NeuroImage: Clinical 15 (2017) 613–624

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Five-class differential diagnostics of neurodegenerative diseases using random undersampling boosting

Tong Tong[a,b,*], Christian Ledig[a], Ricardo Guerrero[a], Andreas Schuh[a], Juha Koikkalainen[c,l], Antti Tolonen[c], Hanneke Rhodius[d], Frederik Barkhof[f,g], Betty Tijms[d], Afina W Lemstra[d], Hilkka Soininen[h,i], Anne M Remes[h,i], Gunhild Waldemar[j], Steen Hasselbalch[j], Patrizia Mecocci[k], Marta Baroni[k], Jyrki Lötjönen[c,l], Wiesje van der Flier[d,e], Daniel Rueckert[a]

[a] Department of Computing, Imperial College London, London, UK
[b] Laboratory for Computational Neuroimaging, Athinoula A. Martinos Center for Biomedical Imaging, MGH/Harvard Medical School, Charlestown, USA
[c] VTT Technical Research Centre of Finland, Tampere, Finland
[d] Alzheimer Center, Department of Neurology, VU University Medical Centre, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands
[e] Department of Epidemiology and Biostatistics, VU University Medical Centre, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands
[f] Department of Radiology and Nuclear Medicine, VU University Medical Centre, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands
[g] Institutes of Neurology and Healthcare Engineering, UCL, London, UK
[h] Department of Neurology, Institute of Clinical Medicine, University of Eastern Finland, Kuopio, Finland
[i] Department of Neurology, Kuopio University Hospital, Kuopio, Finland
[j] Department of Neurology, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark
[k] Section of Gerontology and Geriatrics, University of Perugia, Perugia, Italy
[l] Combinostics Ltd., Tampere, Finland

ARTICLE INFO

Keywords:
Neurodegenerative diseases
Differential diagnosis
MRI
Dementia
Imbalance learning
Multi-class feature selection

ABSTRACT

Differentiating between different types of neurodegenerative diseases is not only crucial in clinical practice when treatment decisions have to be made, but also has a significant potential for the enrichment of clinical trials. The purpose of this study is to develop a classification framework for distinguishing the four most common neurodegenerative diseases, including Alzheimer's disease, frontotemporal lobe degeneration, Dementia with Lewy bodies and vascular dementia, as well as patients with subjective memory complaints. Different biomarkers including features from images (volume features, region-wise grading features) and non-imaging features (CSF measures) were extracted for each subject. In clinical practice, the prevalence of different dementia types is imbalanced, posing challenges for learning an effective classification model. Therefore, we propose the use of the RUSBoost algorithm in order to train classifiers and to handle the class imbalance training problem. Furthermore, a multi-class feature selection method based on sparsity is integrated into the proposed framework to improve the classification performance. It also provides a way for investigating the importance of different features and regions. Using a dataset of 500 subjects, the proposed framework achieved a high accuracy of 75.2% with a balanced accuracy of 69.3% for the five-class classification using ten-fold cross validation, which is significantly better than the results using support vector machine or random forest, demonstrating the feasibility of the proposed framework to support clinical decision making.

1. Introduction

Neurodegeneration is a progressive process that results in the gradual loss of nerve structure and function. The neurodegenerative process occurs with normal aging, but can be accelerated by many neurodegenerative diseases (NDs), including Alzheimer's disease (AD), frontotemporal lobe degeneration (FTLD), dementia with Lewy bodies (DLB), and vascular dementia (VaD). Identifying subjects with a specific dementia type is not only crucial in clinical practice, but also beneficial for developing new treatments and enriching clinical trials. However, the symptoms of different NDs have a high degree of similarity, making differential diagnostics difficult. Although there are established clinical guidelines (Neary et al., 1998; McKhann et al., 1984, 2011; Román et al., 1993; McKeith et al., 2005) for the diagnosis of different NDs, they are relatively general and require significant expertise from clinicians in order to reach a correct diagnosis. Therefore, it is essential to

* Corresponding author at: Biomedical Image Analysis Group, Department of Computing, Imperial College London, London, UK.
E-mail address: t15008@fzu.edu.cn (T. Tong).

http://dx.doi.org/10.1016/j.nicl.2017.06.012
Received 9 March 2017; Received in revised form 7 May 2017; Accepted 8 June 2017
Available online 12 June 2017

develop computer-aided decision support systems that can increase the confidence and accuracy of differential diagnostics of NDs in clinical practice.

Neuroimaging techniques have been widely used to detect pathological changes associated with NDs. For example, magnetic resonance (MR) T1-weighted imaging has been successfully used in the detection of atrophy patterns in subjects with AD (Cuingnet et al., 2011; Tong et al., 2014; Thung et al., 2014), FTLD (Du et al., 2007), DLB (Whitwell et al., 2007) and VaD (Zarow et al., 2005). The atrophy patterns including the affected regions and the atrophy rates associated with different types of dementia are different, thus providing discriminative information for the differential diagnostics of these NDs. Features such as subcortical volumes and cortical thickness which reflect the atrophy patterns can be used as biomarkers for differential diagnostics. In addition, white matter changes (i.e. hyperintensity) are typical for patients with VaD and can be well detected using Fluid Attenuated Inversion Recovery (FLAIR) imaging. This provides additional information for distinguishing VaD from the other dementia types.

Among different NDs, AD is the most common cause of dementia, accounting for 60%–70% of all dementia cases (Frisoni et al., 2010), while every other ND accounting for less than 20% (McKeith et al., 2004; Van Straaten et al., 2004; Du et al., 2007). This imbalance in the prevalence of different dementias poses challenges for learning an effective classification model as most commonly used classification algorithms tend to favour classifying subjects to the majority class (i.e. AD) compared to the minority classes (i.e. VaD and DLB). For example, in our dataset, the number of AD subjects is about ten times that of VaD subjects and about five times that of DLB subjects. Therefore, machine learning techniques that can handle the problem of class imbalance are required in order to learn an effective model for the differential diagnostics. There are two major categories of approaches to handle class imbalance. One category of approaches is based on cost-sensitive learning (Zhang and Zhou, 2010; Thai-Nghe et al., 2010). In these methods, a high cost is assigned to the misclassification of minority classes while simultaneously minimizing the overall cost. The other major category uses a sampling technique (Chawla et al., 2002; Seiffert et al., 2010) to create a balanced dataset for training. In our work, we used the second one to alleviate the effect of an imbalanced dataset on the classifier training. Specifically, the RUSBoost algorithm proposed in Seiffert et al. (2010) was adopted since it is simpler, faster and can achieve better performance (Seiffert et al., 2010) compared to other approaches such as AdaBoost (Rätsch et al., 2001) and SMOTEBoost (Chawla et al., 2003).

In addition to obtaining accurate differential diagnostics, it is also interesting to know which biomarkers are most important in the diagnostics. Multiple biomarkers can be extracted from MR T1-weighted or FLAIR images for analysis. However, not all of them contribute equally to the diagnostic accuracy. Feature selection is an essential step that selects informative biomarkers while eliminating irrelevant biomarkers in order to train effective classifiers. In addition, feature selection may provide interpretable results for understanding the underlying pathologies behind different diseases. Most feature selection methods such as t-tests or Elastic Net sparse regression (Moradi et al., 2015) can select discriminative features that show significant differences between two groups of patients. They are therefore tuned towards a binary classification scenario. However, these approaches do not guarantee that the selected features in binary classification would be useful in multi-class classification. A multi-class feature selection method is required in our differential diagnostic task to identify the useful biomarkers.

In this work, our objective is to develop an effective classification framework for accurate differential diagnostics of different NDs. Previous studies (Varma et al., 2002; Grossman et al., 2004; Davatzikos et al., 2008; Burton et al., 2009; Muñoz-Ruiz et al., 2012; Raamana et al., 2014) have been carried out for the differential diagnostics of dementias in different classification scenarios. However, most previous studies were based on binary classifications. The multi-class classification may be more useful to clinicians than the binary classification as it represents a real-world clinical scenario. In a recent study (Koikkalainen et al., 2016), we have presented a multi-class classification framework on the differential diagnostics of dementias. In comparison with our previous study in Koikkalainen et al. (2016), the major contributions of this work include (1) the extraction of more sophisticated region-wise grading features than those in Koikkalainen et al. (2016). Specifically, in this study, two types of grading features were extracted for analysis. The atrophy grading features extracted from T1-weighted images can capture the atrophy information for classification while the VaD grading features utilize vascular changes from the FLAIR images; (2) the introduction of the RUSBoost algorithm to handle the class imbalance problem; and (3) the integration of a multi-class feature selection step to select useful biomarkers for classification and to show the importance of different biomarkers and regions. Cross-sectional studies are performed on a large dataset of 500 subjects including 118 patients with subjective memory complaints (SMC), 219 AD patients, 92 FTLD patients, 47 DLB patients and 24 VaD patients. In the remainder of the paper, we will first introduce the dataset used in our work in Section 2.1. The calculation of different biomarkers is introduced in Section 2.3 and the classification framework using feature selection and RUSBoost is then presented in Sections 2.4 and 2.5 respectively. The performance of the proposed framework is analyzed in Section 3. Finally, we discuss the strengths and weaknesses of our work in Section 4 and conclude this paper in Section 5.

2. Material and methods

2.1. Dataset

Data used in the preparation of this article were obtained from the Amsterdam Dementia Cohort (van der Flier et al., 2014). A group of 500 patients who visited the Alzheimer Center between 2004 and 2014 were studied in this work. All patients received a standardized and multi-disciplinary work-up, including medical history, physical, neurological and neuropsychological examination, MRI, laboratory test and lumbar puncture to collect cerebrospinal fluid. A MR T1-weighted gradient echo sequence was performed for each patient to capture the atrophy patterns of different NDs. 97 and 317 patients were imaged using 1.5 T and 3.0 T devices respectively. The remaining 86 patients were imaged using 1.0 T device. The voxel sizes of T1-weighted images varied between $0.9 \times 0.9 \times 0.9$ mm^3 and $1.1 \times 1.1 \times 1.5$ mm^3. In addition, a fast FLAIR sequence was also carried out for each patient. The voxel sizes of the obtained FLAIR images varied between $0.4 \times 0.4 \times 1.0$ mm^3 and $1.2 \times 1.2 \times 6.5$ mm^3.

