An intelligent Alzheimer's disease diagnosis method using unsupervised feature learning

Firouzeh Razavi[1], Mohammad Jafar Tarokh[2*] and Mahmood Alborzi[1]

*Correspondence:
mjtarokh@kntu.ac.ir
[2] Department of Industrial
Engineering, K.N. Toosi
University of Technology,
Tehran, Iran
Full list of author information
is available at the end of the
article

Abstract

Today, the diagnosis of Alzheimer's disease (AD) or mild cognitive impairment (MCI) has attracted the attention of researchers in this field owing to the increase in the occurrence of the diseases and the need for early diagnosis. Unfortunately, the nature of high dimension of neural data and few available samples led to the creation of a precise computer diagnostic system. Machine learning techniques, especially deep learning, have been considered as a useful tool in this field. Inspired by the concept of unsupervised feature learning that uses artificial intelligence to learn features from raw data, a two-stage method was presented for an intelligent diagnosis of Alzheimer's disease. At the first stage of learning, scattered filtering, an uncontrolled two-layer neural network was used to directly learn features from raw data. At the second stage, SoftMax regression was used to categorize health statuses based on the learned features. The proposed method was validated by the data sets of Alzheimer's Brain Images. The results showed that the proposed method achieved very good diagnostic accuracy and was better than the existing methods for brain image data sets. The proposed method reduces the need for human work and makes it easy to intelligently diagnose for big data processing, because the learning features are adaptive. In our experiments with the Alzheimer's Disease Neuroimaging Initiative (ADNI) data, a dual and multi-class classification was conducted for AD/MCI diagnosis and the superiority of the proposed method in comparison with the advanced methods was shown.

Keywords: Alzheimer's disease, Sparse filtering, Unsupervised feature learning, Intelligent diagnosis, SoftMax regression

Introduction

In computer systems, especially with the advancement of the Internet and databases, big data is increasingly expanding and is advancing exponentially [1–4]. This is mostly true in medical big data and images. Therefore, the issue of exploding data shows the concept and power of the big data. In the field of medicine, especially magnetic resonance imaging (MRI) images, the issue of big data with high data dimensions is investigated [5].

As people grow older in the community, an untreated disease would be common, which is called Alzheimer's, and it has been proven that it has no treatment, but it can be prevented from development with timely diagnosis. Alzheimer is known as the most common disease among the various causes of dementia and with each passing decade, the number of people infected with the disease is almost doubled. For this reason, timely

diagnosis of this disease for appropriate and quick treatment is important to prevent or at least delay its progress [6].

Early diagnosis of Alzheimer's disease requires the accurate information from patients such as a history of disease, neurological tests, etc. Given the limited time to treat this disease; early diagnosis is critical to prevent its progression. For this purpose, researchers have made many efforts to provide a system that can discover the mechanism and cause of the disease and prevent its development as far as possible. The analysis of various neurological images such as MRI, positron emission tomography (PET), geometric sensitivity gating (GSG), etc., is needed to diagnose Alzheimer's disease [7, 8].

The major challenge in this discussion is the high dimension with the small number of samples in the analysis of brain images. Therefore, machine learning and in their outline, deep learning, has achieved a lot of success. Machine learning methods can overcome these issues (large p, small n). Lately, deep learning in the field of computer, particularly in the analysis of brain images [9], computer vision [10], natural language processing [10], and speech recognition [11] has had a great efficiency as compared to other methods, thus it is considered as a main and powerful tool to analyze the brain images. Generally, the existing methods in these techniques are divided into two categories of dimension reduction and feature selection. Many methods have been proposed to reduce the dimension. These methods by reducing the dimension of the data result in the loss of many features with information load. While, the feature selection methods find features from features with information load and without information load in the main data space. This study was conducted to work on neurological image data from the focus on feature selection techniques.

In this study, the proposed method was considered in two stages. In the first stage, sparse filtering learning, which is like a two-layer network was used, to learn the features that characterize Alzheimer's disease. This algorithm can reduce the effect of features without information load and with low information load in feature selection. To remove features without information load, the sparse filtering is implemented into a global mode. Then, in the second stage of learning, SoftMax regression method was trained for automatic classification.

The main structure of the paper is as follows.

Related work

By reviewing the literatures on diagnosis of Alzheimer's disease, it was concluded that brain data and images with a low sample size and high dimension are one of the most important challenges in this study, and new research can be carried out in this area [12–15]. Most of the recently used methods are deep learning methods, including deep sparse multi-task learning [16], stacked auto-encoder [17], sparse regression models [18], etc., each attempting to overcome the aforementioned challenges. These methods are used more to select features. In the features of this issue, there are features with information load and without information load that should be selected, which can improve the classification accuracy of the disease. In [17], a deep architecture for the removal of the features without information load has been recursively proposed by implementing sparse multi-task learning in a hierarchy. The optimal regression coefficients were assumed to reflect the relative importance of the features

