Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2016, Article ID 9523849, 10 pages
http://dx.doi.org/10.1155/2016/9523849

Research Article

Centered Kernel Alignment Enhancing Neural Network Pretraining for MRI-Based Dementia Diagnosis

David Cárdenas-Peña, Diego Collazos-Huertas, and German Castellanos-Dominguez

Signal Processing and Recognition Group, Universidad Nacional de Colombia, Manizales, Colombia

Correspondence should be addressed to David Cárdenas-Peña; dcardenasp@unal.edu.co

Received 24 December 2015; Accepted 3 March 2016

Academic Editor: Dwarikanath Mahapatra

Dementia is a growing problem that affects elderly people worldwide. More accurate evaluation of dementia diagnosis can help during the medical examination. Several methods for computer-aided dementia diagnosis have been proposed using resonance imaging scans to discriminate between patients with Alzheimer's disease (AD) or mild cognitive impairment (MCI) and healthy controls (NC). Nonetheless, the computer-aided diagnosis is especially challenging because of the heterogeneous and intermediate nature of MCI. We address the automated dementia diagnosis by introducing a novel supervised pretraining approach that takes advantage of the artificial neural network (ANN) for complex classification tasks. The proposal initializes an ANN based on linear projections to achieve more discriminating spaces. Such projections are estimated by maximizing the centered kernel alignment criterion that assesses the affinity between the resonance imaging data kernel matrix and the label target matrix. As a result, the performed linear embedding allows accounting for features that contribute the most to the MCI class discrimination. We compare the supervised pretraining approach to two unsupervised initialization methods (autoencoders and Principal Component Analysis) and against the best four performing classification methods of the 2014 *CADDementia* challenge. As a result, our proposal outperforms all the baselines (7% of classification accuracy and area under the receiver-operating-characteristic curve) at the time it reduces the class biasing.

1. Introduction

In 2010, the number of people aged over 60 years with dementia was estimated at 35.6 million worldwide and this figure had been expected to double over the next two decades [1]. Actually, World Health Organization and the Alzheimer's Disease International had declared dementia as a public health priority, encouraging articulating government policies and promoting actions at international and national levels [2]. Alzheimer's disease (AD) is the most diagnosed dementia-related chronic illness that demands very expensive costs of care, living arrangements, and therapies. Thus, efforts are underway to improve treatment which may delay, at least, one year the AD onset and development, leading to decreasing the number of cases by nine millions [3]. AD can be early diagnosed by predicting the conversion to dementia from a state of mild cognitive impairment (MCI) that especially increases the AD risk [4].

In this regard, early diagnosis is directly related to the effectiveness of interventions [5]. Along with clinical history, neuropsychological tests, and laboratory assessment, the joint clinical diagnosis of AD also includes neuroimaging techniques like positron emission tomography (PET) and magnetic resonance imaging (MRI). These techniques are usually incorporated in the routine workup for excluding secondary pathology causes (e.g., tumors) [6, 7]. However, factors related to image quality and radiologist experience may limit their use [8]. For dealing with this issue, the imaging-based automatic assessment of quantitative biomarkers has been proven to enhance the performance for dementia diagnosis. In the particular case of AD, there are two groups of widely studied biomarkers: (i) patterns of brain amyloid-beta, such as low cerebrospinal fluid (CSF) $A\beta 42$ and amyloid PET imaging, and (ii) measures of neuronal injury or degeneration like CSF tau measurement, fluorodeoxyglucose PET, and atrophy on structural MRI [9]. Thus, structural MRI

has become valuable for biomarker assessment since this noninvasive technique explains structural changes at the onset of cognitive impairment [10].

For the purpose of automated diagnosis, the first stage to implement is the structure-wise feature extraction from available MRI data, including voxel-based morphometry, volume, thickness, shape, and intensity relation. Nonetheless, more emphasis usually focuses on the classification approach due to its strong influence on the entire diagnosis system. With regard to neurodegenerative diseases, the reported classifiers range from straightforward approaches (k-Nearest Neighbors [11], Linear Discriminant Analysis [12], Support Vector Machines [13], Random Forests [14], and Regressions [15]) to the combination of classifiers [16]. Most of the above approaches had been evaluated for the *2014 CADDementia challenge* which aimed to reproduce the clinical diagnosis of 354 subjects in a multiclass classification problem of three diagnostic groups [17], Alzheimer's diagnosed patients, subjects with MCI, and healthy controls (NC), given their T1-weighted MRI scans. As a result, the best-performing algorithm yielded an accuracy of 63.0% and an area under the receiver-operating-characteristic (ROC) curve of 78.8%. Nonetheless, reported true positive rates are 96.9% and 28.7% for NC and MCI, respectively, resulting in class biasing.

Generally speaking, dementia diagnosis from MRI still remains a challenging task, mainly, because of the nature of mild cognitive impairment; that is, there is a heterogeneous and intermediate category between the NC and AD diagnostic groups, from which subjects may convert to AD or return to the normal cognition [4]. For overcoming this shortcoming, machine learning tools as the artificial neural networks (ANN) have been developed to enhance dementia diagnosis, presenting the following advantages [18, 19]: (i) ability to process a large amount of data, (ii) reduced likelihood of overlooking relevant information, and (iii) reduction of diagnosis time.

Nonetheless, an essential procedure for ANN implementation is initializing deep architecture (termed pretraining) which can be carried out by training a deep network to optimize directly only the supervised objective of interest, starting from a set of randomly initialized parameters. However, this strategy performs poorly in practice [20]. With the aim to improve each initial-random guess, a local unsupervised criterion is considered to pretrain each layer stepwise, trying to produce a useful higher-level description based on the adjacent low-level representation output of the previous layer. Particular examples that use unsupervised learning are the following: Restricted Boltzmann Machines [21], autoencoders [22], sparse autoencoders [23], and the greedy layer-wise unsupervised learning which is the most common approach that learns one layer of a deep architecture at a time [24]. Although the unsupervised pretraining generates hidden representations that are more useful than the input space, many of the resulting features may be irrelevant for the discrimination task [25, 26].

In this paper, we benefit from the ANN advantages for complex classification tasks to introduce a novel supervised ANN initialization approach devoted to the automated dementia diagnosis. The proposed pretraining approach searches for a linear projection into a more discriminating space so that the resulting embedding features and labels become as much as possible associated. Consequently, the obtained ANN architecture should match better the nature of supervised training data. Taking into account the fact that the ANN straightforward hybridization with other approaches yields stronger paradigms for solving complex and computationally expensive problems [27, 28], we also incorporate kernel theory for assessing the affinity between projected data and available labels. The use of kernel approaches offers an elegant, functional analysis framework for tasks, gathering multiple information sources (e.g., features and labels) as the minimum variance unbiased estimation of regression coefficients and least squares estimation of random variables [29]. Moreover, we consider the centered kernel alignment criterion as the affinity measure between a data kernel matrix and a target label matrix [30, 31]. As a result, the linear embedding allows accounting for features that contribute the most to the class discrimination.