The diagnosis of patients were made in a multidisciplinary consensus meeting according to several criteria: probable AD was assigned according to the criteria of National Institute on Aging-Alzheimer's Association (McKhann et al., 2011) and the criteria of National Institute for Neurological and Communicative Diseases Alzheimer's Disease and Related Disorders Association (McKhann et al., 1984); FTLD was assigned based on the Neary criteria (Neary et al., 1998) and the revised criteria from Rascovsky et al. (2011); DLB was diagnosed using the McKeith criteria (McKeith et al., 2005); Patients were diagnosed with VaD using the criteria of National Institute of Neurological Disorders and Stroke and Association Internationale pour la Recherche et l'Enseignement en Neurosciences (Román et al., 1993); patients with SMC were diagnosed only when the cognitive complaints and tests could not meet the criteria for mild cognitive impairment (MCI), dementia or other neurological or psychiatric disorder. In addition, follow up of patients with SMC took place by annual routine visits with a mean of 2.5 ± 1.4 years and a minimal of 9 months. Only SMC patients who were confirmed to remain stable during the follow-up were included in this study. Cognitive functions were assessed with a standardized test battery consisting of the Mini Mental State Examination (MMSE), the Cambridge Examination for Mental Disorders of the Elderly (CAMCOG)

Table 1
Demographic information of the dataset used in this study. MMSE: Mini-Mental State Examination; AD: Alzheimer's disease; FTLD: frontotemporal lobe degeneration; DLB: dementia with Lewy bodies; VaD: vascular dementia; SMC: subjective memory complaints as a control group.

Group	Number	Age	MMSE	Diagnosis criteria
AD	219	65.9 ± 7.3	20.6 ± 4.5	McKhann et al. (1984) and McKhann et al. (2011)
FTLD	92	63.2 ± 6.7	23.7 ± 5.2	Neary et al. (1998) and Rascovsky et al. (2011)
DLB	47	68.5 ± 5.7	23.8 ± 4.7	McKeith et al. (2005)
VaD	24	68.1 ± 8.6	23.1 ± 3.7	Román et al. (1993)
SMC	118	60.4 ± 8.5	28.4 ± 1.3	Did not meet the criteria for MCI, dementia or other neurological or psychiatric disorder

forward and backward conditions of Digit Span, the Visual Association Test (VAT), the Rey Auditory Verbal Learning Test (RAVLT), the Category Fluency Test (CFT) (animals), the Trail Making Test (TMT), the Frontal Assessment Battery (FAB), the Stroop test and the Rey figure copy test. Depressive symptoms were assessed by the Geriatric Depression Scale (GDS), behavioural and psychological symptoms by the Neuropsychiatric Inventory (NPI) and activities of daily living using the Disability Assessment for Dementia (DAD). These neuropsychological tests were used in the above criteria for diagnosis. The study was approved by the local Medical Ethical Committee. All patients have signed written informed consent for their clinical data to be used for research purposes. The demographic information of the dataset is shown in Table 1.

2.2. Image preprocessing

All T1-weighted images were bias corrected using the N4 bias field correction algorithm (Tustison et al., 2010) and skull-stripped using the pincram method (Heckemann et al., 2015). After that, non-rigid registration based on B-spline free-form deformation (Rueckert et al., 1999) with a final control point spacing of 10 mm was performed to align all T1-weighted images to the MNI152 template space. The approach proposed in Nyúl and Udupa (1999) was used to normalize the image intensities between the subjects and the template. Each FLAIR image was aligned to the T1-weighted image of the same subject using affine registration, followed by a non-rigid transformation to the template space. The intensity normalization approach (Nyúl and Udupa, 1999) was also applied to all the FLAIR images. After preprocessing, all the images were in the same template space and the intensities of images from the same modality were at a comparable scale. In addition, it was noted that the images acquired with the 1.0 T MRI device produced systematic differences as compared to images acquired with the 1.5 T and 3 T MRI. Consequently, a regression step was added to remove this systematic error from the intensities, making it possible to simultaneously compare images acquired with different MRI devices.

2.3. Extraction of features

For each patient, a total of 824 features was extracted as biomarkers for analysis as shown in Table 2, including 138 volume features, 682 grading features, 3 CSF measures and age. How these features were extracted is described in the following:

Volume features: T1-weighted images were segmented into 138 anatomical regions using multi-atlas label propagation with expectation-maximization (MALPEM) (Ledig et al., 2015). 30 T1-weighted images which were manually segmented by Neuromorphometrics, Inc. (http://Neuromorphometrics.com) were used as atlases. First, all 30 atlases were transformed to a target image space using non-rigid registration (Rueckert et al., 1999). Then, atlas label maps were transformed using the obtained transformations. Finally, a label fusion step with an expectation-maximization (EM) refinement (Ledig et al., 2015) was conducted to obtain a consensus segmentation. The output of the segmentation includes 138 brain structures whose volumes were used as features. These volume features can capture the atrophy patterns of different NDs, providing discriminative information for classification. The cerebral white matter region was further segmented into 130 subregions using an extension of the method (Ledig et al., 2015), where a white matter atlas is affinely aligned to provide a finer white matter parcellation. Here, we employed the white matter atlas presented in Oishi et al. (2009). The white matter subregions were used for calculating the VaD grading features as described in the following paragraph. Fig. 1 shows an example of the segmentation results.

Table 2
A summary of the features used in our study.

Features	Modality	Dimension
Volume	T1-weighted	138
SMC grading	T1-weighted	138
AD grading	T1-weighted	138
DLB grading	T1-weighted	138
FTLD grading	T1-weighted	138
VaD grading	FLAIR	130
CSF	CSF	3
Age	–	1

Grading features: Although volume-based imaging biomarkers provide good characterization of brain atrophy patterns, their capability is limited by the inter-subject variability of brain anatomy (Coupé et al., 2012). The grading features (Coupé et al., 2012), which calculate scoring values for each test subject by estimating its similarity to different training populations, have been shown to allow for a better characterization of structural atrophy for the detection of AD (Coupé et al., 2012) than the volume features. In this work, we propose to calculate region-wise grading features using sparse regression techniques. Specifically, a grading value was calculated within each brain structure. Given the intensities of a test subject $I_{test} \in R^{k\times 1}$ and the intensities of n training subjects $\mathbf{I}_{training} \in R^{k\times n}$ within a structure i, the structure-specific grading score g_i of this test subject can be calculated by minimizing the following cost function (Tong et al., 2017):

$$\begin{cases} \hat{\alpha}_i = \underset{\alpha_i}{\operatorname{argmin}} \frac{1}{2}\|I_{test} - \mathbf{I}_{training}\alpha_i\|_2^2 + \lambda_1\|\alpha_i\|_1 + \frac{\lambda_2}{2}\|\alpha_i\|_2^2 \\ g_i = \frac{\sum_{j=1}^{n}\hat{\alpha}_i(j)y_j}{\sum_{j=1}^{n}\hat{\alpha}_i(j)} \end{cases} \quad (1)$$

Here $\hat{\alpha}_i$ are the coding coefficients of the test subject for structure i and y_j is the disease label vector for the jth training subject. Each training label vector is defined as $y_j = [0,\ldots 1\ldots,0]$, where the non-zero entry position indicates the disease label of a specific group. Most of the coefficients in $\hat{\alpha}$ are zero due to the L_1 regularization over α. By adding the L_2 norm in Eq. (1), a grouping effect can be obtained over the sparse coding coefficients. Qualitatively speaking, an algorithm exhibits the grouping effect if the coding coefficients of a group of highly correlated subjects tend to be equal. For example, there are two training subjects with very similar or identical intensity patterns. If we just use the L_1 norm, it will select one of them while eliminating the other one. However, in calculating the grading biomarkers, both subjects are similar to the target subject within structure i and should be used to propagate their disease information to the target subject. After adding the L_2 norm as in Eq. (1), both subjects can be selected in calculating the grading biomarkers. If the jth coefficient in $\hat{\alpha}_i$ is not zero, it indicates that the corresponding jth training subject has been selected to propagate its clinical label information to the test subject.In our work, two types of grading

