ultimately progress to AD. Neuroimaging has been playing an increasingly important role for clinical AD diagnostics. As the search for effective therapies to arrest and slow the progression of AD intensifies, there is a need for better diagnostic and prognostic tools to identify individuals at high risk to progress to AD.

The hippocampus is one of the first brain structures affected by AD and undergoes severe structural changes [2]. The structural variation between the hippocampus of AD patients and healthy subjects has been studied intensively. Most of previous studies considered the hippocampus as a singular structure and focused on volumetric measures. Recently, more attention has been given to the fact that the hippocampus is heterogeneous in its formation, function, and relation to cognitive aging and neurodegeneration, with distinctive hippocampal subfields [3]. Studies have shown that early AD neurodegeneration was associated with initially

focal atrophy in the first cornu ammonis subfield (CA1), a potential marker for early AD detection [4]. Hippocampal subfields definition has been established through histological inspection based on a relatively small number of samples [5]. However, it remains largely unknown to what extent this heuristic definition helps improve early prediction of AD. The advances in neuroimaging techniques and analytics enabled to visualize the hippocampal and parahippocampal subregions *in vivo* as well. However, a major limitation in neuroimaging analysis of hippocampus subfields is the lack of general agreement on MRI-based protocol for subfields definition. Investigators have proposed a variety of manual delineation protocols with overall greatest disagreement in the definition of CA1 boundaries among other disagreements [6] and heuristic geometrical rules [6]. The disagreement in hippocampal subfields definitions could lead to inconsistent results [4] and limit our understanding in the biological mechanisms involved in hippocampal changes in early AD and across AD stages. Further important aspect has been largely ignored in hippocampal subfields investigation methods is the issue of reliability and reproducibility [7]. Most of reported intraclass correlation coefficients for inter/intra-rater disagreement for hippocampal subfields assessment were variable between methods and with unsatisfactory values for small regions [7]. Such disagreement makes the implementation of hippocampal subfields as relevant clinical biomarkers more of a future goal, waiting for an accurate localization by leveraging advanced methods for *in vivo* MRI [8]. Furthermore, most studies focused on hippocampal subfields volume assessment and complementary important features such as texture and shape have been largely ignored, although promising performance of hippocampus shape [9-11] and texture [12] has been demonstrated in AD prediction [13].

To aid AD diagnosis and distinguish MCI patients with higher risk of conversion to AD (progressive MCI, pMCI) from stable MCI individuals (sMCI), machine learning techniques have been proposed to build classifiers upon imaging data and clinical information [13-22], and identified prominent structural differences between pMCI and sMCI subjects at medial temporal lobe (MTL), including regions such as hippocampus and entorhinal cortex. Many studies have specifically focused on the hippocampus for early diagnosis of AD and build predictive models upon anatomical features including volume and shape based measures, and image intensity texture features [17, 18, 23-26]. Following the success of deep learning techniques in pattern recognition, convolutional neural networks (CNNs) have also been adopted to learn informative imaging features from 2D image patches of the hippocampus region for early diagnosis [26]. However, such a method does not fully take advantage of 3D information of the MRI data.

Most existing classification studies of AD have been focusing on two-category classification problems, e.g., distinguishing AD patients from NC elderly, MCI from NC or pMCI from sMCI. However, the early diagnosis of AD is essentially a multi-category classification problem, i.e., we need to identify individuals with AD, pMCI, and sMCI in a single setting. The multi-category classification problem associated with early diagnosis of AD can be solved in a typical multi-category classification framework, using strategies of one-against-one or one-against-the-rest [27]. However, such typical multi-category classification methods may overlook the ordinal information of the damage rendered by normal aging, MCI and AD [28]. Roughly speaking, brain changes rendered by normal aging, sMCI, pMCI, and AD comes with an increased severity of brain damage that is ranked, but difficult to be assigned with a metric value. The inter-subject variability might obliterate relatively small differences between NC and sMCI, between sMCI and pMCI, as well as between pMCI and AD, which makes a difficult task for discriminating different stages of AD progression. Since no proper metric distance can be defined for the ordinal damage severity, metric regression methods might be not good for the problem too.

To address these limitations and achieve early prediction of AD based on the hippocampal MRI data, we develop an ordinal ranking based deep learning method to simultaneously learn reproducible and discriminative features from the hippocampal MRI data and classify AD, pMCI, sMCI, and NC subjects, by making the best of inherent ordinal severity

of brain damage at AD's different stages. Since deep convolutional neural networks (CNNs) based feature learning is potentially able to capture complex relationship between imaging data and the ordinal severity of brain damages of AD, we adopt the CNNs to learn informative features from structural MRI data by optimizing a multi-output logistic regression model which encodes the ranking information of different stages of AD. We have evaluated the proposed method based on a large cohort of subjects from Alzheimer's Disease Neuroimaging Initiative (ADNI), and compared its performance with the state-of-the-art methods with multi-category classification capability. Experimental results have demonstrated that the proposed method could achieve improved prediction performance.

Materials and Methods

Image dataset

The data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu), consisting of baseline MRI scans of 1776 subjects from ADNI 1, Go and 2. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

The MRI data from ADNI 1 were used to train the proposed classification model and all other classification models under comparison, and these classification models were validated using MRI data of ADNI Go & 2. In particular, baseline MRI data in shared collections of the ADNI 1 were used as the training data, including 228 NC, 236 sMCI, 161 pMCI, and 192 AD subjects. Baseline MRI data of 959 subjects were obtained from the ADNI Go and 2, including 311 NC, 395 sMCI, 94 pMCI and 158 AD subjects. MCI subjects that converted to AD from 0.5 to 3 years from the baseline scan were labelled as pMCI, otherwise labelled as sMCI. The demographic information of all the data used is summarized in table 1.

Table 1. Demographic and clinical diagnosis information of the subjects used in this study.