The present paper is organized as follows: Section 2 firstly describes the mathematical background on learning projections using CKA and ANN for classification. Section 3 introduces all the carried out experiments for tuning the algorithm parameters and the evaluation scheme with blinded data. Then, achieved results are discussed in Section 4. Finally, Section 5 presents the concluding remarks and future research directions.

2. Materials and Methods

2.1. Classification Using Artificial Neural Networks. Within the classification framework, an L-layered ANN is assumed to predict the needed class label set through a battery of feedforward deterministic transformations, which are implemented by the hidden layers $\mathbf{h}^l$, which map the input space $\mathbf{x}$ to the network output $\mathbf{h}^L$ as follows [27]:

$$\begin{aligned} \mathbf{h}^l &= \phi\left(\mathbf{b}^l + \mathbf{W}^l \mathbf{h}^{l-1}\right), \quad \forall l = 1, \ldots, L-1, \\ \mathbf{h}^0 &= \mathbf{x}, \end{aligned} \tag{1}$$

where $\mathbf{b}^l \in \mathbb{R}^{m_{l+1}}$ is the lth offset vector, $\mathbf{W}^l \in \mathbb{R}^{m_{l+1} \times m_l}$ is the lth linear projection, and $m_l \in \mathbb{Z}^+$ is the size of the lth layer. The function $\phi(\cdot) \in \mathbb{R}$ applies saturating, nonlinear, element-wise operations. Here, we choose the standard sigmoid, $\phi(z) = \text{sigmoid}(z)$, expressed as follows:

$$\text{sigmoid}(z) = \frac{(\tanh(z) + 1)}{2}. \tag{2}$$

The first layer in (1) (i.e., $\mathbf{h}^0 \in \mathbb{R}^D$) is conventionally adjusted to the input feature vector. In turn, the output layer $\mathbf{h}^L \in [0, 1]^C$ predicts the class when combined with a provided target $t \in \{1, \ldots, C\}$ into a loss function $\mathscr{L}(\mathbf{h}^L, t)$. In practice, the output layer can be carried out by the nonlinear softmax function described as follows:

$$h_c^L = \frac{\exp\left(b_c^L + \mathbf{w}_c^L \mathbf{h}^{L-1}\right)}{\sum_j \exp\left(b_c^L + \mathbf{w}_c^L \mathbf{h}^{L-1}\right)}, \tag{3}$$

where b_c^L is the cth element of $\mathbf{b}^L$, $\mathbf{w}_c^L$ is the cth row of $\mathbf{W}^L$, $\mathbf{h}^L$ is positive, and $\sum_c h_c^L = 1$.

The rationale behind the choice of softmax function is that each yielded output h_c^L can be used as an estimator of $P(t_i = c \mid \mathbf{x}_i)$, so that the interpretation of t_i relates to the class associated with input pattern $\mathbf{x}_i$. In this case, the softmax loss function corresponds often to the negative conditional log-likelihood:

$$\mathscr{L}\left(\mathbf{h}^L, t\right) = -\log \sum_c P\left(t = c \mid \mathbf{x}\right). \tag{4}$$

Therefore, the expected value over $(\mathbf{x}, t)$ pairs is minimized with respect to the biases and weighting matrices.

2.2. ANN Pretraining Using Centered Kernel Alignment. Let $\mathbf{X} \in \{\mathbf{x}_i \in \mathbb{R}^D : i \in N\}$ be the input feature matrix with size $\mathbb{R}^{D \times N}$ which holds N trajectories and let $\mathbf{x}_i \subset \mathscr{X}$ be a D-dimensional random process. In order to encode the affinity between a couple of trajectories, $\{\mathbf{x}_i, \mathbf{x}_j\}$, we determine the following kernel function:

$$\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left\langle \varphi\left(\mathbf{x}_i\right), \varphi\left(\mathbf{x}_j\right)\right\rangle, \quad \forall i, j \in N. \tag{5}$$

$\langle \cdot, \cdot \rangle$ stands for the inner product and $\varphi(\cdot) : \mathbb{R}^D \to \mathscr{H}$ maps from the original domain, $\mathbb{R}^D$, into a Reproduced Kernel Hilbert Space (RKHS), $\mathscr{H}$. As a rule, it holds that $|\mathscr{H}| \to \infty$, so that $|\mathbb{R}^D| \ll |\mathscr{H}|$ can be assumed. Nevertheless, there is no need for computing $\varphi(\cdot)$ directly. Instead, the well-known *kernel trick* is employed for computing (5) through the positive definite and infinitely divisible kernel function as follows:

$$k_{ij} = \kappa\left(d\left(\mathbf{x}_i, \mathbf{x}_j\right)\right), \tag{6}$$

where $d : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{R}^+$ is a distance operator implementing the positive definite kernel function $\kappa(\cdot)$. A kernel matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$ that results from the application of κ over each sample pair in $\mathbf{X}$ is assumed as the covariance estimator of the random process $\mathscr{X}$ over the RKHS.

With the purpose of improving the system performance in terms of learning speed and classification accuracy, we introduce the prior label knowledge into the initialization process. Thus, we compute the pairwise relations between the feature vectors through the introduced feature similarity kernel matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$ which has elements as follows:

$$k_{ij} = \kappa_{\mathbf{x}}\left(d_W\left(\mathbf{x}_i, \mathbf{x}_j\right)\right), \quad \forall i, j \in \{1, \ldots, N\}, \tag{7}$$

with $d_W : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{R}^+$ being a distance operator that implements the positive definite kernel function $\kappa_{\mathbf{x}}(\cdot)$, and $\{(\mathbf{x}_i, t_i) : i = 1, \ldots, N\}$ is a set of input-label pairs with $\mathbf{x}_i \in \mathbb{R}^D$ and $t_i \in \{1, C\}$, with C being the number of classes to identify.

Since we look for a suitable weighting matrix for initializing the ANN optimization, we rely on the Mahalanobis distance that is defined on a D-dimensional space by the following inverse covariance matrix $\mathbf{W}^\top \mathbf{W}$:

$$d_W\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^\top \mathbf{W}^\top \mathbf{W}\left(\mathbf{x}_i - \mathbf{x}_j\right), \tag{8}$$

where matrix $\mathbf{W} \in \mathbb{R}^{m_1 \times D}$ holds the linear projection $\mathbf{y}_i = \mathbf{W}\mathbf{x}_i$, with $\mathbf{y}_i \in \mathbb{R}^{m_1}$, $m_1 \leq D$.