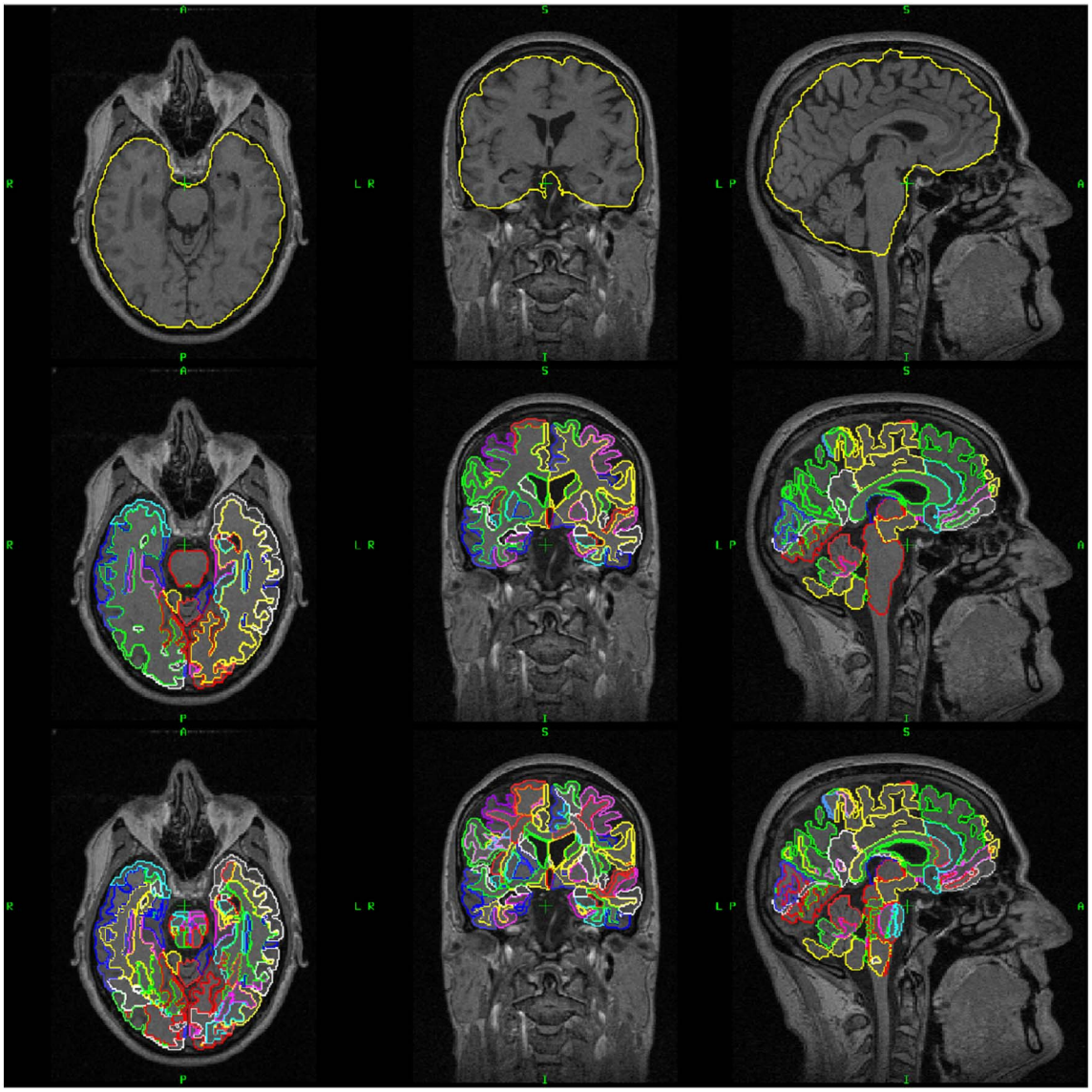

Fig. 1. An example of the calculated segmentations of an T1-weighted MR image. The top row presents the original image and the yellow contour illustrates the result of skull stripping. The second row shows the segmentation of 138 regions and the bottom row adds the further segmentation of the white matter region into 130 subregions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

features were extracted separately: atrophy grading and VaD grading. The atrophy grading features were calculated using T1-weighted images only. They include SMC grading, AD grading, FTLD grading and DLB grading. These grading features were calculated within each of the 138 structures in the brain region, generating 552(4×138) atrophy grading features in total. In order to characterize the white matter changes that are typical to VaD, VaD grading features were calculated separately using FLAIR images. This was carried out in order to differentiate VaD patients from other patients. A VaD grading value was computed within each of the 130 white matter subregions. This provides us with additional 130 grading features. Fig. 2 shows the mean grading maps of patients in different groups.

CSF features: In addition to the features extracted from MR images, we also utilized the CSF biomarkers including $A\beta 42$, total-tau (t-tau), and phosphorylated-tau (p-tau) in our study. These biomarkers have been shown to be highly related to Alzheimer's neuropathology, and can provide useful information in differentiating AD from other types of dementia (Schoonenboom et al., 2012).

2.4. Multi-class feature selection

For each patient, 824 features were extracted as described in the previous section. Some of these features may not be relevant to some of the pathological changes occurring in NDs and therefore do not provide useful information for the multi-class classification task. In order to train more effective classifiers, these features should be eliminated. However, it does not necessarily mean that a feature which captures the pathological changes of NDs is always useful for multi-class classification. In an extreme case, for example, if a feature is equally affected by different NDs (i.e. same atrophy pattern in a specific region), this feature will not provide discriminative information for differential diagnostics even though it reflects the pathological changes of NDs. Therefore, it is essential to apply a multi-class feature selection approach to select those discriminative features which show differences among all classes.

Most existing feature selection approaches (Saeys et al., 2007) such as t-test or sparse regression techniques (Moradi et al., 2015) aim to select discriminative features between two groups of patients. However, the selected features in binary classification are not guaranteed to be useful in multi-class classification. Argyriou et al. (2008) proposed a

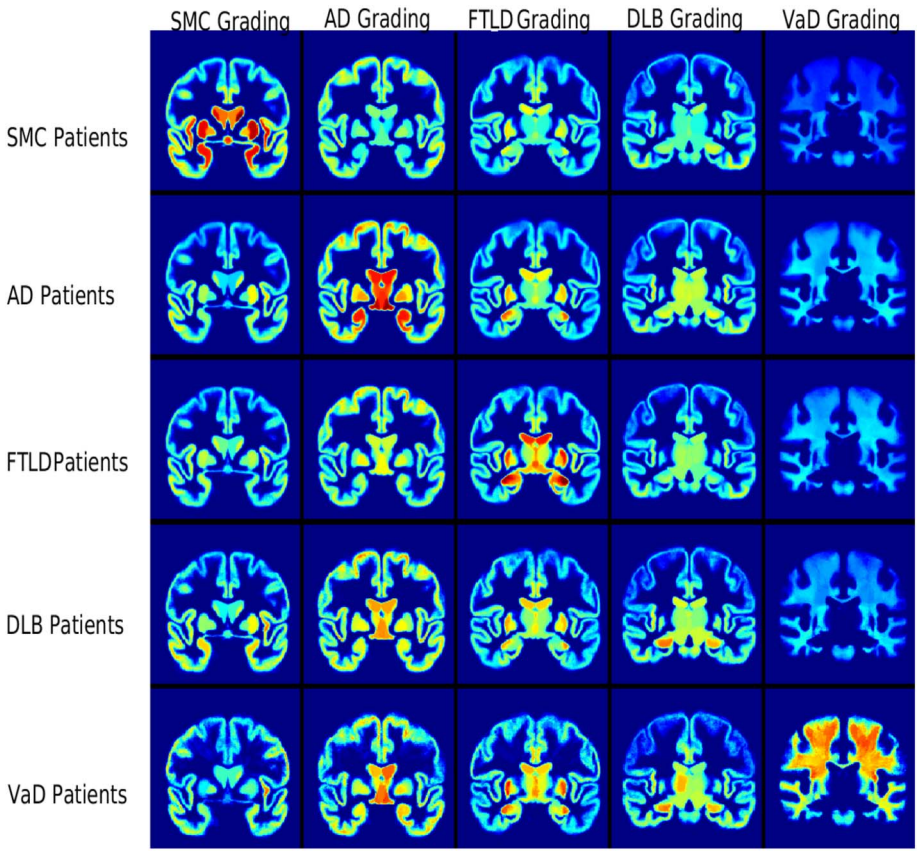

Fig. 2. The average grading maps for each patient group. For each subject, 682 grading features were extracted, including 138 SMC grading features (the first column), 138 AD grading features (the second column), 138 FTLD grading features (the third column), 138 DLB grading features (the fourth column) and 130 VaD grading features (the fifth column). The VaD grading features were extracted in the white matter subregions using FLAIR images. The other grading features were extracted in the grey matter subregions using T1-weighted images.

multi-class feature selection method by assuming that a small subset of features are shared by different classes and formulated the problem as the $l_{2,1}$-norm regularized non-smooth optimization problem:

$$\hat{W} = \arg\min_{W} \left\| X^T W - Y \right\|_F^2 + \beta \left\| W \right\|_{2,1} \tag{2}$$

Here $\|.\|_F$ denotes the Frobenius norm of a matrix. $X \in R^{d \times N}$ are the features with a dimensionality of d and $Y \in R^{N \times c}$ are the corresponding label vectors with c classes. W is the selection matrix. The $l_{2,1}$-norm regularization on W penalizes each row of W as a whole and enforces sparsity among the rows, which allows the selection of informative features shared by multiple groups for classification. Eq. (2) can be efficiently solved using the first-order black-box method (Liu et al., 2009). The $l_{2,1}$-norm regularization is based on a strict assumption that all groups share a common set of discriminative features. However, in many cases, the common set is shared by many groups, but not all. In order to alleviate this strong assumption, we add a l_1-norm regularizer as proposed in Wang et al. (2011) to impose the sparsity among all elements in W:

$$\hat{W} = \arg\min_{W} \left\| X^T W - Y \right\|_F^2 + \beta_1 \left\| W \right\|_1 + \beta_2 \left\| W \right\|_{2,1} \tag{3}$$

After the above equation is solved using an iterative algorithm (Wang et al., 2011), the selected features are determined according to non-zero coefficients in $\hat{W}$.

2.5. Multi-class classification using RUSBoost

As shown in Table 1, our dataset is imbalanced (the numbers of VaD and DLB subjects are much smaller than the other three groups), which poses challenges for traditional classifiers such as support vector machine (SVM) to learn effective classification models. The RUSBoost algorithm (Seiffert et al., 2010) combines the random under-sampling (RUS) technique with the AdaBoost algorithm (Rätsch et al., 2001) to tackle the imbalanced training problem and to learn strong classifiers. In this method, an ensemble of classifiers are trained using a randomly under-sampled subset of the available data. In each iteration of RUSboost, the weight of each sample is adjusted. The weights of misclassified samples are increased while the weights of correctly classified samples are decreased. Therefore, the misclassified samples are more likely to be correctly classified in subsequent iterations. The final classification is a weighted combination of the results of all classifiers in the ensemble. Since minority class samples are most likely to be misclassified at the first iteration, they will receive higher weights during the subsequent iterations and be correctly classified in the boosting process.