ADNI		NC	sMCI	pMCI	AD
1	Age	75.97±5.02	75.03±7.67	74.58±7.00	75.34±7.45
	Sex (M/F)	118/110	156/80	100/61	101/91
	MMSE	29.11±1.00	27.31±1.78	26.63±1.69	23.31±2.04
GO& 2	Age	72.98±6.09	71.44±7.56	72.60±7.02	74.85±8.09
	Sex (M/F)	142/169	213/182	53/41	91/67
	MMSE	29.00±1.25	28.24±1.60	27.23±1.84	23.1±2.07

NC: cognitively normal control; MCI: mild cognitive impairment; sMCI: stable MCI; pMCI: progressive MCI; AD: Alzheimer's diease. MCI subjects that converted to AD from 0.5 to 3 years from the baseline scan were labelled as pMCI, otherwise labelled as sMCI.

Hippocampus extraction

T1 MRI scans of all the subjects were registered to the MNI space using affine registration. Left and right hippocampus regions were then extracted from the T1 images for each subject using LLL [29] algorithm with 100 hippocampus atlases obtained from a preliminary release of the EADC-ADNI harmonized segmentation protocol project (www.hippocampal-protocol.net) [30]. A 3D bounding box of size $29 \times 21 \times 55$ was adopted to extract hippocampus regions from the T1 image using the segmentation label of left and right hippocampus for each subject. These

hippocampus regions, referred to as hippocampus MRI images hereafter, were used as the input to the proposed deep ordinal ranking model.

Ordinal ranking

To make the best of the ordinal severity of brain change/damage rendered by normal aging, MCI, and AD, we propose an ordinal ranking method within an ordinal regression framework by transferring the ordinal ranking problem into a set of binary "larger than" problems [28, 31, 32]. Particularly, NC, sMCI, pMCI and AD are labeled using an ordinal order $y \in \{1,2,3,4\}$, corresponding to their severity of brain change/damage. Three binary "larger than" problems associated with the ordinal ranking problem are brain damage "larger than normal aging?" ($y > 1$), "larger than sMCI?" ($y > 2$), and "larger than pMCI"($y > 3$). The binary "larger than" problems are solved separately and then the binary codes obtained are fused to obtain the final multi-category classification label.

Given training data $\{(x_i, y_i), i = 1,2, \dots, n\}$, where x_i is the feature vector for subject i and $y_i \in \{1,2,3,4\}$ its associated label, the 4-category label could be transformed into a binary label for each binary "larger than" problem. For the k-th binary problem ($y > k$), its positively labeled training dataset X_k^+ and negatively labeled training dataset X_k^- could be constructed as

$$X_k^+ = \{(x_i, 1) | y_i > k\}, X_k^- = \{(x_i, 0) | y_i \leq k\}. \quad (1)$$

Based on the training dataset, a binary classifier f_k could be trained using any pattern classification techniques, such as support vector machine (SVM) [33] and random forests (RF) [34]. Once all the three binary classifiers are obtained, the ordinal ranking rule is constructed as

$$r(x) = 1 + \sum_{k=1}^{3} [\![f_k(x) > 0]\!], \quad (2)$$

where $[\![\cdot]\!]$ is 1 if the inner condition is true and 0 otherwise.

Deep ordinal ranking

Given the imaging data of hippocampus of each subject, different kinds of feature representations could be extracted, such as shape representation and radiomic characterization of image texture measures within the hippocampus regions [13]. Although these representations have been investigated and achieved promising performance, as hand-crafted features they might be not optimal and less discriminative for the AD diagnosis. Other than the hand-crafted feature extractors, the success of deep learning techniques in pattern recognition [35] in recent years have witnessed promising performance of deep learning techniques in learning imaging features for a variety of pattern recognition tasks [36-38]. In these studies, the convolutional neural networks (CNNs) are widely adopted to learn informative imaging features by optimizing a pattern recognition cost function. Ordinal regression based on CNNs has also been adopted for age estimation, and achieved better performance than state-of-the-art alternative techniques [39]. Therefore, we propose a deep ordinal model for AD diagnosis based on CNNs to learn informative and discriminative feature representation of the hippocampus and the mapping between the deep features and ordinal ranking in a data-driven way simultaneously.

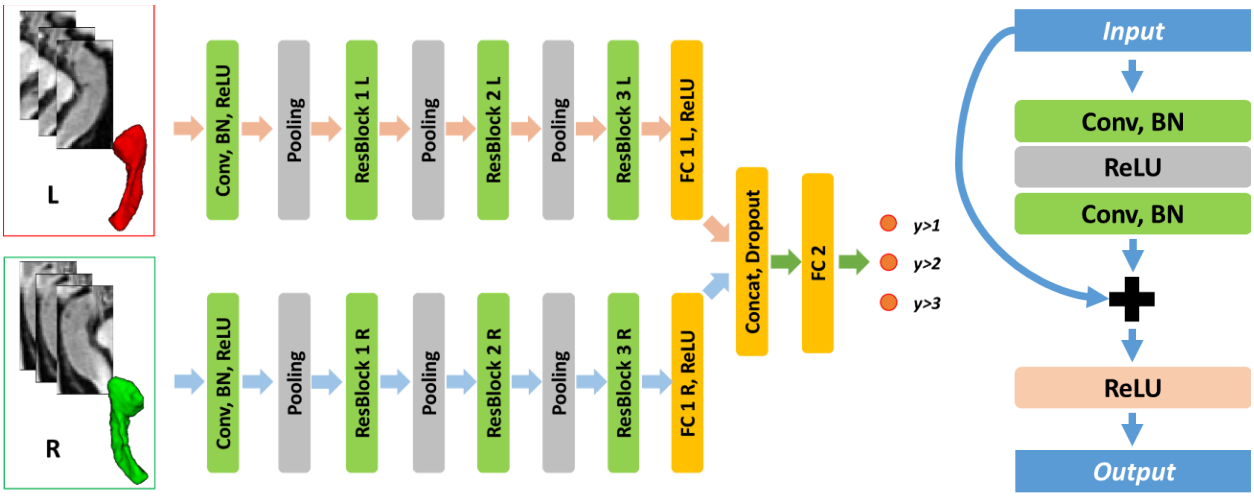

Fig. 1. Deep ordinal ranking model for early AD diagnosis. (a) schematic architecture of the deep network, (b) schematic residual block.