Based on the already estimated feature similarities, we propose further to learn the matrix $\mathbf{W}$ by adding the prior knowledge about the feasible sample membership (e.g., healthy or diseased groups) enclosed in a matrix $\mathbf{B} \in \mathbb{R}^{N \times N}$ with elements $b_{ij} = \delta(t_i - t_j)$. Thus, we measure the similarity between the matrices $\mathbf{K}$ and $\mathbf{B}$ through the following function of centered kernel alignment (CKA) [32]:

$$\rho\left(\mathbf{K}, \mathbf{B}\right) = \frac{\langle \mathbf{HKH}, \mathbf{HBH}\rangle_F}{\|\mathbf{HKH}\|_F \|\mathbf{HBH}\|_F}, \quad \rho \in [0, 1], \tag{9}$$

where $\mathbf{H} = \mathbf{I} - N^{-1}\mathbf{1}\mathbf{1}^\top$, with $\mathbf{H} \in \mathbb{R}^{N \times N}$, is a centering matrix, $\mathbf{1} \in \mathbb{R}^N$ is an all-ones vector, and $\langle \cdot, \cdot \rangle_F$ and $\|\cdot, \cdot\|_F$ stand for the Frobenius inner product and norm, respectively.

Therefore, the centered version of the alignment coefficient leads to better correlation estimation compared to its uncentered version [31]. Therefore, the CKA cost function, described in (9), highlights relevant features by learning the matrix $\mathbf{W}$ that best matches all relations between the resulting feature vectors and provided target classes. Consequently, we state the following optimization problem to compute the projection matrix:

$$\mathbf{W}^\star = \arg\max_{\mathbf{W}} \rho\left(\mathbf{K}_{\mathbf{W}}, \mathbf{B}\right), \tag{10}$$

and we thus initialize the first layer of the ANN with $\mathbf{W}^\star$.

Additionally, the weighting matrix allows analyzing the contribution of the input feature set for building the projection matrix by computing the feature relevance vector $\boldsymbol{\varrho} \in \mathbb{R}^D$ in the following form:

$$\varrho_d = \mathscr{E}\left\{w_{ud}^2 : \forall u \in \left[1, m_1\right]\right\}, \tag{11}$$

where $w_{ud} \in \mathbb{R}$ is the weight that associates each dth feature to uth hidden neuron. $\mathscr{E}\{\cdot\}$ stands for the averaging operator. The main assumption behind the introduced relevance in (11) is that the larger the values of ϱ_d the larger the dependency of the estimated embedding on the input attribute.

3. Experimental Setup

An automated, computer-aided diagnosis system based on artificial neural networks is introduced to classify structural magnetic resonance imaging (MRI) scans in accordance with the following three neurological classes: normal control (NC), mild cognitive impairment (MCI), and Alzheimer's disease (AD). Figure 1 illustrates the methodological development of the proposed approach.

3.1. ADNI Data. Data used in the preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/) which was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and nonprofit

Table 1: Demographic and clinical details of the selected ADNI cohort.

	"best" quality			"partial" quality		
	NC	MCI	AD	NC	MCI	AD
N	655	825	513	465	130	34
Age	74.9 ± 5.0	74.4 ± 7.4	74.0 ± 7.4	76.6 ± 6.4	76.0 ± 6.3	74.3 ± 6.5
Male	47.5%	39.5%	47.6%	70.1%	62.3%	70.6%
MMSE	29.0 ± 1.0	27.1 ± 2.5	21.9 ± 4.4	27.5 ± 2.0	21.2 ± 1.6	14.4 ± 2.8

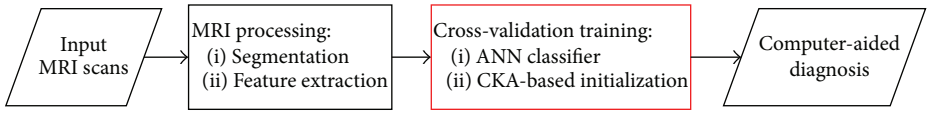


Figure 1: General processing pipeline: FreeSurfer independently segments and extracts features from given MRIs. Centered kernel alignment is proposed to learn a projection matrix initializing the NN training in a 5-fold cross-validation scheme. Tuned model is used for classification task.

organizations. The primary goal of ADNI is 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). From the ADNI 1, ADNI 2, and ADNI GO phases, we selected a subset of 633 subjects with scans that had been noted with the "best" quality mark. As a result, the selected subset holds $N = 1993$ images with three class labels described above; $C = 3$. Besides, a random subset of 70% data was chosen for tuning and training stages, while the remaining 30% is for the test purpose. In addition, 629 images with a "partial" quality mark were selected in order to assess the performance under more complicated imaging conditions. Table 1 briefly describes the demographic information for the ADNI selected cohort.

3.2. Processing of MRI Data. We used FreeSurfer, version 5.1 (a free available (http://surfer.nmr.mgh.harvard.edu/), widely used and extensively validated brain MRI analysis software package), to process the structural brain MRI scans and compute the morphological measurements [33]. FreeSurfer morphometric procedures have been demonstrated to show good test-retest reliability across scanner manufacturers and across field strengths [34]. The FreeSurfer pipeline is fully automatic and includes the next procedures: a watershed-based skull stripping [35], a transformation to the Talairach, an intensity normalization and bias field correction [36], tessellation of the gray/white matter boundary, topology correction [37], and a surface deformation [38]. Consequently, a representation of the cortical surface between white and gray matters, of the pial surface, and segmentation of white matter from the rest of the brain are obtained. FreeSurfer computes structure-specific volume, area, and thickness measurements. Cortical Volumes and Subcortical Volumes are normalized to each subject's Total Intracranial Volume (eTIV) [39]. Table 2 summarizes the five feature sets extracted for each subject, which are concatenated into the feature matrix $\mathbf{X}$ with dimensions $N = 1993$ and $D = 324$.

Table 2: FreeSurfer extracted features. # stands for the number of features.