Assume we are given a set of n training samples $\mathbf{S}$, including $\{\mathbf{X},\mathbf{Y}\} = \{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$. $\mathbf{X}$ are the input features and $\mathbf{Y}$ are corresponding labels. Each training sample in $\mathbf{S}$ is assigned an initial weight $D_1(i) = \frac{1}{n}$, $i = 1, 2, ..., n$. For each iteration $t = 1, 2, ..., T$, the RUSBoost algorithm performs the following steps:

1. Create a subset of training samples $\mathbf{S}_t$ with distribution D_t using random under-sampling
2. Learn a weak classifier $h_t : \mathbf{S}_t \rightarrow Y_t$ using decision trees, given an input of $\mathbf{S}_t$
3. Do predictions on all training samples S using h_t
4. Calculate the pseudo-loss for S:
   $\epsilon_t = \sum_{(i,y):y_i \neq y} D_t(i)(1 - h_t(x_i, y_i) + h_t(x_i, y))$
5. Calculate the weight for the weak classifier h_t:
   $\alpha_t = \frac{\epsilon_t}{1-\epsilon_t}$
6. Update the weight $D_t(i)$ for each sample:
   $D_{t+1}(i) = D_t(i)\alpha_t^{\frac{1}{2}(1+h_t(x_i,y_i)-h_t(x_i,y:y\neq y_i))}$
7. Normalize the weights D_{t+1}:
   $D_{t+1}(i) = \frac{D_{t+1}(i)}{\sum_{i=1}^{n} D_{t+1}(i)}$

After T iterations of the above steps, T weak classifiers h_t with weights α_t are obtained. The final output can be calculated as

$$H(x) = \underset{y \in Y}{\operatorname{argmax}} \sum_{t=1}^{T} h_t(x, y) \log \frac{1}{\alpha_t} \tag{4}$$

2.6. Implementation details

For training classifiers, SVM, random forest (RF), multi-class cost learning kNN (mc*k*NN), the synthetic minority over-sampling technique (SMOTE) and RUSBoost were used. The implementation of SVM was performed using libsvm in Matlab (https://www.csie.ntu.edu.tw/~cjlin/libsvm/). The random forest classifier was implemented via (http://code.google.com/p/randomforest-matlab). The number of trees in random forest was set to 500 and the number of features randomly selected at each tree node was set to the square root of the total number of features as suggested in Liaw and Wiener (2002). mc*k*NN is a cost learning method proposed in Zhang and Zhou (2010) to handle the class imbalance problem. The implementation of mc*k*NN from (http://lamda.nju.edu.cn/code_mcKLR.ashx) was utilized. SMOTE proposed in Chawla et al. (2002) is an over-sampling technique to handle the imbalance problem, and the implementation from (http://lamda.nju.edu.cn/code_CSNN.ashx) was used. To perform RUSBoost, the number of iterations T was set to 500 in all experiments. All features were normalized to have zero mean and unit variance before the classification. To evaluate the classification performance, 10-fold cross validation was carried out. The reported results in terms of overall accuracy and balanced accuracy are averages over 100 runs, which are calculated as

$$\text{Overall Accuracy} = \frac{\text{number of all correctly classified subjects}}{\text{total number of all subjects}} \tag{5}$$

$$\text{Balanced Accuracy} = \frac{1}{c} \sum_{i=1}^{c} \frac{\text{number of correctly classified subjects in group}i}{\text{number of subjects in group}i} \tag{6}$$

To assess the statistical significance of different results, the Mann Whitney U Test were performed using the accuracies of 100 runs. In addition, the area under this curve (AUC) was used as another performance measure. In contrary to accuracy, AUC measurement does not require a threshold on the classifier's output probabilities and thus does not rely on the class priors. The multi-class AUC (MAUC) was calculated based on the method proposed in Hand and Till (2001).

3. Results

3.1. The advantage of RUSBoost

Using all the 824 features, we validated the classification performance of different classifiers: SVM, RF, mc*k*NN and RUSBoost. Fig. 3 compares the classification results using different classifiers. In addition, the confusion matrices of the classification results are shown in Fig. 4 for comparison. As can be seen from Fig. 3, the overall accuracies of RF and RUSBoost are significantly better than that using SVM. Although RF achieves the best overall accuracy, the balanced accuracy of RF is the worst. The imbalanced training set causes RF to be biased towards the majority groups. This bias can be confirmed by the confusion matrix of RF in Fig. 4. RF achieved very high sensitivities in the classification of AD and SMC, while the sensitivities in detecting VaD and DLB are nearly zero. Interestingly, SVM is biased to minority groups as shown in Fig. 4. The highest sensitivity using SVM was observed for the VaD group. In SVM, Support Vectors (SVs) are selected from each group. Only SVs are used for building classification models while many samples of majority groups far from the decision boundary are not used. The percentage of SVs in minority groups (i.e. VaD group) is much higher than that in majority groups (i.e. AD group), which may result in the bias of SVM towards minority groups. In comparison with SVM and RF, the results using mc*k*NN and RUSBoost are less biased since they were proposed to handle the class imbalance problem. The sensitivities of each group using mc*k*NN and RUSBoost are sensible as shown in their confusion matrices. Overall, RUSBoost achieved a good overall accuracy with much higher balanced accuracy than the other three approaches.

3.2. Comparison of results using different features

In a further experiment, we investigated and compared the classification performance when different types of features were used. RUSBoost was used for classifier training and testing. Fig. 5 shows the classification results using different feature types. The classification results using grading features are more accurate than those using volume features. Both the grading features and volume features were extracted from MR images. However, the volume features were calculated based on the T1-weighted images only, which capture the atrophy pattern of different NDs for classification. The grading features were extracted using information from both the T1-weighted and FLAIR images, which not only capture the atrophy pattern from T1-weighted images but also the white matter changes from FLAIR images. This provides additional information for classification, resulting in a more accurate classification performance as compared to volume features. Moreover, the combination of all features resulted in the best classification performance, indicating complementary information between the MR features and the CSF biomarkers.

3.3. Benefit of feature selection

In the above classification experiments, all the 824 features were used for classification. However, some features may not be informative for the five-class classification. Thus, we applied the multi-class feature selection step as described in Section 2.4. The results after feature selection are compared to those without feature selection. As shown in Table 3, the performance of RUSBoost can be significantly improved after feature selection (the p-value is less than 0.001). The feature selection can improve both the overall and balanced accuracies of all classifiers except for the balanced accuracy of SVM. The overall accuracy of SVM was increased after feature selection since SVM is biased to minority groups as demonstrated in Section 3.1, while this may hamper the improvement of the balanced accuracy. Overall, RUSBoost achieved significantly higher balanced accuracy than the other classifiers. The average confusion matrix of 100 cross validations using RUSBoost after feature selection is presented in Fig. 6. As shown in Fig. 6, the VaD patients were mostly misclassified as AD patients. The most challenging group are the DLB patients, which were frequently misclassified as SMC and AD patients. In addition, we carried out the classification by just using imaging biomarkers including the volume and grading features. A classification accuracy of 70.0% with a balance accuracy of 67.2% was achieved.

3.4. Most discriminative features and regions

In addition to obtaining accurate differential diagnostics, it is also interesting to investigate which features and structures are most important and informative for the diagnostics. In order to estimate the importance of different features, feature selection was repeatedly carried out on subsets of the data in the 100 runs of ten-fold cross validation. The selection frequency, which was normalized into [0,1], provides a measure of importance for each feature and was used in our work. If the selection frequency of a feature is 1, it means that this feature is consistently selected across all runs. Features which are consistently selected can be considered as more important than those which are only selected occasionally. We ranked the importance of different features according to their selection frequencies in the feature selection step. The ranked list of the feature importance can be found at

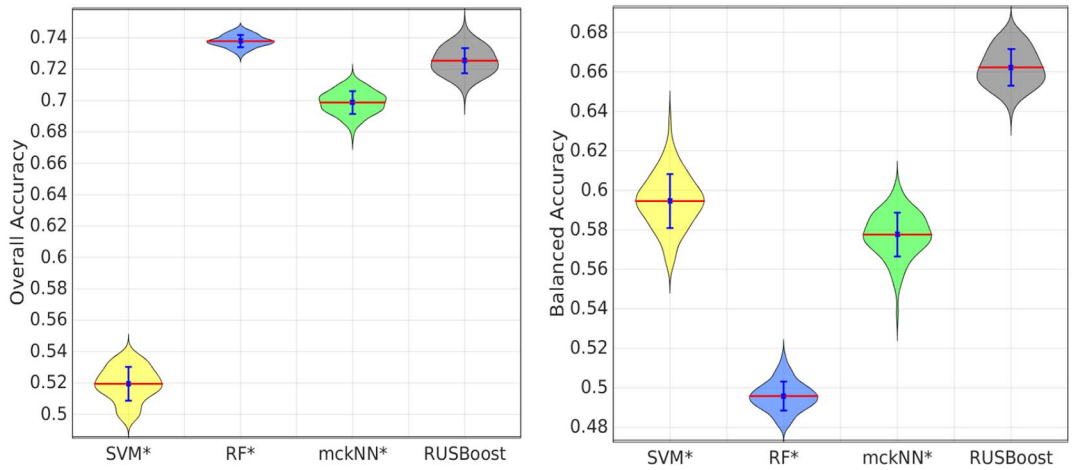

Fig. 3. Comparison of classification performance using different classifiers. Results were obtained using 100 runs of 10-fold cross validation. The statistical tests using the Mann Whitney U Test were performed between the results using RUSBoost and those using other methods. * means that the results are significantly different from those using RUSBoost with p-value < 0.001.

http://scholar.harvard.edu/files/ttong/files/featureimportance.txt. The top 30 features consist of 3 volume features, 23 grading features, age and 3 CSF biomarkers. In addition, we mapped the importance of the grading features and the volume features into the MNI152 template space for visualization as shown in Fig. 7.

In order to estimate the importance of each structure, the selection frequencies of different features within each structure were added up, including the selection frequencies of the volume features and the five grading features. The summed frequency was treated as the importance of each structure. We ranked the summed frequencies of all structures.