The network architecture of the proposed deep learning model is illustrated in Fig. 1a. Our model contains 1 convolutional layer (Conv), followed by 3 residual blocks (ResBlock), 1 fully connected layer (FC), and an output layer for the ordinal ranking. Rectified linear units (ReLU) is used as a nonlinear activation function for the convolutional and fully connected layers, batch normalization (BN) is adopted to accelerate deep network training [40], and max pooling layers are adopted to obtain features at multiple scales. As illustrated in Fig. 1b, the residual network structure has been adopted widely since its invention [41] and achieved promising performance in many challenging pattern recognition tasks. Several studies have also demonstrated that the residual connection would accelerate the convergence and improve the performance of the CNNs [42]. The left and right hippocampus regions are adopted as two-stream inputs to the deep model, which are gradually convolved by multiple 3D kernels within the subsequent Conv layer and ResBlock layers. The high-level feature representations of each hippocampus region are then flatten and connected to the FC layers, and concatenated and fed into the output layer.

To learn imaging features informative for the ordinal ranking with binary "larger than" classification problems, we formulate the ordinal ranking as a multi-label classification problem. In our study, the 4-category label of each subject $y_i \in \{1,2,3,4\}$ is transformed into a 3-bit binary label encoding its status corresponding to the 3 binary "larger than" problems, i.e., brain damage "larger than normal aging?", "larger than sMCI?", and "larger than pMCI?". For example, one AD patient will be labeled as $[1,1,1]$ in the deep ordinal ranking setting while labeled as $[0,0,0,1]$ in the regular 4-category classification setting. The output layer has 3 nodes corresponding to the three binary "larger than" problems. Sigmoid cross entropy loss is adopted to optimize the deep learning model.

Data augmentation

To boost the deep learning model's performance and robustness to image alignment and hippocampus segmentation errors, data augmentation is adopted to generate artificial training data [35]. Particularly, augmented image data were generated using image translation and non-rigid deformable image registration techniques. In particular, each hippocampus image along with its corresponding hippocampus masks in the training dataset was translated by 2 voxels along 26 directions of 3D image space separately, yielding augmented images that account for translation invariance for training the deep learning model. A non-rigid deformable image

registration method, namely ANTs [43], was adopted with its default parameter setting to register one hippocampus MRI image, referred to as moving image, to another of the same side (left to left and right to right) within the same disease category (NC to NC, sMCI to sMCI, pMCI to pMCI, and AD to AD), and the resulting deformation field was used to deform the moving hippocampus image and its hippocampus label to generate deformed hippocampus image and label. In total, 21242 spatial translated images, and 84824 non-rigid registered images were generated as the augmented dataset for training the deep learning model.

Validation and comparisons

We evaluated the proposed method and compared it with state-of-the-art alternative methods based on the same training and validation datasets.

State-of-the-art alternative methods under comparison

We compared the proposed method with the state-of-the-art feature extraction methods for hippocampal MRI images with the multi-category classification and ordinal ranking settings. The feature extraction methods for the hippocampal MRI images include shape characterization, radiomics, and deep CNNs. Details of these classification schemes are as following.

- Hippocampal representation. (1) shape characterization: 11 shape related features are extracted from the segmentation label of left and right hippocampi respectively, including volume, maximum 3D diameter, maximum 2D diameter (column, row, and slice), surface area, surface volume ratio, flatness, sphericity, elongation, and spherical disproportion [44]. (2) radiomics: image texture features are extracted from the hippocampal images and their counterparts after wavelet decomposition, including the first order features, gray level co-occurrence matrix (GLCM) features, gray level size zone matrix (GLSZM) features, and gray level run length matrix (GLRLM) features, and there are 711 textures in total for each hippocampus region. The shape and texture representation is calculated using the pyradiomics packages (http://pyradiomics.readthedocs.io) [44]. (3) deep CNNs: informative and discriminative features are automatically learned during the training procedure of forward convolution and back-propagation of deep CNNs.
- Classifier construction. (1) shallow classifier: Random forests (RF) [34] is adopted to construct classifier using shape and radiomics representation respectively. Its inherent feature selection and decision ensemble techniques lead to robust classification and better generalization. Moreover, RF can handle multi-category classification. (2) deep classifier: CNNs with the architecture shown in Fig. 1 is adopted for the prediction tasks based on the learned features in the data-driven way.
- Classification strategy. (1) multi-category classification: the early diagnosis of AD is formulated as a 4-category classification problem, RF using shape representation, RF using radiomics representation, and CNNs with 4-category output are evaluated in this study. For the CNNs, the network architecture is the same as that in Fig.1 except that the output layer is replaced with 4-node output layer for multi-category classification. (2) ordinal ranking classification: For the shallow classifiers, 3 “larger than” binary classifiers are constructed using RF based on shape and radiomics representation respectively. For the deep classifier, the proposed deep ordinal model as illustrated in Fig.1 is adopted. Note that the same network architecture and parameter configuration is adopted for deep classifier under both multi-category and ordinal ranking setting except the differences of the output layer.

In addition to the 4-category classification, we have also performed a binary classification to distinguish AD patients from NC elderly based on shape, radiomics, and deep representation as baseline experiments, in order to investigate the discriminative power of hippocampal representation for AD diagnosis.

The performance of the classification is evaluated with the following metrics: (1) normalized confusion matrix, (2) adjusted classification accuracy, (3) receiver operating characteristic curve (ROC), and area under ROC (AUC). A normalized confusion matrix illustrates not only the sensitivity and specificity for the multi-category classification results, but also the pattern of misclassification reflecting the severity of different stages of AD disease. Adjusted classification accuracy is calculated as the mean sensitivity value of the 4 categories, which takes the imbalance of sample sizes of different categories into consideration. For the binary AD versus NC prediction, ROC and AUC are adopted for the evaluation.