in representing the target response variables. Also, combined two different conceptual approaches of sparse regression and deep learning to diagnose and predict Alzheimer's disease. Especially, first sparse regression models train some models [18]. Each of these models was trained with different values from an adjustment control parameter. Therefore, multiple sparse regression models potentially select the subset of features from the sum of the main features. Therefore, they have different powers to predict the response values. Another literature [19], presented a method for representing a linear feature based on SAE's deep learning. The combination of hidden information with the main feature in this article has helped to create an improved model for Alzheimer's disease/mild cognitive impairment (AD/MCI) classification with high diagnosis accuracy. In addition, thanks to the unsupervised feature and the pre-training of deep learning, they have been able to use non-related samples with the goal of initializing SAE.
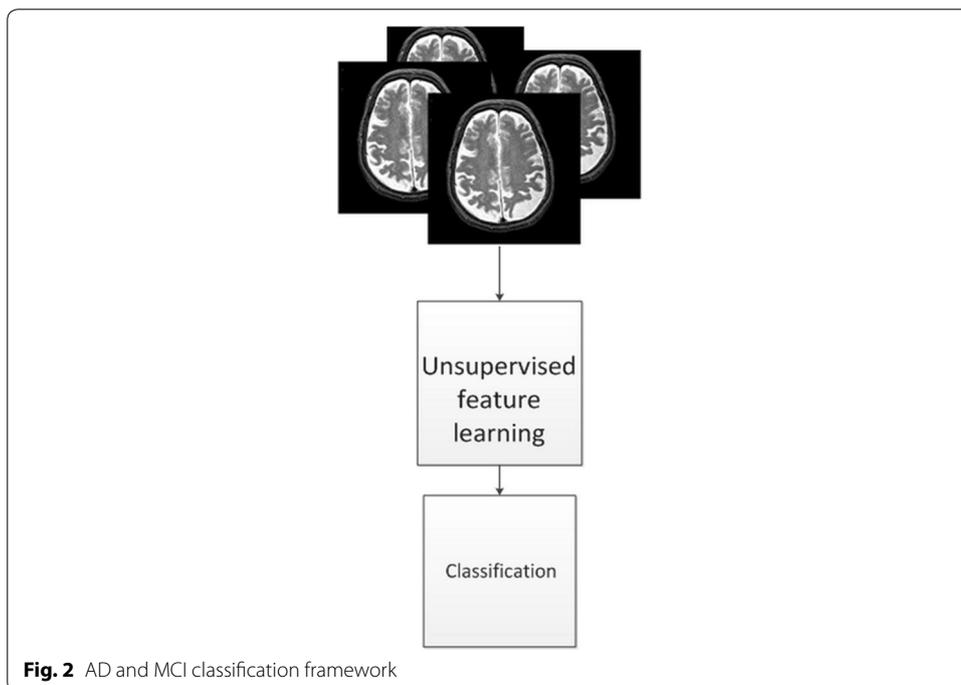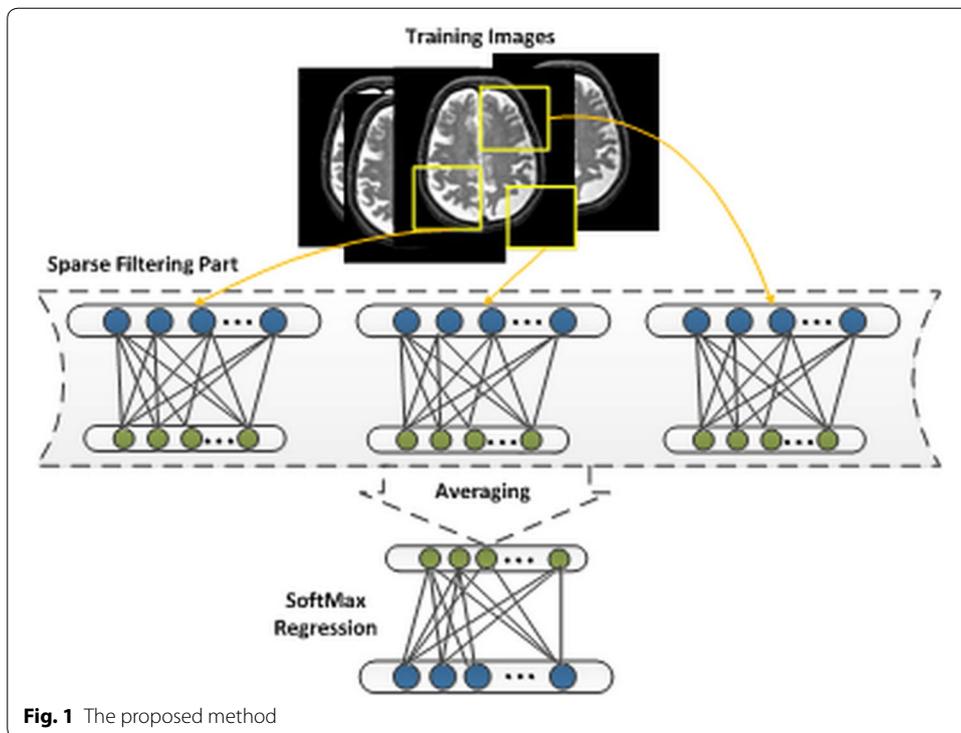
There are other methods, which are based on deep learning that uses the convolutional neural network. In [20], a three-dimensional convolutional neural network was used to predict Alzheimer's disease that can learn general features by taking biological signs and adapt it to the data sets from different domains. Three-dimensional convolutional neural network (3D-CNN) is built from a 3D convolutional auto-encoder that has already been trained to get anatomical changes in brain structure of MRI scans. Hosseini-Asl et al. [21] used the convolutional neural network to build a system for diagnosing Alzheimer's disease in 2016.