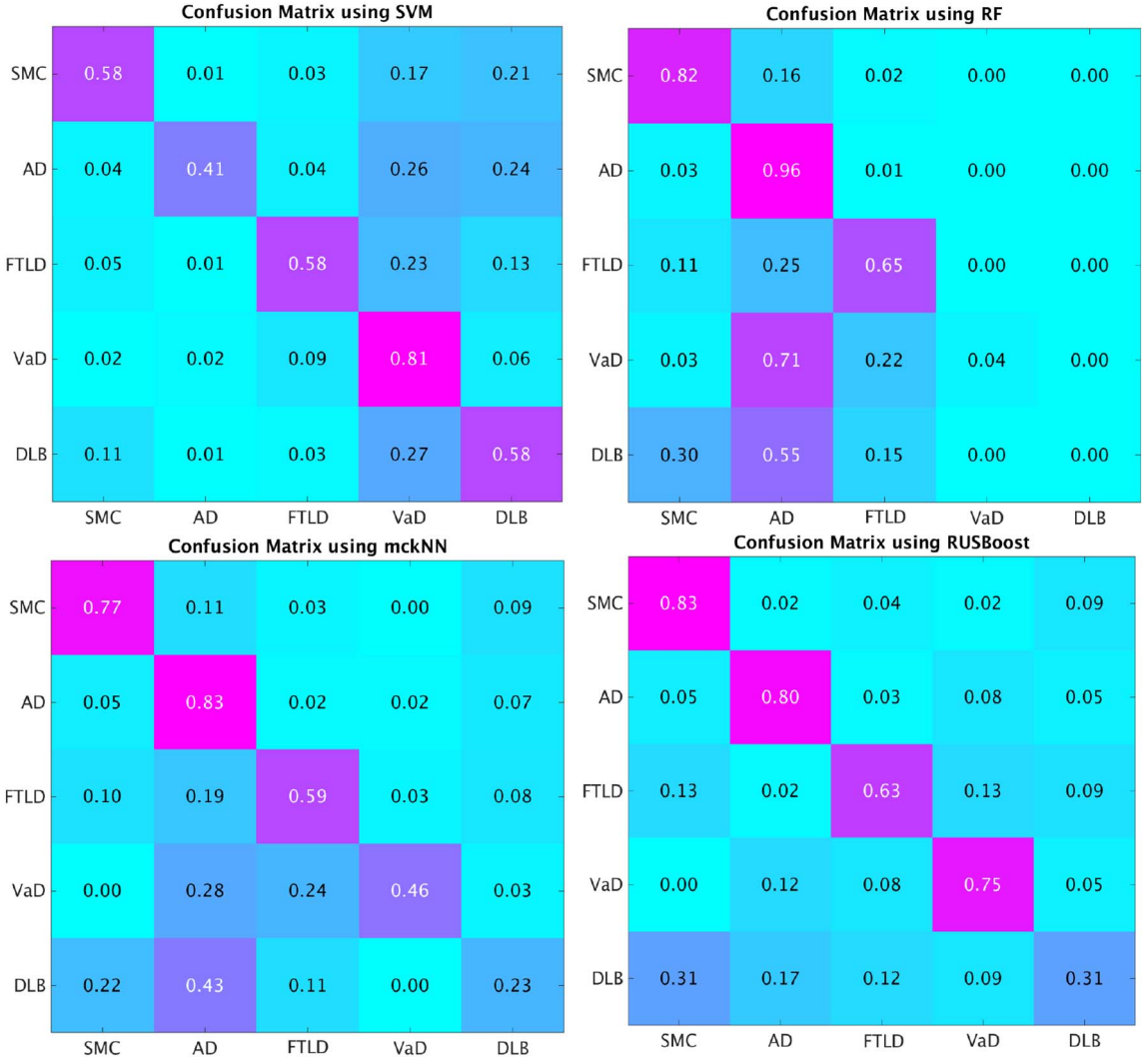

Fig. 4. Confusion matrix of the classification results using different classifiers.

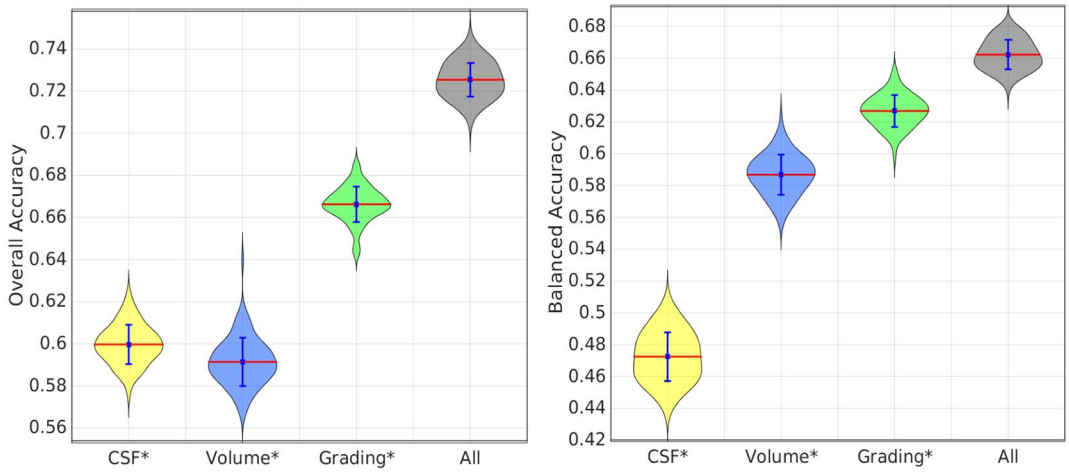

Fig. 5. Comparison of classification performance using different types of features. RUSBoost was used for training and testing. Results were obtained using 100 runs of 10-fold cross validation. The statistical tests using the Mann Whitney U Test were performed between the results using all features and those using each individual type of features. * means that the results are significantly different from those using all features with p-value < 0.001.

Table 3
The classification results before and after feature selection. FS represents the method after feature selection. MAUC represents multi-class area under the curve. The statistical tests using the Mann Whitney U Test were performed between the overall accuracies using methods without feature selection and those with feature selection.

Classifier	Overall accuracy (%)	Balanced accuracy (%)	MAUC (%)
SVM	51.9 ± 1.1	59.5 ± 1.4	81.8 ± 0.7
RF	73.8 ± 0.4	49.6 ± 0.7	86.2 ± 0.5
SMOTE	71.0 ± 0.6	60.1 ± 1.0	85.0 ± 1.1
mc*k*NN	69.8 ± 0.7	57.7 ± 1.1	–
RUSBoost	72.5 ± 0.8	66.2 ± 0.9	88.1 ± 0.3
SVM_FS*	65.4 ± 1.4	55.6 ± 1.9	84.7 ± 0.9
RF_FS*	75.2 ± 0.6	52.2 ± 0.9	87.9 ± 0.5
SMOTE_FS*	73.1 ± 1.1	63.8 ± 1.6	86.7 ± 1.0
mc*k*NN_FS*	73.4 ± 1.1	60.0 ± 1.7	–
RUSBoost_FS*	75.2 ± 0.8	69.3 ± 1.0	89.3 ± 0.5

* Means that the results with feature selection are significantly different from those without feature selection with p-value < 0.001. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The ranked list is available at http://scholar.harvard.edu/files/ttong/files/regionimportance.txt. The most selected structures include the hippocampus, the entorhinal cortex, amygdala, the middle occipital gyrus and the accumbens area, implying that these structures are important indicators for the differential diagnostics of NDs. In addition, the highly selected structures have been shown to be more discriminative in the left brain than those in the right brain. Take the hippocampus as an example, the right hippocampus volumes of the FTLD patients have similar average size as those of AD and VaD as shown in Fig. 8. However, the left hippocampus of the FTLD patients has smaller average size than those of all the other groups. This indicates that the FTLD patients have similar atrophy in the right hippocampus as the AD and VaD patients, but have larger atrophy in the left hippocampus than other groups. Therefore, the left hippocampus can provide useful information for distinguishing the FTLD patients from the other groups. This is why both the grading feature and the volume feature of the left hippocampus were consistently selected for classification as shown in the ranked list of the feature importance. Last but not least, VaD grading features show highly differential ability of VaD from other dementias in the structures of the subcortical precentral gyrus (PrCG) and the inferior fronto-occipital (IFO) fasciculus of white matter as shown in Fig. 9. This is consistent with previous findings which show white matter hyperintensity in the PrCG (Ngai et al., 2007) and IFO structures (Sheline et al., 2008), indicating that these structures are important locations for differentiating VaD from other dementias.

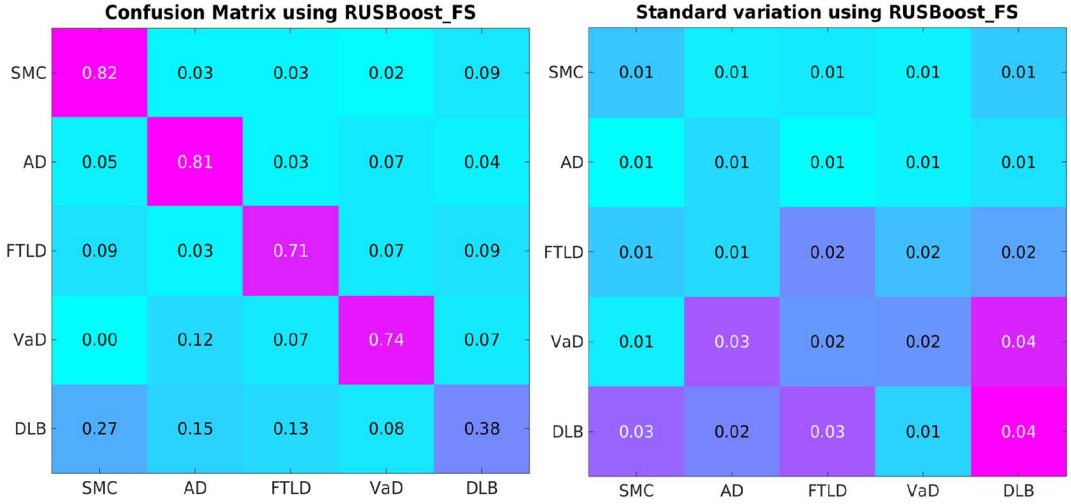

Fig. 6. Confusion matrix of RUSBoost after feature selection using all available features. The values in the right figure are the corresponding standard variation of the values in the confusion matrix.

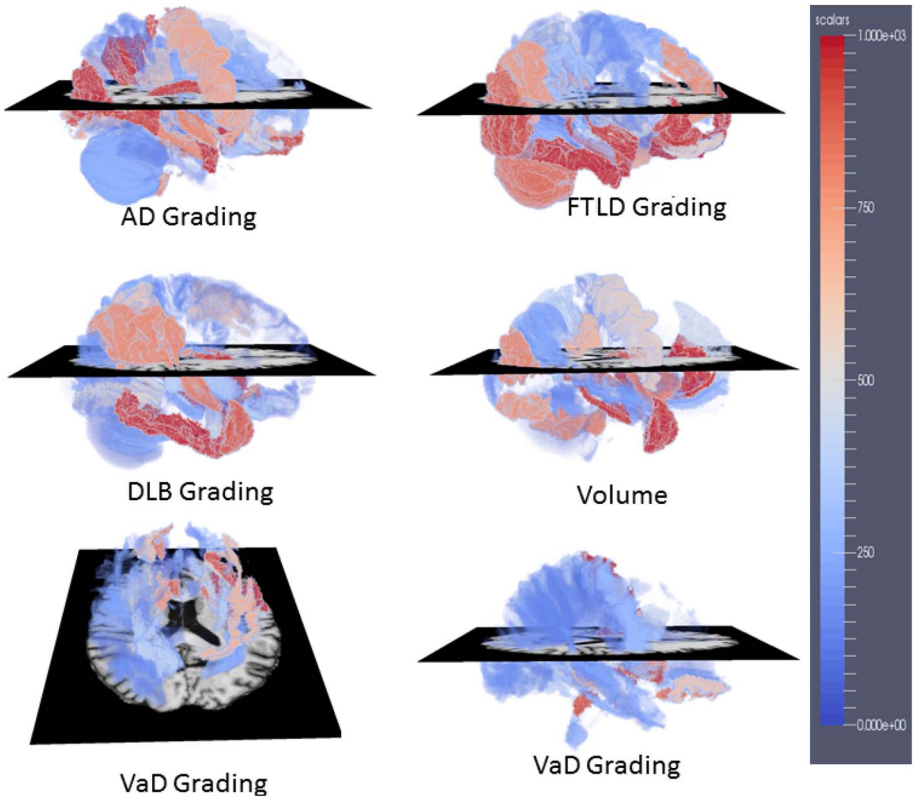

Fig. 7. The importance of different features in the MNI152 template space. A high value (in red regions) in the maps means that the corresponding feature extracted in that region was selected with a high frequency, indicating that the feature is important for the five-class classification.