Experimental settings

The deep learning model's network architecture is illustrated by Fig. 1, with 1 Conv layer, 3 ResBlocks, 1 FC layer, and an output layer. In particular, the Conv layer contains 64 kernels, while the ResBlock 1, 2, and 3 contains 64, 128, and 128 kernels respectively. The kernel size for all the kernels is $3 \times 3 \times 3$. A stride of 2 and kernel size of 2 is used for the max pooling layer. The fully connected layer FC1 contains 256 nodes, which extract a 256-dimensional features for left and right hippocampus respectively. The two 256-dimensional feature vector is concatenated and fed to FC2 with 3 output nodes for the deep ordinal ranking model (4 output nodes for the deep multi-category classification model). A dropout operation with a ratio of 0.5 is applied before the features fed into the last FC layer.

The deep learning model was optimized using stochastic gradient descent (SGD) algorithm [45], the momentum was set to 0.9, and the base learning rate was set to 5×10^{-5}. The learning rate was updated using a stepwise policy, which drops the learning rate by a factor of 0.1 after every 40000 steps. The maximum iteration of the training procedure was set to 120000. Batch size of 32 was adopted to update weights in the model. The deep learning models was implemented using Caffe [46], and trained on a Nividia Titan X (Pascal) graphics processing unit (GPU).

For the RF based on shape representation and radiomics representation, 1000 decision trees were adopted for the forest, and the minimum leaf size of the tree was set to 5. Sample weight for each training image was set to the ratio between total number of training images and the number of images within the same category, and the training images were sampled with replacement during the training procedure. The built-in RF implementation TreeBagger in Matlab (R2013a) was adopted to train the model, and default values were used for other parameters.

Results

The hippocampus volumes of all the subjects are illustrated in Fig. 2. These plots indicated that AD patients and NC elderly could be roughly separated based on their hippocampus volumes. However, the hippocampus volumes of MCI individuals scatter in-between the AD group and NC group, making it non-trivial to distinguish all the 4 groups based on the hippocampus volume only.

Two experiments were conducted to evaluate the performance of the proposed method. We first performed a binary classification task for distinguishing the AD patients from the NC individuals using the shape representation, radiomics representation and deep representation respectively with a two-fold purpose. On one hand, we would like to check the power of hippocampus based representation for the AD diagnosis, on the other hand, we would like to investigate if the representation learned based on the deep CNN model is more discriminative for the prediction task. We then performed the 4-category prediction based on the 3 kinds of hippocampus representation under multi-category classification and ordinal ranking setting, to investigate if improved prediction performance could be achieved by our proposed deep ranking model. It is worth noting that all the prediction models were trained using the ADNI I dataset, and validated using the ADNI Go & 2 dataset.

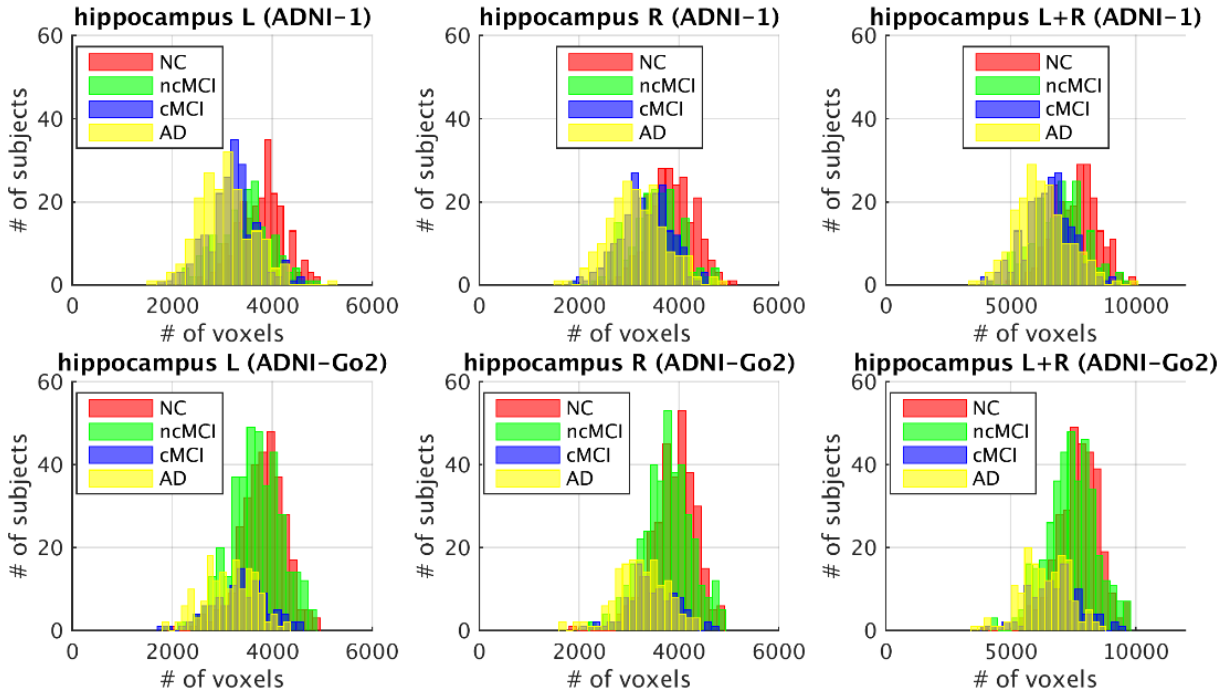

Fig. 2. Histograms of hippocampus volume measures of all the subjects (top row: ADNI-1 and bottom row: ADNI-GO & 2) used in this study. Volume measures of left and right hippocampi and their combination are shown from left to right. Each voxel has a spatial resolution of $1x1x1mm^3$.