Type	# features	Units
Cortical Volumes (CV)	70	mm^3
Subcortical Volumes (SV)	42	mm^3
Surface Area (SA)	72	mm^2
Thickness Average (TA)	70	mm
Thickness Std. (TS)	70	mm
Total	324	

3.3. Tuning of ANN Model Parameter. Given input $D = 324$ MRI features for classification of the 3 neurological classes, we use the feedforward ANNs with one hidden layer: 324-input and 3-output neurons. An exhaustive search is carried out for tuning the single free parameter, namely, the number of neurons in the hidden layer (m_1). We also compare our proposal against autoencoders (AEN) [20] and the well-known Principal Components Analysis (PCA) for the initialization stage. All of these approaches (AEN, PCA, and CKA) provide a projection matrix with an output dimension that, in this case, equates the hidden layer size. Thus, resulting projections are used as the initial weights for the first layer. Also, biases and output layer weights are randomly initialized. For a different number of neurons, Figure 2 shows the accuracy results obtained by each considered strategy of initialization using 5-fold cross-validation scheme. Since we look for the most accurate and stable network configuration, we chose the optimal net as the one with the highest mean-to-deviation ratio. The resulting search indicates that the best number of hidden neurons is accomplished at $m_1 = 20$, $m_1 = 16$, and $m_1 = 14$ for AEN, PCA, and CKA approaches, respectively.

We further analyze the influence of each feature to the initialization process regarding the relevance criterion introduced in (11). Obtained results of relevance in Figure 3 show that the proposed CKA approach enhances the Subcortical

Table 3: Best performing algorithms in the 2014 CADDementia challenge [17].

Algorithm	Features	Classifier
Abdulkadir	Voxel-based morphometry	Support Vector Machine
Ledig	Volume and intensity relations	Random Forest classifier
Sørensen	Volume, thickness, shape, and intensity relations	Regularized Linear Discriminant Analysis
Wachinger	Volume, thickness, and shape	Generalized Linear Model