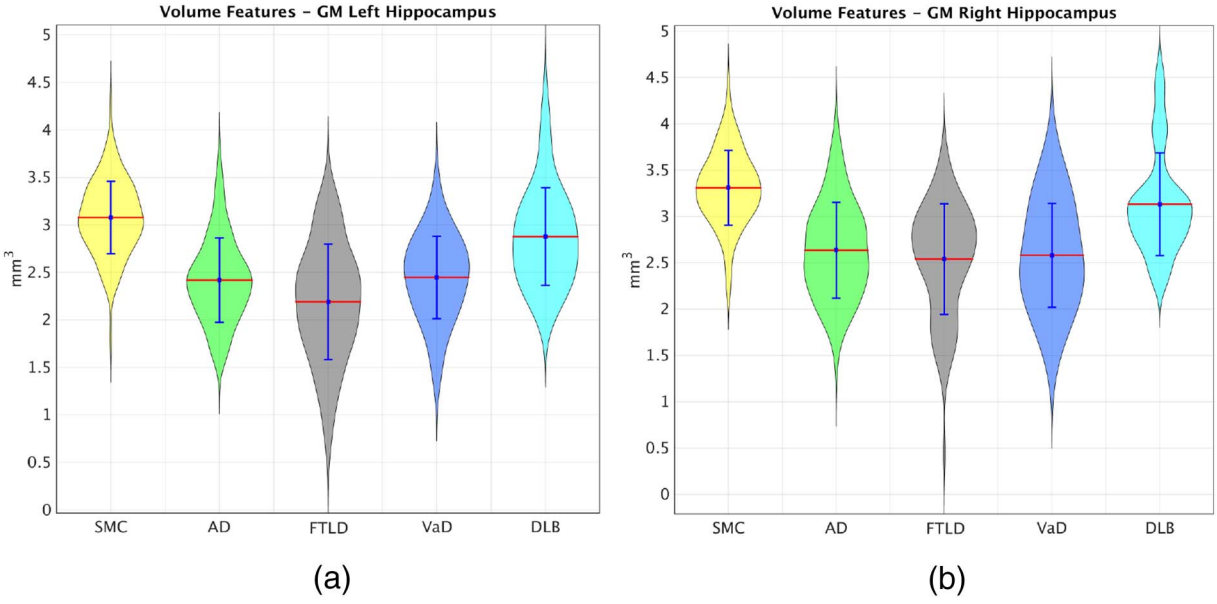

Fig. 8. The distributions of the left and right hippocampus volumes for different groups.

4. Discussion

In this study, we have developed a novel classification framework for the differential diagnostics of dementias. First, volume features were extracted from 138 structures using MR T1-weighted images. In addition, two types of region-wise grading features were extracted from the T1-weighted and FLAIR images respectively to capture both the atrophy information and the vascular changes for classification. Finally, three CSF features and age were added, resulting in a total of 824 features for each subject. The RUSBoost classifier with a multi-class feature selection method was then applied to the differential diagnostics. In the end, an overall classification of 75% with a balanced accuracy of 69% was obtained using a dataset of 500 subjects. SMC and AD patients can be well differentiated with sensitivities over 80%. The sensitivities of differentiating FTLD and VaD are over 70%, while the most challenging group are the DLB patients with a sensitivity of 38%. A large number of DLB patients were misclassified as SMC or AD, indicating that there is a large variation in the DLB group. Some DLB patients are more healthy-like while some are more atrophy-like.

Previous studies (Varma et al., 2002; Grossman et al., 2004; Davatzikos et al., 2008; Burton et al., 2009; Muñoz-Ruiz et al., 2012; Raamana et al., 2014) have reported results for the differential diagnostics of dementias in different classification scenarios. Since the utilized data are different and the classification scenarios vary in these

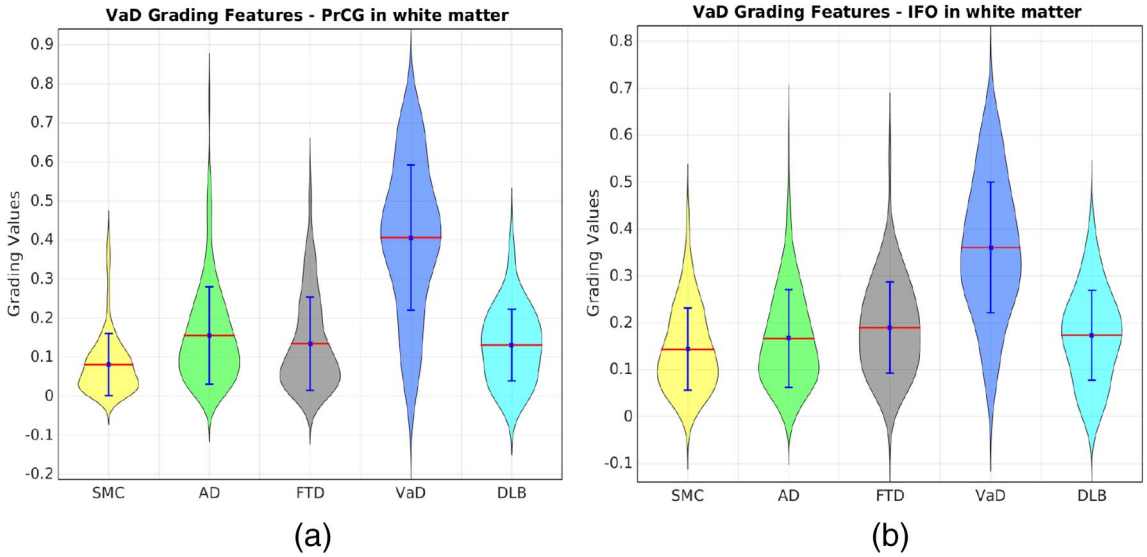

Fig. 9. The distributions of VaD grading features in PrCG and IFO of white matter for different groups.

Table 4
Comparison of the results using different classifiers and features.

Features	Method	Classifier	Overall accuracy	Balanced accuracy
Volume features	Koikkalainen et al. (2016)	DSI	50.4	50.7
	Proposed	RUSBoost	58.6	58.7
Grading features	Koikkalainen et al. (2016)	DSI	58.3	51.5
	Proposed	RUSBoost	66.6	62.7

studies, it is difficult to make a direct comparison with these studies. In comparison with the study in Koikkalainen et al. (2016), more sophisticated region-wise grading features were extracted in this study. These grading features can capture both the atrophy information from T1-weighted images and the vascular changes from the FLAIR images. As shown in Table 4, an overall classification accuracy of 58.3% was obtained in Koikkalainen et al. (2016) with a balanced accuracy of 51.5% when the grading features were used. In this work, the classification accuracy was improved to 66.6% and the balanced accuracy to 62.7% using the proposed grading features. The improvement resulted from the calculation of grading features in an anatomical region-wise way and the introduction of the vascular grading features. In addition, the RUSBoost algorithm was introduced in this work to handle the class imbalance problem. The balanced accuracy was 50.7% in Koikkalainen et al. (2016) when only the volume features were used. This was increased to 58.7% in this work, demonstrating the effectiveness of the RUSBoost method in handling the class imbalance problem. Moreover, a multi-class feature selection method was adopted in this work to select useful features for the five-class classification. The multi-class feature selection step not only improves the classification accuracy but also allows us to investigate which features and structures are most important in the classification. We have shown that the grey matter structures including the hippocampus, the entorhinal cortex, amygdala, the middle occipital gyrus and the accumbens area are important regions for capturing atrophy differences for differential diagnostics. These structures have been reported as important regions for identifying dementias in previous studies (Barber et al., 2000; Schott and Fox, 2007; Frisoni et al., 2010). For example, the early structural changes in AD have been found to occur in medial temporal lobe, particularly atrophy of hippocampus and entorhinal cortex (Schott and Fox, 2007; Frisoni et al., 2010). The atrophy in hippocampus and medial temporal lobe are relatively preserved in FTLD as compared to AD (Duara et al., 1999). DLB patients have shown significant atrophy as compared to controls, but relatively preserved in medial temporal lobe, thalamus, hippocampal, and amygdala volumes as compared to AD (Barber et al., 2000, 1999). In addition, the PrCG and IFO structures in white matter have been shown in our study to provide essential information for differentiating VaD, which is consistent with previous findings (Ngai et al., 2007; Sheline et al., 2008). Moreover, Varma et al. (2002) reported asymmetrical atrophy only in FTLD patients. In our work, we found asymmetrical hippocampal volumes in the FTLD patients as shown in Fig. 8, which provides discriminative information for differentiating FTLD from other dementias. These findings may provide valuable insights for clinicians to better understand the characteristics of different NDs.