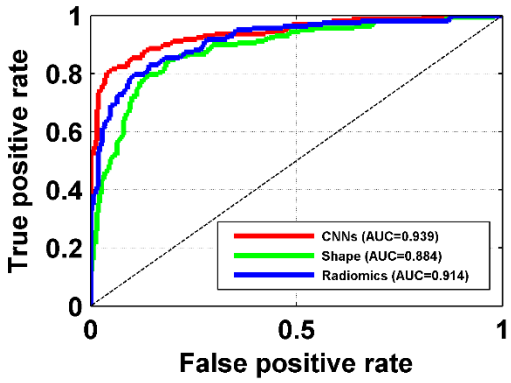

Fig. 3. ROC curves obtained based on different hippocampus representations for AD versus NC prediction.

Fig. 3 shows the ROC curve of the binary classification on the validation dataset using shape, radiomics and deep hippocampus representation respectively. The AUC obtained by the deep representation was 0.939, while that based on shape representation and radiomic representation were 0.884 and 0.914 respectively. As demonstrated in the figure, all the 3 representations were quite powerful for distinguishing the patients from health individuals, indicating that hippocampus based representation could indeed characterize the anatomical alternations along the progression of disease. Moreover, radiomics representation got better performance than the shape based representation, and CNNs based representation got the best

performance, indicating that intensity variations within hippocampus could provide more discriminative information, and features learned in the data-driven manner could capture the task related characteristics better than the conventional hand-crafted features.

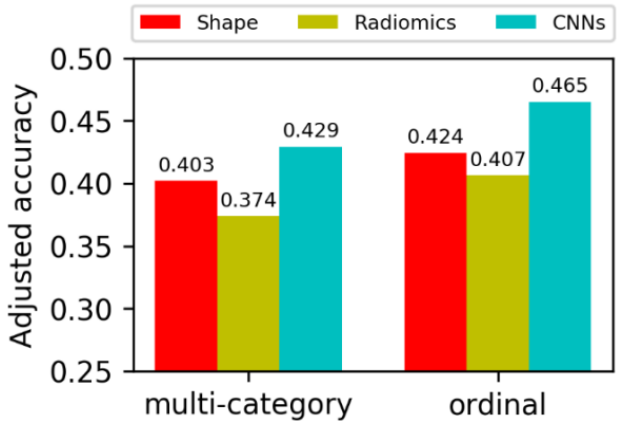

Fig. 4. Adjusted accuracy for the 4-category prediction under different classification setting.

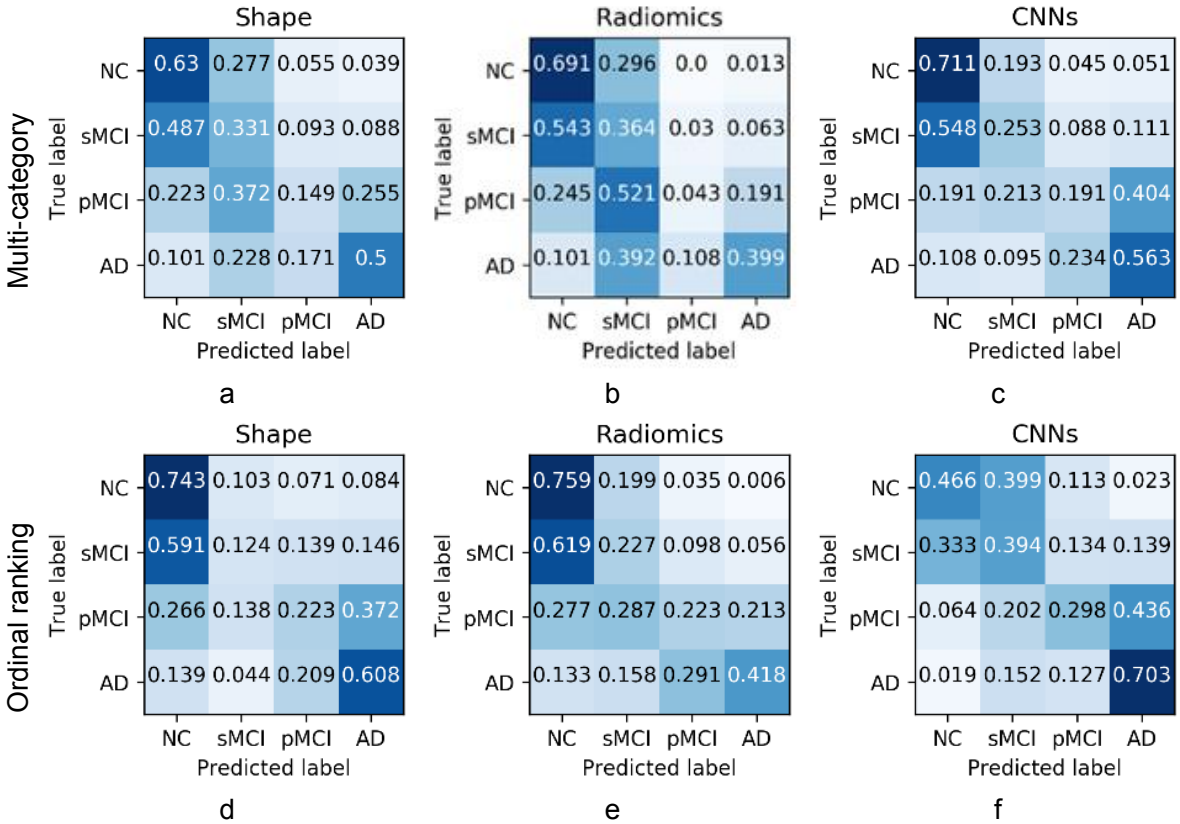

Fig. 5. Confusion matrices of 4-category prediction. (a-c) results obtained based on different hippocampus representations under multi-category classification setting, (d-f) corresponding results under ordinal ranking setting.