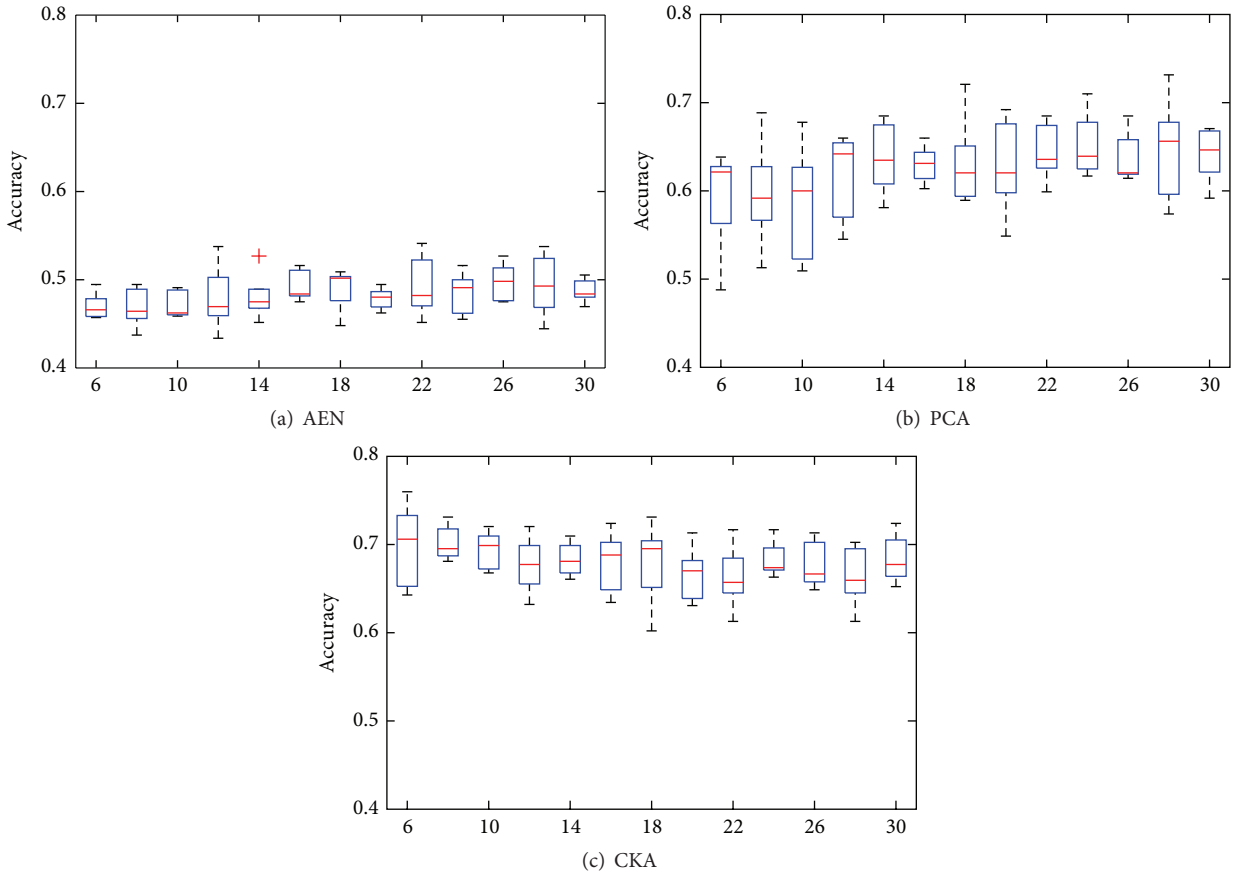


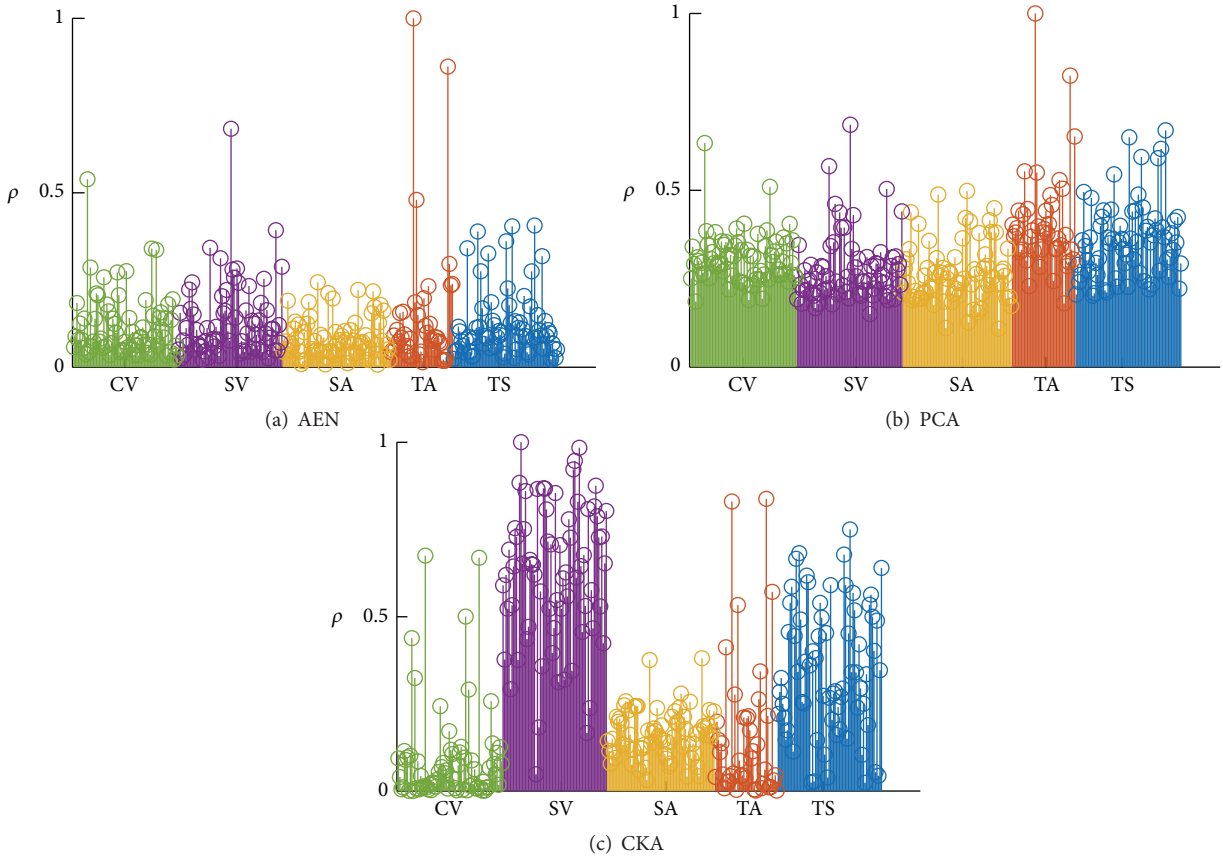
Figure 2: Artificial neural network performance along the number of nodes in the hidden layer (m_1) for the three initialization approaches: autoencoder, PCA-based projection, and CKA-based projection. Results are computed under 5-fold cross-validation scheme.

Volume features at the time it diminishes the influence of most Cortical Volumes and Thickness Averages. The relevance of each feature set provided by AEN and PCA is practically the same. Hence, CKA allows the selection of relevant biomarkers from MRI.

3.4. Classifier Performance of Neurological Classes. As shown in Table 3, the ANN models that have been tuned for the three initialization strategies are contrasted with the best four performing approaches of the 2014 CADDementia challenge [17]. The compared algorithms are evaluated in terms of their classification performance, accuracy (α), area under the receiver-operating-characteristic curve (β), and class-wise true positive rate (τ_p^c) criteria, respectively, which are defined as

$$\begin{aligned} \alpha &= \frac{\sum_c \left(t_p^c + t_n^c\right)}{\sum_c N^c}, \\ \tau^c &= \frac{t_p^c}{N^c}, \\ \beta &= \frac{\sum_c \beta^c \cdot N^c}{\sum_c N^c}, \end{aligned} \tag{12}$$

where $c \in \{\text{NC}, \text{MCI}, \text{AD}\}$ indexes each class and N^c, t_p^c, and t_n^c are the number of samples, true positives, and true negatives for the cth class, respectively. The area under the



(a) AEN (b) PCA

(c) CKA

Figure 3: Relevance indexes grouped by feature type: Cortical Volume (CV), Subcortical Volume (SV), Surface Area (SA), Thickness Average (TA), and Thickness Std. (TS).

curve β is the weighted average of the area under the ROC curve of each class β^c. Presented results for the baseline approaches are the ones reported on the challenge for 354 images. Although the testing groups on the challenge and on this paper are not exactly the same, the amount of data, their characteristics, and the blind setup make those two groups equivalent for evaluation purposes.

As seen in Table 4 which compares the classification performance on the 30% "best" quality test set for considered algorithms, the proposed approach, besides outperforming other compared approaches of initialization, also performs better than other computer-aided diagnosis methods as a whole. For the "partial" quality images, as expected, the general performance diminishes in all ANN approaches. Nonetheless, the overall accuracy and AUC are still competitive with respect to the challenge winner. Based on the displayed ROC curves and confusion matrices for the ANN-based classifiers with the optimum parameter set (see Figure 4), we also infer that the proposed approach improves MCI discrimination.

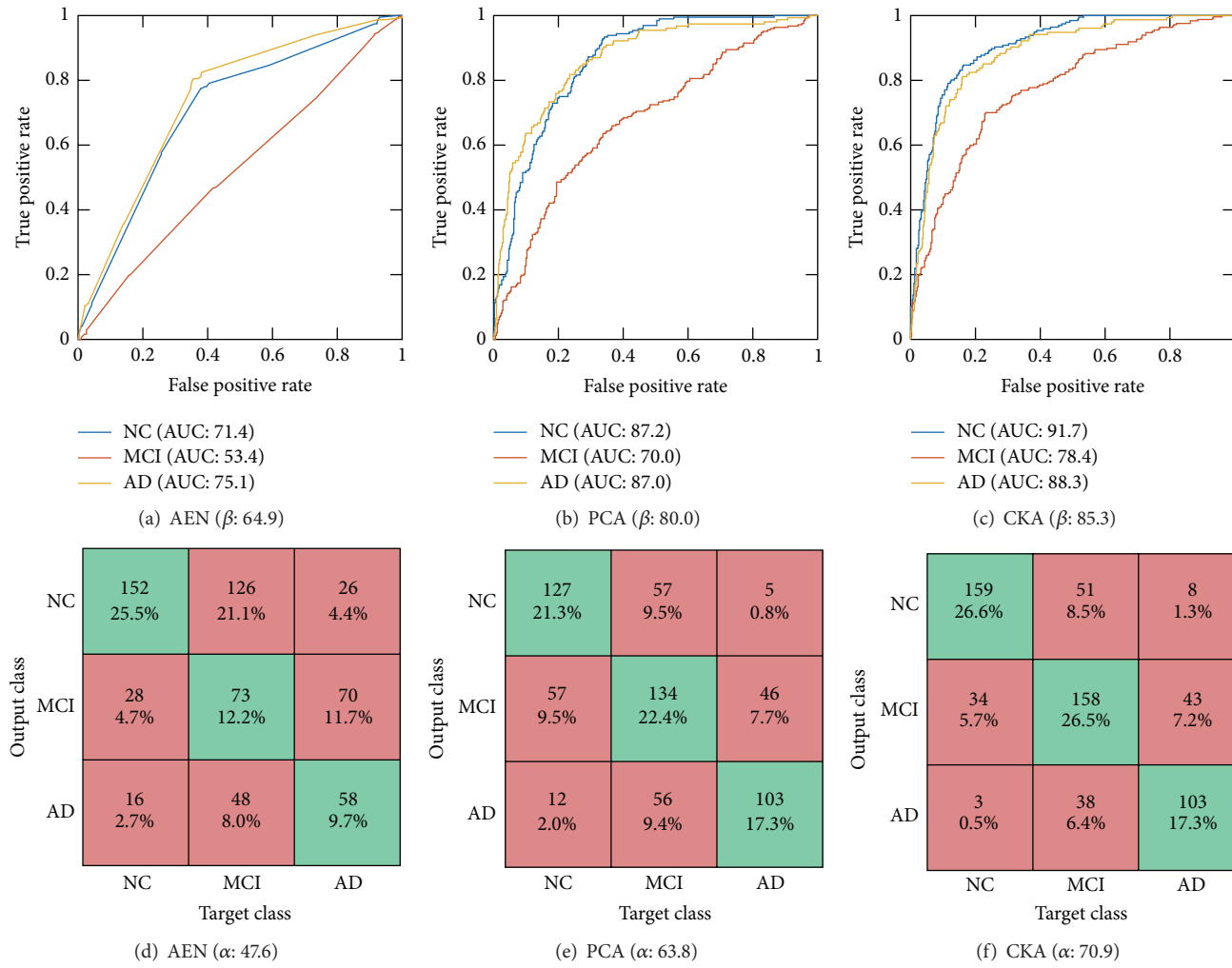
Table 4: Classification performance on the testing groups for considered algorithms under evaluation criteria. Top: baseline approaches. Bottom: ANN pretrainings.