We have demonstrated that the combination of different features including the grading features, the volume features and the CSF features can improve the classification accuracy. Age is also an important predictor in the differential diagnosis of dementias. It ranked top 30 in the utilized 824 features, highlighting that age is a risk factor for ND. It is reported in previous studies (Tong et al., 2017) that older subjects are more likely to develop AD than younger subjects. Other biomarkers such as cortical thickness (Hartikainen et al., 2012; Blanc et al., 2015), voxel-based morphometry (Davatzikos et al., 2008) and deformation-based morphometry (Muñoz-Ruiz et al., 2012) measures can be combined for classification. However, these structural biomarkers have shared information as our volumetric and grading features, thus may provide limited complementary information for improving the results. In contrast with these structural measures, biomarkers from other modalities such as functional imaging, neuropsychological tests and genetic data may have more valuable information for improving the classification performance. A further validation was carried out by combining the proposed features with additional cognitive scores, including MMSE, FAB, VAT, RAVLT, TMT, NPI and the Stroop test. The obtained overall accuracy has been significantly improved to 81% with a balanced accuracy of 78%. These cognitive scores add valuable information for classification. In addition to the combination of multimodal biomarkers, developing disease-specific biomarkers is also essential for improving the classification performance. For example, a VaD-specific biomarker called vascular burden was developed in Koikkalainen et al. (2016) to characterize the vascular changes so that the VaD patients can be separated from other groups. Although it requires the segmentation of lesions to quantify the vascular changes, the vascular burden biomarker can significantly improve the detection sensitivity of VaD patients to 96% as shown in Koikkalainen et al.

(2016), which is much higher than 74% obtained in our work, leading to a similar balanced accuracy as in our work. In addition, the sensitivity for detecting the DLB patients is low since there are no sufficient DLB-specific changes found in MR T1-weighted and FLAIR images. By adding neuropsychological tests, the sensitivity for diagnosing the DLB patients was improved from 38% to 70%. These cognitive scores are important features for improving the diagnosis accuracy of the DLB patients. In addition, it would be interesting to develop other DLB-specific biomarkers using modalities such as FDG-PET and DaTscan (McKeith et al., 2007) in future work.

In addition to the extraction of discriminative and complementary features, the adopted classifier also plays an important factor in the multi-class classification. In this work, we have shown that traditional classifiers including SVM and RF do not work well for the five-class classification because both classifiers are highly affected by the class imbalance problem. The RUSBoost method has been shown to provide a good solution to this problem in our work, which utilizes a sampling technique to handle class imbalance. Other recent approaches (Mac Aodha and Brostow, 2013; Bahnsen et al., 2015) may provide an alternative way for handling the class imbalance problem. Furthermore, in this work, different types of features were concatenated for classification. The feature concatenation provides a straightforward way to fuse multiple biomarkers, but may not optimally exploit the complementary information among different features for classification. How different types of biomarkers can be efficiently integrated requires further investigations. Finally, it should be mentioned that a patient diagnosed with a specific type of dementia may actually have mixed dementia. One previous study (Alzheimer's Association et al., 2015) shows that 54% of 141 volunteers, who had been diagnosed with AD, showed evidence of another type of dementia such as VaD or DLB according to their autopsies. Therefore, it would be interesting to generate probabilistic maps to visualize the likelihood of a region belonging to different NDs or shared by different NDs.

5. Conclusions

In this paper, a feature extraction and classification framework was proposed for the automatic differential diagnostics of four neurodegenerative diseases including AD, FTLD, VaD, and DLB, as well as patients with SMC. Structural volumes, structure-specific grading values and CSF features were extracted as biomarkers. The experimental results show that the best performance of the proposed framework was achieved when all the biomarkers were combined, indicating that there is complementary information among these biomarkers. In addition, the class imbalance problem poses challenges to traditional classifiers such as SVM and RF. The introduced RUSBoost algorithm has been shown to provide a good solution to this problem. In addition, a multi-class feature selection step has been demonstrated to improve the classification performance of the proposed framework. We also provide a way to quantify and investigate the importance of different features and regions. The proposed framework achieved a high classification accuracy of 75% for the five-class differential diagnostics of 500 patients, which is notable better than that obtained with visual MRI ratings (44.6%) as presented in Koikkalainen et al. (2016). Although the proposed framework is required a further improvement on accuracy before clinical use and further validations on different data sets are needed, it is encouraging to note that there is an increasing trend in the diagnosis accuracy with the efforts on this challenging task, which will finally lead to a prediction rate suitable for clinical use. Cost-efficiency analysis will then be investigated to decide what level of accuracy is acceptable in clinical practice in future studies.

Acknowledgements

This work was funded under the Seventh Framework Programme by the European Commission (http://cordis.europa.eu; EU-Grant-611005-PredictND; Name: From Patient Data to Clinical Diagnosis in Neurodegenerative Diseases).

References

Alzheimer's Association et al., 2015. 2015 Alzheimer's disease facts and figures. Alzheimers Dement. 11 (3) (332–332).
Argyriou, A.A., Evgeniou, T.T., Pontil, M.M., 2008. Convex multi-task feature learning. Mach. Learn. 73 (3), 243–272.
Bahnsen, A.C.A.C., Aouada, D.D., Ottersten, B.B., 2015. Ensemble of Example-dependent Cost-sensitive Decision Trees. (arXiv preprint arXiv:1505.04637).
Barber, R.R., Ballard, C.C., McKeith, I.I., Gholkar, A.A., O'brien, J.J., 2000. MRI volumetric study of dementia with Lewy bodies: a comparison with AD and vascular dementia. Neurology 54 (6), 1304–1309.
Barber, R.R., Gholkar, A.A., Scheltens, P.P., Ballard, C.C., McKeith, I.I., O'Brien, J.J., 1999. Medial temporal lobe atrophy on MRI in dementia with Lewy bodies. Neurology 52 (6) (1153–1153).
Blanc, F.F., Colloby, S.J.S.J., Philippi, N.N., de Pétigny, X.X., Jung, B.B., Demuynck, C.C., Phillipps, C.C., et al., 2015. Cortical thickness in dementia with Lewy bodies and Alzheimer's disease: a comparison of prodromal and dementia stages. PLoS One 10 (6), e0127396.
Burton, E.E., Barber, R.R., Mukaetova-Ladinska, E.E., Robson, J.J., Perry, R.R., Jaros, E.E., Kalaria, R.R., O'Brien, J.J., 2009. Medial temporal lobe atrophy on MRI differentiates Alzheimer's disease from dementia with Lewy bodies and vascular cognitive impairment: a prospective study with pathological verification of diagnosis. Brain 132 (1), 195–203.
Chawla, N.V.N.V., Bowyer, K.W.K.W., Hall, L.O.L.O., Kegelmeyer, W.P.W.P., 2002. Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. 321–357.
Chawla, N.V.N.V., Lazarevic, A.A., Hall, L.O.L.O., Bowyer, K.W.K.W., 2003. SMOTEBoost: improving prediction of the minority class in boosting. In: Knowledge Discovery in Databases. Springer, pp. 107–119.
Coupé, P.P., Eskildsen, S.F.S.F., Manjón, J.V.J.V., Fonov, V.S.V.S., Pruessner, J.C.J.C., Allard, M.M., Collins, D.L.D.L., et al., 2012. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. NeuroImage: Clin. 1 (1), 141–152.
Cuingnet, R.R., Gerardin, E.E., Tessieras, J.J., Auzias, G.G., Lehéricy, S.S., Habert, M.-O.M.-O., Chupin, M.M., Benali, H.H., Colliot, O.O., et al., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage 56 (2), 766–781.
Davatzikos, C.C., Resnick, S.M.S.M., Wu, X.X., Parmpi, P.P., Clark, C.M.C.M., 2008. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. NeuroImage 41 (4), 1220–1227.
Du, A.-T.A.-T., Schuff, N.N., Kramer, J.H.J.H., Rosen, H.J.H.J., Gorno-Tempini, M.L.M.L., Rankin, K.K., Miller, B.L.B.L., Weiner, M.W.M.W., 2007. Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. Brain 130 (4), 1159–1166.
Duara, R.R., Barker, W.W., Luis, C.A.C.A., 1999. Frontotemporal dementia and Alzheimer's disease: differential diagnosis. Dement. Geriatr. Cogn. Disord. 10 (Suppl. 1), 37–42.
Frisoni, G.B.G.B., Fox, N.C.N.C., Jack, C.R.C.R., Scheltens, P.P., Thompson, P.M.P.M., 2010. The clinical use of structural MRI in Alzheimer disease. Nat. Rev. Neurol. 6 (2), 67–77.
Grossman, M.M., McMillan, C.C., Moore, P.P., Ding, L.L., Glosser, G.G., Work, M.M., Gee, J.J., 2004. What's in a name: voxel-based morphometric analyses of MRI and naming difficulty in Alzheimer's disease, frontotemporal dementia and corticobasal degeneration. Brain 127 (3), 628–649.
Hand, D.J.D.J., Till, R.J.R.J., 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach. Learn. 45 (2), 171–186.
Hartikainen, P.P., Räsänen, J.J., Julkunen, V.V., Niskanen, E.E., Hallikainen, M.M., Kivipelto, M.M., Vanninen, R.R., Remes, A.M.A.M., Soininen, H.H., 2012. Cortical thickness in frontotemporal dementia, mild cognitive impairment, and Alzheimer's disease. J. Alzheimers Dis. 30 (4), 857–874.
Heckemann, R.A.R.A., Ledig, C.C., Gray, K.R.K.R., Aljabar, P.P., Rueckert, D.D., Hajnal, J.V.J.V., Hammers, A.A., 2015. Brain extraction using label propagation and group agreement: pincram. PLoS One 10 (7), e0129211.
Koikkalainen, J.J., Rhodius-Meester, H.H., Tolonen, A.A., Barkhof, F.F., Tijms, B.B., Lemstra, A.W.A.W., Tong, T.T., Guerrero, R.R., Schuh, A.A., Ledig, C.C., et al., 2016. Differential diagnosis of neurodegenerative diseases using structural MRI data. NeuroImage: Clin. 11, 435–449.
Ledig, C.C., Heckemann, R.A.R.A., Hammers, A.A., Lopez, J.C.J.C., Newcombe, V.F.V.F., Makropoulos, A.A., Lötjönen, J.J., Menon, D.K.D.K., Rueckert, D.D., 2015. Robust whole-brain segmentation: application to traumatic brain injury. Med. Image Anal. 21 (1), 40–58.
Liaw, A.A., Wiener, M.M., 2002. Classification and regression by random forest. R News 2 (3), 18–22.
Liu, J.J., Ji, S.S., Ye, J.J., 2009. Multi-task feature learning via efficient $l_{2,1}$-norm minimization. In: International Conference on Uncertainty in Artificial Intelligence. AUAI Press, pp. 339–348.
Mac Aodha, O.O., Brostow, G.G., 2013. Revisiting example dependent cost-sensitive learning with decision trees. In: IEEE International Conference on Computer Vision, pp. 193–200.
McKeith, I.I., Dickson, D.D., Lowe, J.J., Emre, M.M., O'Brien, J.J., Feldman, H.H., Cummings, J.J., Duda, J.J., Lippa, C.C., Perry, E.E., et al., 2005. Diagnosis and management of dementia with Lewy bodies: third report of the DLB consortium. Neurology 65 (12), 1863–1872.