The results of 4-category prediction were illustrated in Fig. 4 and Fig. 5. Fig. 4 shows the adjusted accuracy for the prediction. It could be seen that the CNNs model obtained better performance than that obtained by RF using shape and radiomics representations, and the performance under ordinal ranking setting were generally better than their counterparts under

multi-category classification setting. The best performance was obtained by our proposed deep ordinal ranking model, and the overall accuracy is 0.465. Fig. 5 illustrates the confusion matrices of all the 6 prediction models. Generally speaking, the AD group and NC group were separated pretty well for all the models, our proposed deep ordinal ranking model captured the progressive patterns of the AD better than other models, as the larger coefficients of the confusion matrix located at the nearby positions along the diagonal of the matrix, indicating that misclassified subjects were assigned to adjacent categories in the progression spectrum, instead of the distant categories.

To investigate how the different hippocampus representations contributed to the classification, we have projected the different hippocampus representations onto a 2D plane using the t-SNE algorithm [47], as shown in Fig. 6. The 4 subplots correspond to the shape representation, radiomics representation, deep representation learned by multi-category CNNs, and deep representation learned by our ordinal ranking CNNs. As illustrated in the figure, for the shape and radiomics representations, the distribution of the sMCI and pMCI individuals were largely overlapped with that of AD and NC individuals, which limited the discriminative power of the corresponding prediction models. For the deep representations, a relative clear progressive pattern could be observed, where the AD and NC individuals distributed around two poles while the sMCI and pMCI individuals spanned in between. The visualization also indicated that the learned representations were more informative and facilitated the subsequent prediction.

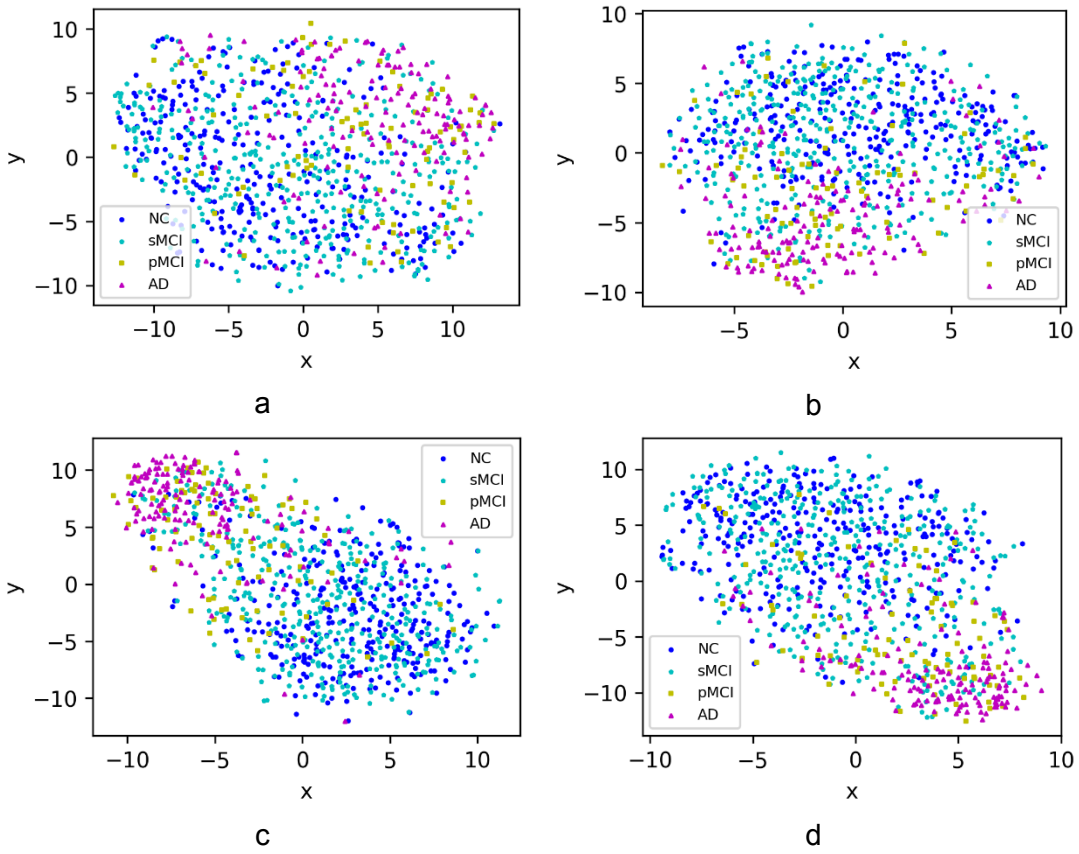

Fig. 6. t-SNE visualization of different hippocampus representation. (a) shape, (b) raidomics, (c) multi-category CNNs, and (d) ordinal ranking CNNs.

Discussions

As one of the first brain structures affected by AD, the structural variation of hippocampus between AD patients and healthy subjects has been studied intensively. Although several studies have been proposed to extract different feature representations of hippocampus from structural MR imaging data for computer-aided AD diagnosis, most of them focus on shape related features or conventional hand-crafted features, which might not be optimal and most discriminative for the diagnosis task. Moreover, these studies generally focus on binary classification instead of 4-category prediction covering all the stages of AD progression which is more clinical oriented, and do not take the intrinsic ordinal severity of different stages of AD into account. To this end, we develop a deep ordinal ranking framework to automatically extract hippocampus representation from MR images in a data-driven way and the mapping between them and the ordinal AD staging information simultaneously. In particular, we get the deep representations for left and right hippocampus respectively using deep CNNs which extract imaging features hierarchically. The regular 4-category labels are transformed into 3 ordinal labels which are used for optimizing the multi-output loss function to drive the whole deep learning model. Experimental results on the large ADNI dataset suggest that the proposed method could help boost the prediction performance. It would be straightforward to integrate multimodal imaging data and biological/clinical measures using the convolutional layers and fully connected layers.