Algorithm	α	τ^{NC}	τ^{MCI}	τ^{AD}	β	β^{NC}	β^{MCI}	β^{AD}
2014 CADDementia								
Sørensen	**63.0**	**96.9**	28.7	**61.2**	**78.8**	86.3	**63.1**	87.5
Wachinger	59.0	72.1	51.6	51.5	77.0	83.3	59.4	**88.2**
Ledig	57.9	89.1	41.0	38.8	76.7	**86.6**	59.7	84.9
Abdulkadir	53.7	45.7	**65.6**	49.5	77.7	85.6	59.9	86.7
"best" quality testing								
NN-AEN	47.6	73.4	33.1	38.1	64.9	71.4	53.4	75.1
NN-PCA	63.8	70.4	56.7	66.9	80.0	87.2	70.0	87.0
NN-CKA	**70.9**	78.4	**66.6**	**68.3**	**85.3**	**91.7**	**78.4**	**88.3**
"partial" quality								
NN-AEN	62.9	64.6	46.4	32.0	77.0	82.5	65.6	72.5
NN-PCA	64.4	67.6	**49.3**	26.0	78.4	82.3	67.5	79.2
NN-CKA	**65.2**	68.6	38.6	42.0	**81.6**	85.7	**70.1**	82.4

4. Discussion

From the validation carried out above for MRI-based dementia diagnosis, the following aspects emerge as relevant for the developed proposal of ANN pretraining:

(i) As commonly implemented by the state-of-the-art ANN algorithms, the proposed initialization approach also has one free model parameter which is the number of hidden neurons. Tuning of this parameter is proposed to be carried out heuristically



(a) AEN (β: 64.9) (b) PCA (β: 80.0) (c) CKA (β: 85.3)

(d) AEN (α: 47.6) (e) PCA (α: 63.8) (f) CKA (α: 70.9)

FIGURE 4: Receiver-operating-characteristic curve ((a), (b), and (c)) and confusion matrix ((d), (e), and (f)) on the 30% test data for AEN ((a) and (d)), PCA ((b) and (e)), and CKA ((c) and (f)) initialization approaches at the best parameter set of the ANN classifier.

by an exhaustive search so as to reach the highest accuracy on a 5-fold cross-validation (see Figure 2). Thus, 24, 20, and 16 hidden neurons are selected for CKA, AEN, and PCA, respectively. As a result, the suggested CKA approach improves other pretraining ANN approaches (in about 10%) with the additional benefit of decreasing the performed parameter sensitivity.

(ii) We assess the influence of each MRI feature at the pretraining procedure regarding the relevance criterion introduced in (11). As follows from Figure 3, AEN and PCA ponder every feature evenly, restraining their ability to extract biomarkers. By contrast, CKA enhances the influence of Subcortical Volumes and Thickness Standard deviations at the time it diminishes the contribution of Cortical Volumes and Thickness Averages. Consequently, the proposed approach is also suitable for feature selection tasks.

(iii) In the interest of comparing, we contrast the developed ANN pretraining approach with the best four classification strategies of the 2014 CADDementia, devoted especially to dementia classification. From the obtained results, summarized in Table 4, it follows that proposed CKA outperforms other algorithms in most of the evaluation criteria and imaging conditions, providing the most balanced performance over all classes. Particularly for the 30% testing images, CKA increases by 7%-points the classification accuracy and average area under the ROC curve. It is worth noting that although Sørensen's approach accomplishes a τ^{NC} value that is 18.5%-points higher than the proposal, its performance turns out to be biased towards the NC, yielding a worse value of MCI. That is, CKA carries out unbiased class performance of the dementia classification. In the case of "partial" quality images, in spite of the general performance reduction, CKA remains as the best ANN initialization approach. Moreover, the overall measures are still competitive with the results provided by the CADDementia challenge.

(iv) Figure 4 shows the per-class ROC curves and confusion matrices obtained by the contrasted approaches. In all cases, the area under the curve and accuracy for NC and AD classes are higher than the ones achieved by the MCI class (Figures 4(a)–4(c)). Hence, MCI classification from the incorporated MRI features remains a challenging task due to the following facts: the widely known MCI heterogeneity, the MCI being an intermediate class between healthy individuals and those diagnosed with Alzheimer's disease, and the possibility of MCI subjects eventually converting to AD or NC. Moreover, confusion matrices displayed in Figures 4(d)–4(f) confirm that NC and AD are suitable for distinction in most of the cases. Nevertheless, the MCI class introduces the most errors when considered as both target and output class. Therefore, particular studies on the mild cognitive impairment should improve the diagnosis [5, 40].