McKeith, I.I., Mintzer, J.J., Aarsland, D.D., Burn, D.D., Chiu, H.H., Cohen-Mansfield, J.J., Dickson, D.D., Dubois, B.B., Duda, J.E.J.E., Feldman, H.H., et al., 2004. Dementia with Lewy bodies. Lancet Neurol. 3 (1), 19–28.
McKeith, I.I., O'Brien, J.J., Walker, Z.Z., Tatsch, K.K., Booij, J.J., Darcourt, J.J., Padovani, A.A., Giubbini, R.R., Bonuccelli, U.U., Volterrani, D.D., et al., 2007. Sensitivity and specificity of dopamine transporter imaging with 123 I-FP-CIT SPECT in dementia with Lewy bodies: a phase III, multicentre study. Lancet Neurol. 6 (4), 305–313.
McKhann, G.G., Drachman, D.D., Folstein, M.M., Katzman, R.R., Price, D.D., Stadlan, E.M.E.M., 1984. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 34 (7) (939–939).
McKhann, G.M.G.M., Knopman, D.S.D.S., Chertkow, H.H., Hyman, B.T.B.T., Jack, C.R.C.R., Kawas, C.H.C.H., Klunk, W.E.W.E., Koroshetz, W.J.W.J., Manly, J.J.J.J., Mayeux, R.R., et al., 2011. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 7 (3), 263–269.
Moradi, E.E., Pepe, A.A., Gaser, C.C., Huttunen, H.H., Tohka, J.J., et al., 2015. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. NeuroImage 104, 398–412.
Muñoz-Ruiz, M.Á.M.Á., Hartikainen, P.P., Koikkalainen, J.J., Wolz, R.R., Julkunen, V.V., Niskanen, E.E., Herukka, S.-K.S.-K., Kivipelto, M.M., Vanninen, R.R., Rueckert, D.D., et al., 2012. Structural MRI in frontotemporal dementia: comparisons between hippocampal volumetry, tensor-based morphometry and voxel-based morphometry. PLoS One 7 (12), e52531.
Neary, D.D., Snowden, J.S.J.S., Gustafson, L.L., Passant, U.U., Stuss, D.D., Black, S.A.S.A., Freedman, M.M., Kertesz, A.A., Robert, P.P., Albert, M.M., et al., 1998. Frontotemporal lobar degeneration a consensus on clinical diagnostic criteria. Neurology 51 (6), 1546–1554.
Ngai, S.S., Tang, Y.Y., Du, L.L., Stuckey, S.S., 2007. Hyperintensity of the precentral gyral subcortical white matter and hypointensity of the precentral gyrus on fluid-attenuated inversion recovery: variation with age and implications for the diagnosis of amyotrophic lateral sclerosis. Am. J. Neuroradiol. 28 (2), 250–254.
Nyúl, L.G.L.G., Udupa, J.K.J.K., 1999. On standardizing the MR image intensity scale. Magn. Reson. Med. 42 (6), 1072–1081.
Oishi, K.K., Faria, A.A., Jiang, H.H., Li, X.X., Akhter, K.K., Zhang, J.J., Hsu, J.T.J.T., Miller, M.I.M.I., van Zijl, P.C.P.C., Albert, M.M., et al., 2009. Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and Alzheimer's disease participants. NeuroImage 46 (2), 486–499.
Raamana, P.R.P.R., Rosen, H.H., Miller, B.B., Weiner, M.W.M.W., Wang, L.L., Beg, M.F.M.F., 2014. Three-class differential diagnosis among Alzheimer disease, frontotemporal dementia, and controls. Front. Neurol. 5, 71.
Rascovsky, K.K., Hodges, J.R.J.R., Knopman, D.D., Mendez, M.F.M.F., Kramer, J.H.J.H., Neuhaus, J.J., van Swieten, J.C.J.C., Seelaar, H.H., Dopper, E.G.E.G., Onyike, C.U.C.U., et al., 2011. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain 134 (9), 2456–2477.
Rätsch, G.G., Onoda, T.T., Müller, K.-R.K.-R., 2001. Soft margins for AdaBoost. Mach. Learn. 42 (3), 287–320.
Román, G.C.G.C., Tatemichi, T.K.T.K., Erkinjuntti, T.T., Cummings, J.J., Masdeu, J.C.J.C., Garcia, J.H.J.H., Amaducci, L.L., Orgogozo, J.-M.J.-M., Brun, A.A., Hofman, A.A., et al., 1993. Vascular dementia diagnostic criteria for research studies: report of the NINDS-AIREN International Workshop. Neurology 43 (2) (250–250).
Rueckert, D.D., Sonoda, L.I.L.I., Hayes, C.C., Hill, D.L.G.D.L.G., Leach, M.O.M.O., Hawkes, D.J.D.J., 1999. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imaging 18 (8), 712–721.
Saeys, Y.Y., Inza, I.I., Larrañaga, P.P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23 (19), 2507–2517.
Schoonenboom, N.N., Reesink, F.F., Verwey, N.N., Kester, M.M., Teunissen, C.C., Van De Ven, P.P., Pijnenburg, Y.Y., Blankenstein, M.M., Rozemuller, A.A., Scheltens, P.P., et al., 2012. Cerebrospinal fluid markers for differential dementia diagnosis in a large memory clinic cohort. Neurology 78 (1), 47–54.
Schott, J.M.J.M., Fox, N.C.N.C., 2007. Structural imaging in the dementias. Psychiatry 6 (12), 503–507.
Seiffert, C.C., Khoshgoftaar, T.M.T.M., Van Hulse, J.J., Napolitano, A.A., 2010. RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Trans. Syst. Man Cybern. Syst. Hum. 40 (1), 185–197.
Sheline, Y.I.Y.I., Price, J.L.J.L., Vaishnavi, B.S.N.B.S.N., Mintun, M.A.M.A., Barch, D.M.D.M., Epstein, A.A.A.A., Wilkins, C.H.C.H., Snyder, A.Z.A.Z., Couture, L.L., Schechtman, K.K., et al., 2008. Regional white matter hyperintensity burden in automated segmentation distinguishes late-life depressed subjects from comparison subjects matched for vascular risk factors. Am. J. Psychiatry 165, 524–532.
Thai-Nghe, N.N., Gantner, Z.Z., Schmidt-Thieme, L.L., 2010. Cost-sensitive learning methods for imbalanced data. In: International Joint Conference on Neural Networks. IEEE, pp. 1–8.
Thung, K.-H.K.-H., Wee, C.-Y.C.-Y., Yap, P.-T.P.-T., Shen, D.D., et al., 2014. Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion. NeuroImage 91, 386–400.
Tong, T.T., Gao, Q.Q., Guerrero, R.R., Ledig, C.C., Chen, L.L., Rueckert, D.D., et al., 2017. A novel grading biomarker for the prediction of conversion from mild cognitive impairment to Alzheimer's disease. IEEE Trans. Biomed. Eng. 64 (1), 155–165.
Tong, T.T., Wolz, R.R., Gao, Q.Q., Guerrero, R.R., Hajnal, J.V.J.V., Rueckert, D.D., et al., 2014. Multiple instance learning for classification of dementia in brain MRI. Med. Image Anal. 18 (5), 808–818.
Tustison, N.J.N.J., Avants, B.B.B.B., Cook, P.P., Zheng, Y.Y., Egan, A.A., Yushkevich, P.P., Gee, J.C.J.C., et al., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320.
van der Flier, W.M.W.M., Pijnenburg, Y.Y., Prins, N.N., Lemstra, A.W.A.W., Bouwman, F.H.F.H., Teunissen, C.E.C.E., van Berckel, B.B., Stam, C.J.C.J., Barkhof, F.F., Visser, P.J.P.J., et al., 2014. Optimizing patient care and research: the Amsterdam Dementia Cohort. J. Alzheimers Dis. 41 (1), 313–327.
Van Straaten, E.E., Scheltens, P.P., Barkhof, F.F., 2004. MRI and CT in the diagnosis of vascular dementia. J. Neurol. Sci. 226 (1), 9–12.
Varma, A.A., Adams, W.W., Lloyd, J.J., Carson, K.K., Snowden, J.J., Testa, H.H., Jackson, A.A., Neary, D.D., 2002. Diagnostic patterns of regional atrophy on MRI and regional cerebral blood flow change on SPECT in young onset patients with Alzheimer's disease, frontotemporal dementia and vascular dementia. Acta Neurol. Scand. 105 (4), 261–269.
Wang, H.H., Nie, F.F., Huang, H.H., Risacher, S.S., Ding, C.C., Saykin, A.J.A.J., Shen, L.L., 2011. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In: IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 557–562.
Whitwell, J.L.J.L., Weigand, S.D.S.D., Shiung, M.M.M.M., Boeve, B.F.B.F., Ferman, T.J.T.J., Smith, G.E.G.E., Knopman, D.S.D.S., Petersen, R.C.R.C., Benarroch, E.E.E.E., Josephs, K.A.K.A., et al., 2007. Focal atrophy in dementia with Lewy bodies on MRI: a distinct pattern from Alzheimer's disease. Brain 130 (3), 708–719.
Zarow, C.C., Vinters, H.V.H.V., Ellis, W.G.W.G., Weiner, M.W.M.W., Mungas, D.D., White, L.L., Chui, H.C.H.C., 2005. Correlates of hippocampal neuron number in Alzheimer's disease and ischemic vascular dementia. Ann. Neurol. 57 (6), 896–903.
Zhang, Y.Y., Zhou, Z.-H.Z.-H., 2010. Cost-sensitive face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 32 (10), 1758–1769.