Several quantitative measures about hippocampus have been explored to investigate their discriminative power to distinguish AD patients from NC individuals, from simple volume, to geometric shape measures, to intensity based imaging features such as textures. Though statistical differences between patients and health individuals and promising classification performance have been reported based on these measures, which has also been evaluated as in our binary classification experiment, the discriminative power of these hand-crafted measures are still limited, especially when used for more complex classification tasks, such as the 4-category classification. As illustrated in Fig. 6 (a, b), they are unable to separate the distribution of sMCI and pMCI individuals as the intermediate stage between AD and NC, and therefore the corresponding prediction performance are hindered, as shown in Fig. 5(a, b, d, e). Unlike the fixed feature extractor of the hand-crafted measures, the deep CNNs could extract relevant features tailored for the specific requested task, the feature extractor could be optimized during the procedure of model training. As illustrated in Fig. 6 (c, d), individuals of AD and NC are further separated compare to the hand-crafted features, while the overall distribution of 4 categories shows relatively more clear transition from one pole (NC) to the other (AD). The more informative features also promote the prediction performance as shown in Fig. 5 (c, f).

Instead of formulating the AD diagnosis as binary classification that accounts for 2 out 4 stages of AD progression, or as regular multi-category classification ignoring the progressive property of adjacent stages, we formulate the diagnosis task under an ordinal ranking framework. The ordinal ranking framework can naturally consider the degrees of brain degeneration along with the disease progression. While under multi-category classification setting, one subject might be misclassified into one arbitrary category, more penalty would be introduced to the prediction model if one pMCI individual is assigned to the NC instead of sMCI, as NC is distant from pMCI on the ordinal list. This has also been demonstrated in Fig. 4 and 5. All the prediction model under ordinal setting outperforms their multi-category counterparts, and the pattern of the prediction results follows the disease progression better, as shown in Fig. 5f in particular, most of the incorrectly assigned individuals were located at adjacent categories of their true category.

Although the proposed deep ordinal ranking model has achieved promising performance for AD diagnosis, further efforts are needed for the following aspects. First, the current study focuses on the role of hippocampus in AD diagnosis, and obtains comparable performance with

that based on whole-brain features [48], incorporating more region of interests (ROIs) affected by the disease would provide complementary information and lead to improved prediction performance. Second, hyper-parameters of the deep ordinal modeling need further optimization, including network architecture, convolutional filter size, learning rate, batch size, number of filters per convolution layer, and so on [35]. Bayesian optimization methods could be used to tune our models [49, 50], and we have faith in that potential performance improvement could be obtained. Moreover, the definition of pMCI category might influence the performance of the diagnosis. Conversion to AD within 2 or 3 years are generally used for the identification of pMCI in the literature, but there is still no common sense regarding this, more concerns need to be considered. Also, imaging data from ADNI 1 were used as training dataset and that from ADNI Go & 2 were used as independent validation dataset. While the imaging data from ADNI I are acquired using 1.5T scanner and that of ADNI Go & 2 using 3T scanner, it is also interesting to investigate how this affects the imaging based AD diagnosis.

Conclusion

In this paper, we have presented a deep ordinal ranking model for classifying AD's different stages using structural imaging data focusing on the hippocampus morphology, built on CNNs and ordinal ranking techniques. The comparison with the traditional multi-category classification methods on the large cohort of ADNI dataset shows that our method can achieve promising performance, indicating that the utilization of inherent ordinal severity of brain damage rendered by AD's different stages can help achieve improved classification performance. Moreover, the deep hippocampus representation learned by the deep model also outperform relatively simple imaging representations, i.e., shape and radiomics features. Benefiting from the flexible architecture of proposed deep model, the performance of our method might be further improved if multi-modality information is taken into account, e.g., PET imaging and CSF biomarkers. Besides classification, our proposed method is also a better fit for regression studies of AD associated clinical score estimation than simple metric regression, since most of the clinical score measures, e.g., mini mental state examination (MMSE), are not continuous variables.

References

1. Grundman, M., et al., *Mild cognitive impairment can be distinguished from alzheimer disease and normal aging for clinical trials.* Archives of Neurology, 2004. **61**(1): p. 59-66.
2. Braak, H. and E. Braak, *Neuropathological stageing of Alzheimer-related changes.* Acta neuropathologica, 1991. **82**(4): p. 239-259.
3. Small, S.A., et al., *A pathophysiological framework of hippocampal dysfunction in ageing and disease.* Nature reviews. Neuroscience, 2011. **12**(10): p. 585-601.
4. de Flores, R., R. La Joie, and G. Chetelat, *Structural imaging of hippocampal subfields in healthy aging and Alzheimer's disease.* Neuroscience, 2015. **309**: p. 29-50.
5. Duvernoy, H.M., *The human hippocampus: functional anatomy, vascularization and serial sections with MRI*. 2005: Springer Science & Business Media.
6. Yushkevich, P.A., et al., *Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment.* Human brain mapping, 2015. **36**(1): p. 258-287.
7. Giuliano, A., et al., *Hippocampal subfields at ultra high field MRI: An overview of segmentation and measurement methods.* Hippocampus, 2017. **27**(5): p. 481-494.
8. van Strien, N.M., et al., *Imaging hippocampal subregions with in vivo MRI: advances and limitations.* Nature reviews. Neuroscience, 2011. **13**(1): p. 70.