5. Conclusion and Future Work

In this paper, we propose a supervised method for initializing the training of artificial neural networks, aiming to improve the computer-aided diagnosis of dementia. Given a set of volume, area, surface, and thickness features extracted from the subject's brain MRI, the examined dementia diagnosis task consists of assigning subjects to the next neurological groups: normal control, mild cognitive impairment (MCI), or Alzheimer's disease. This dementia classification task is particularly challenging because MCI is a heterogeneous and intermediate category between NC and AD. Also, MCI subjects may convert to AD or come back to NC.

To improve the classification performance, we incorporate a matrix projecting the samples into a more discriminating feature space so that the affinity between projected features and class labels is maximized. Such a criterion is implemented by the centered kernel alignment (CKA) between the feature and target label kernels, providing two key benefits: (i) the only free parameter is the hidden dimension; (ii) a relevance analysis can be introduced to find biomarkers. As a result, our proposal of ANN pretraining outperforms the contrasted algorithms (7% of classification accuracy and area under the ROC curve) and reduces the class biasing, resulting in better MCI discrimination.

Nonetheless, the use of CKA implies a couple of restrictions. Firstly, the number of samples should be larger than input and output dimensions to avoid overfitted linear projections. We cope with this drawback by considering a large enough subset of samples for training purposes (about 1300). Secondly, attained projections must always be of lower dimension compared to the original feature space. In this case, the enhancement on class discrimination is due to the affinity between labels and features, not due to an increase of the dimension.

As future work, we plan to evaluate the CKA discriminative capabilities in other neuropathological tasks from MRI as predicting Alzheimer's conversion from MCI and attention deficit hyperactivity disorder classification. We also expect to develop a neural network training scheme using CKA as the cost function.

Appendix

Gradient Descend-Based Optimization of CKA Approach

The explicit objective function of the empirical CKA in (9) yields [32]

$$\widehat{\rho}_{\mathrm{CKA}}\left(\mathbf{K}_{\mathbf{W}}, \mathbf{B}\right) = \log\left(\mathrm{tr}\left(\mathbf{K}_{\mathbf{W}}\mathbf{H}\mathbf{B}\mathbf{H}\right)\right) - \frac{1}{2}\log\left(\mathrm{tr}\left(\mathbf{K}_{\mathbf{W}}\mathbf{H}\mathbf{K}_{\mathbf{W}}\mathbf{H}\right)\right) + \rho_0, \tag{A.1}$$

with $\rho_0 \in \mathbb{R}$ being a constant independent of $\mathbf{W}$. We then consider the gradient descent approach to iteratively solve the optimization problem. To this end, we compute the gradient of the explicit function in (A.1) with respect to $\mathbf{W}$ as

$$\nabla_{\mathbf{W}}\left(\widehat{\rho}_{\mathrm{CKA}}\left(\mathbf{K}_{\mathbf{W}}, \mathbf{B}\right)\right) = -4\mathbf{W}\left(\left(\mathbf{G}\circ\mathbf{K}_{\mathbf{W}}\right) - \mathrm{diag}\left(\mathbf{1}^{\top}\left(\mathbf{G}\circ\mathbf{K}_{\mathbf{W}}\right)\right)\right)\left(\mathbf{X}\mathbf{W}\right)^{\top}, \tag{A.2}$$

where $\mathrm{diag}(\cdot)$ and $\circ$ denote the diagonal operator and the Hadamard product, respectively. $\mathbf{G} \in \mathbb{R}^{N\times N}$ is the gradient of the objective function with respect to the kernel matrix $\mathbf{K}_{\mathbf{W}}$:

$$\mathbf{G} = \nabla_{\mathbf{K}_{\mathbf{W}}}\left(\widehat{\rho}_{\mathrm{CKA}}\left(\mathbf{K}_{\mathbf{A}}, \mathbf{B}\right)\right) = \frac{\mathbf{H}\mathbf{B}\mathbf{H}}{\mathrm{tr}\left(\mathbf{K}_{\mathbf{W}}\mathbf{H}\mathbf{B}\mathbf{H}\right)} - \frac{\mathbf{H}\mathbf{K}_{\mathbf{W}}\mathbf{H}}{\mathrm{tr}\left(\mathbf{K}_{\mathbf{W}}\mathbf{H}\mathbf{K}_{\mathbf{W}}\mathbf{H}\right)}. \tag{A.3}$$

As a result, the updating rule for $\mathbf{W}$, given the initial guess $\mathbf{W}^0$, becomes

$$\mathbf{W}^{t+1} = \mathbf{W}^{t} - \mu_{W}^{t}\nabla_{\mathbf{W}^t}\left(\widehat{\rho}_{\mathrm{CKA}}\left(\mathbf{K}_{\mathbf{W}}, \mathbf{B}\right)\right), \tag{A.4}$$

with $\mu_{W}^{t} \in \mathbb{R}^{+}$ being the step size of the updating rule and $\mathbf{W}^{t}$ being the estimated projection matrix at iteration t.

Competing Interests

The authors declare that there are no competing financial, professional, or personal interests influencing the performance or presentation of the work described in this paper.

Acknowledgments

This work was supported by *Programa Nacional de Formación de Investigadores "Generación del Bicentenario" 2011* and the research Project no. 111956933522, both funded by COLCIENCIAS. Besides, this research would not have been possible without the funding of the E-health project *"Plataforma tecnológica para los servicios de teleasistencia, emergencias médicas, seguimiento y monitoreo permanente de pacientes y apoyo a los programas de prevención"* Eje 3-ARTICA.

References

[1] M. Prince, R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, and C. P. Ferri, "The global prevalence of dementia: a systematic review and metaanalysis," *Alzheimer's & Dementia*, vol. 9, no. 1, pp. 63.e2–75.e2, 2013.

[2] M. Wortmann, "Dementia: a global health priority—highlights from an ADI and World Health Organization report," *Alzheimer's Research & Therapy*, vol. 4, no. 5, article 40, 2012.

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

[4] S. Klöppel, J. Peter, A. Ludl et al., "Applying automated MR-based diagnostic methods to the memory clinic: a prospective study," *Journal of Alzheimer's Disease*, vol. 47, no. 4, pp. 939–954, 2015.

[5] R. Wolz, V. Julkunen, J. Koikkalainen et al., "Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease," *PLoS ONE*, vol. 6, no. 10, Article ID e25446, pp. 1–9, 2011.

[6] B. M. Tijms, A. M. Wink, W. de Haan et al., "Alzheimer's disease: connecting findings from graph theoretical studies of brain networks," *Neurobiology of Aging*, vol. 34, no. 8, pp. 2023–2036, 2013.