9. Li, S., et al., *Hippocampal shape analysis of Alzheimer disease based on machine learning methods.* AJNR. American journal of neuroradiology, 2007. **28**(7): p. 1339-1345.
10. Costafreda, S.G., et al., *Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment.* NeuroImage, 2011. **56**(1): p. 212-219.
11. Gerardin, E., et al., *Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging.* NeuroImage, 2009. **47**(4): p. 1476-1486.
12. Sorensen, L., et al., *Early detection of Alzheimer's disease using MRI hippocampal texture.* Human brain mapping, 2016. **37**(3): p. 1148-1161.
13. Rathore, S., et al., *A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages.* NeuroImage, 2017. **155**: p. 530-548.
14. Desikan, R.S., et al., *Automated MRI measures predict progression to Alzheimer's disease.* Neurobiol Aging, 2010. **31**(8): p. 1364-74.
15. Filipovych, R., C. Davatzikos, and I. Alzheimer's Disease Neuroimaging, *Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI).* Neuroimage, 2011. **55**(3): p. 1109-19.
16. Moradi, E., et al., *Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects.* Neuroimage, 2015. **104**: p. 398-412.
17. de Vos, F., et al., *Combining multiple anatomical MRI measures improves Alzheimer's disease classification.* Hum Brain Mapp, 2016. **37**(5): p. 1920-9.
18. Hu, K., et al., *Multi-scale features extraction from baseline structure MRI for MCI patient classification and AD early diagnosis.* Neurocomputing, 2016. **175, Part A**: p. 132-145.
19. Fan, Y., et al., *Structural and functional biomarkers of prodromal Alzheimer's disease: A high-dimensional pattern classification study.* Neuroimage, 2008. **41**(2): p. 277-285.
20. Davatzikos, C., et al., *Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging.* Neurobiology of Aging, 2008. **29**(4): p. 514-523.
21. Fan, Y., et al., *Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline.* Neuroimage, 2008. **39**(4): p. 1731-1743.
22. Misra, C., 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, 2009. **44**(4): p. 1415-1422.
23. Chupin, M., et al., *Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI.* Hippocampus, 2009. **19**(6): p. 579-87.
24. Devanand, D.P., et al., *MRI hippocampal and entorhinal cortex mapping in predicting conversion to Alzheimer's disease.* Neuroimage, 2012. **60**(3): p. 1622-9.
25. Ben Ahmed, O., et al., *Classification of Alzheimer's disease subjects from MRI using hippocampal visual features.* Multimedia Tools and Applications, 2015. **74**(4): p. 1249-1266.
26. Aderghal, K., et al., *Classification of sMRI for AD Diagnosis with Convolutional Neuronal Networks: A Pilot 2-D+ε Study on ADNI*, in *MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part I*, L. Amsaleg, et al., Editors. 2017, Springer International Publishing: Cham. p. 690-701.
27. Chih-Wei, H. and L. Chih-Jen, *A comparison of methods for multiclass support vector machines.* IEEE Transactions on Neural Networks, 2002. **13**(2): p. 415-425.
28. Fan, Y., *Ordinal Ranking for Detecting Mild Cognitive Impairment and Alzheimer's Disease Based on Multimodal Neuroimages and CSF Biomarkers*, in *Multimodal Brain Image Analysis: First*

International Workshop, MBIA 2011, Held in Conjunction with MICCAI 2011, Toronto, Canada, September 18, 2011. Proceedings, T. Liu, et al., Editors. 2011, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 44-51.
29. Hao, Y., et al., *Local label learning (LLL) for subcortical structure segmentation: Application to hippocampus segmentation.* Human brain mapping, 2014. **35**(6): p. 2674-2697.
30. Boccardi, M., et al., *Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol.* Alzheimers & Dementia, 2015. **11**(2): p. 175-183.
31. Li, L. and H.-T. Lin, *Ordinal regression by extended binary classification*, in *Proceedings of the 19th International Conference on Neural Information Processing Systems*. 2006, MIT Press: Canada. p. 865-872.
32. Chang, K.Y., C.S. Chen, and Y.P. Hung. *Ordinal hyperplanes ranker with cost sensitivities for age estimation*. in *CVPR 2011*. 2011.
33. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine Learning, 1995. **20**(3): p. 273-297.
34. Tin Kam, H., *The random subspace method for constructing decision forests.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998. **20**(8): p. 832-844.
35. Goodfellow, I., Y. Bengio, and A. Courville, *Deep Learning*. 2016: MIT Press.
36. Nie, D., et al., *3D Deep Learning for Multi-modal Imaging-Guided Survival Time Prediction of Brain Tumor Patients.* Med Image Comput Comput Assist Interv, 2016. **9901**: p. 212-220.
37. Esteva, A., et al., *Dermatologist-level classification of skin cancer with deep neural networks.* Nature, 2017. **542**(7639): p. 115-118.
38. Gulshan, V., et al., *Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs.* JAMA, 2016. **316**(22): p. 2402-2410.
39. Niu, Z., et al. *Ordinal Regression with Multiple Output CNN for Age Estimation*. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
40. Ioffe, S. and C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, in *ICML*, F.R. Bach and D.M. Blei, Editors. 2015, JMLR.org. p. 448-456.
41. He, K., et al. *Deep Residual Learning for Image Recognition*. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
42. Szegedy, C., et al. *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*. 2017.
43. Avants, B.B., et al., *A reproducible evaluation of ANTs similarity metric performance in brain image registration.* Neuroimage, 2011. **54**(3): p. 2033-44.
44. Griethuysen, J.J.v., et al., *Computational Radiomics System to Decode the Radiographic Phenotype.* Cancer Research, 2017.
45. Boyd, S. and L. Vandenberghe, *Convex optimization*. 2004: Cambridge university press.
46. Jia, Y., et al. *Caffe: Convolutional architecture for fast feature embedding*. in *Proceedings of the 22nd ACM international conference on Multimedia*. 2014. ACM.
47. Maaten, L.v.d. and G. Hinton, *Visualizing data using t-SNE.* Journal of Machine Learning Research, 2008. **9**(Nov): p. 2579-2605.
48. Liu, S., et al., *Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease.* IEEE Trans Biomed Eng, 2015. **62**(4): p. 1132-40.
49. Snoek, J., H. Larochelle, and R.P. Adams, *Practical Bayesian Optimization of Machine Learning Algorithms*, in *Advances in Neural Information Processing Systems*. 2012. p. 1-9.
50. SPEARMINT. *https://github.com/JasperSnoek/spearmint/tree/master/spearmint*. 2017.