[7] S. Lithfous, A. Dufour, and O. Després, "Spatial navigation in normal aging and the prodromal stage of Alzheimer's disease: insights from imaging and behavioral studies," *Ageing Research Reviews*, vol. 12, no. 1, pp. 201–213, 2013.

[8] B. Dubois, H. H. Feldman, C. Jacova et al., "Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria," *The Lancet Neurology*, vol. 13, no. 6, pp. 614–629, 2014.

[9] G. M. McKhann, D. S. Knopman, H. Chertkow et al., "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," *Alzheimer's and Dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[10] C. R. Jack, D. S. Knopman, W. J. Jagust et al., "Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers," *The Lancet Neurology*, vol. 12, no. 2, pp. 207–216, 2013.

[11] G. A. Papakostas, A. Savio, M. Graña, and V. G. Kaburlasos, "A lattice computing approach to Alzheimer's disease computer assisted diagnosis based on MRI data," *Neurocomputing*, vol. 150, pp. 37–42, 2015.

[12] L. Sørensen, A. Pai, C. Igel, and M. Nielsen, "Hippocampal texture predicts conversion from MCI to Alzheimer's disease," *Alzheimer's & Dementia*, vol. 9, no. 4, p. P581, 2013.

[13] S. Klöppel, A. Abdulkadir, C. R. Jack, N. Koutsouleris, J. Mourão-Miranda, and P. Vemuri, "Diagnostic neuroimaging across diseases," *NeuroImage*, vol. 61, no. 2, pp. 457–463, 2012.

[14] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, and J. Tohka, "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects," *NeuroImage*, vol. 104, pp. 398–412, 2015.

[15] S. F. Eskildsen, P. Coupé, V. S. Fonov, J. C. Pruessner, and D. L. Collins, "Structural imaging biomarkers of Alzheimer's disease: predicting disease progression," *Neurobiology of Aging*, vol. 36, supplement 1, pp. S23–S31, 2015.

[16] S. Farhan, M. A. Fahiem, and H. Tauseef, "An ensemble-of-classifiers based approach for early diagnosis of Alzheimer's disease: classification using structural features of brain images," *Computational and Mathematical Methods in Medicine*, vol. 2014, Article ID 862307, 11 pages, 2014.

[17] E. E. Bron, M. Smits, W. M. van der Flier et al., "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge," *NeuroImage*, vol. 111, pp. 562–579, 2015.

[18] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," *Journal of Applied Biomedicine*, vol. 11, no. 2, pp. 47–58, 2013.

[19] D. Chyzhyk, A. Savio, and M. Graña, "Evolutionary ELM wrapper feature selection for Alzheimer's disease CAD on anatomical brain MRI," *Neurocomputing*, vol. 128, pp. 73–80, 2014.

[20] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. 3, pp. 3371–3408, 2010.

[21] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[22] Y. Bengio and P. Lamblin, "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems 19*, pp. 153–160, MIT Press, 2007.

[23] M. Ranzato, C. Poultney, S. Chopra, and Y. L. Cun, "Efficient learning of sparse representations with an energy-based model," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '07)*, pp. 1137–1144, 2007.

[24] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade*, vol. 7700 of *Lecture Notes in Computer Science*, pp. 437–478, Springer, Berlin, Germany, 2012.

[25] J. Weston, F. Ratle, H. Mobahi, and R. Collobert, "Deep learning via semi-supervised embedding," in *Neural Networks: Tricks of the Trade*, G. Montavon, G. B. Orr, and K.-R. Müller, Eds., vol. 7700 of *Lecture Notes in Computer Science*, pp. 639–655, Springer, Berlin, Germany, 2012.

[26] A.-R. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '11)*, pp. 5060–5063, Prague, Czech Republic, May 2011.

[27] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[28] I. A. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," *Journal of Microbiological Methods*, vol. 43, no. 1, pp. 3–31, 2000.

[29] J.-W. Xu, A. R. Paiva, I. Park, and J. C. Principe, "A reproducing kernel Hilbert space framework for information-theoretic learning," *IEEE Transactions on Signal Processing*, vol. 56, no. 12, pp. 5891–5902, 2008.

[30] M. Orbes-Arteaga, D. Cárdenas-Peña, M. A. Álvarez, A. A. Orozco, and G. Castellanos-Dominguez, "Kernel centered alignment supervised metric for multi-atlas segmentation," in *Image Analysis and Processing—ICIAP 2015: 18th International Conference, Genoa, Italy, September 7–11, 2015, Proceedings, Part I*, vol. 9279 of *Lecture Notes in Computer Science*, pp. 658–667, Springer, Berlin, Germany, 2015.

[31] C. Cortes, M. Mohri, and A. Rostamizadeh, "Algorithms for learning kernels based on centered alignment," *Journal of Machine Learning Research*, vol. 13, pp. 795–828, 2012.

[32] A. J. Brockmeier, J. S. Choi, E. G. Kriminger, J. T. Francis, and J. C. Principe, "Neural decoding with kernel-based metric learning," *Neural Computation*, vol. 26, no. 6, pp. 1080–1107, 2014.

[33] B. Fischl, "FreeSurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, 2012.

[34] X. Han, J. Jovicich, D. Salat et al., "Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer," *NeuroImage*, vol. 32, no. 1, pp. 180–194, 2006.

[35] F. Ségonne, A. M. Dale, E. Busa et al., "A hybrid approach to the skull stripping problem in MRI," *NeuroImage*, vol. 22, no. 3, pp. 1060–1075, 2004.

[36] J. G. Sied, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in mri data," *IEEE Transactions on Medical Imaging*, vol. 17, no. 1, pp. 87–97, 1998.

[37] F. Ségonne, J. Pacheco, and B. Fischl, "Geometrically accurate topology-correction of cortical surfaces using nonseparating loops," *IEEE Transactions on Medical Imaging*, vol. 26, no. 4, pp. 518–529, 2007.

[38] B. Fischl, A. van der Kouwe, C. Destrieux et al., "Automatically parcellating the human cerebral cortex," *Cerebral Cortex*, vol. 14, no. 1, pp. 11–22, 2004.

[39] R. L. Buckner, D. Head, J. Parker et al., "A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume," *NeuroImage*, vol. 23, no. 2, pp. 724–738, 2004.

[40] J. Ramírez, J. M. Górriz, A. Ortiz, P. Padilla, and F. J. Martínez-Murcia, "Ensemble tree learning techniques for magnetic resonance image analysis," in *Innovation in Medicine and Healthcare 2015*, vol. 45 of *Smart Innovation, Systems and Technologies*, pp. 395–404, Springer, Berlin, Germany, 2016.