Another method that focuses on Alzheimer's disease, and its diagnosis are manifold-based learning method. In [22], manifold-based learning method was used to classify Alzheimer's disease. It has been assumed that manifold space is linear and needs to define the similarity of measurement or the approximation of the graph. In [22], manifold-based learning was presented based on deep belief network (DBN) to discover similarities in a group of DBN architecture images together with several layers of restricted Boltzmann.

This paper is organized as follows: In "Related work" section, a method design that includes sparse filtering and SoftMax regression is briefly described. "Methods" section describes the two-stage learning method. In "Proposed method" and "Data set" sections, Alzheimer's disease data set classification was studied using the proposed method. Finally, "Results and discussion" section concludes the study.

Methods

The proposed method is as shown in Fig. 1. A two-stage learning approach was presented. In the first stage of learning, sparse filtering, which is seen as a two-layer network, was used to learn the expressive features of brain images. Then, in the second stage of learning, SoftMax regression, which is also a two-layer network, was trained to automatically categorize the conditions of healthy and unhealthy individuals. Due to the use of the neural network to learn the features, the proposed method is not dependent on prior knowledge and human work and is much more suitable for processing large signals in the field of monitoring the conditions. Here, the classification of Alzheimer's

Fig. 1 The proposed method

Fig. 2 AD and MCI classification framework

disease from brain images was studied, and the results showed the superiority of the proposed method. AD and MCI classification framework is as shown in Fig. 2.

Unsupervised learning

Unsupervised feature learning techniques, such as scattered self-encryptions, limited Boltzmann machines, dispersed coding, etc., are extensively trying to model an approximate fair distribution of collected data. They frequently require tuning different parameters for a platform, which is a major challenge. For example, in a limited Boltzmann machine, the parameters must be set, such as the number of features, weight degradation, dispersion compensation, rate of learning, and amount of movement. When these parameters are adjusted inappropriately, the acquired features may lead to low diagnosis accuracy. Hence, Nigam et al. [18] presented an unsupervised feature learning technique, which is sparse filtering. The main key to the sparse filtering is that the only parameter required is the number of features. Hence, sparse filtering necessarily involves tuning the parameters. Sparse filtering, as an uncontrolled two-layer network, presents dispersed distribution of the features computed by the data collected rather than modeling the distribution of data.

Assuming the training set of $\{x^i\}^M i = 1$, where $x^i \in R^{N \times 1}$ are a sample and M is the number of samples, the sparse filtering of samples is mapped to the features of $f^i \in R^{L \times 1}$ using the matrix of $W \in R^{N \times L}$. Here is a situation where sparse filtering calculates the linear feature of each sample.

$$f_l^i = W_l^T x^i \tag{1}$$

where, f_l^i corresponds to the *lth* feature of the *ith* sample. The sparse filtering optimizes a function using the normalized feature of l2. It should be noted that the lp norm of t is formulated as $||t||_p = \sqrt[p]{|t_1|^p + \cdots + |t_n|^p}$ where, $t = [t_1, t_2 \ldots t_n]$.

The f_l^i makes the feature matrix. First, each of the rows of the feature matrix is normalized with the l2 norm in all the samples.

$$\tilde{f}_l = f_l / \left\| f_l \right\|_2 \tag{2}$$

Then, each column is normalized with the l2 norm, so that the features are placed on the l2-ball unit.

$$\hat{\tilde{f}}^i = \tilde{f}^i / \left\| \tilde{f}^i \right\|_2 \tag{3}$$

Finally, the weight matrix in Eq. 1 is solved with the optimization of the l1 norm of cost function restriction for each sample, which is as follows:

$$\min_w \sum_{i=1}^M \left\| \hat{\tilde{f}}^i \right\|_1 \tag{4}$$

The l2 norm is usually used to measure the dispersion of components and the dispersion means that most of the components are zero. The $\left\| \hat{f}^i \right\|_1$ word in Eq. (4) measures the features of the *ith* sample. As the normalized feature of *l*2 of $\hat{f}^i$ is limited to the *l*2-ball unit, the cost function in Eq. (4) is minimized when these features are dispersed. Sparse filtering can be converted to a nonlinear filtering using an activation function and the activation function of $g(.) = |.|$ is recommended. Therefore, Eq. (1) is expanded as:

$$f_l^i = g\left(W_l^T x^i \right) \tag{5}$$

By optimizing the cost function of the features in Eq. 5, the trained features can be used to find nonlinear information from the input samples and have good unreliability. More details of the sparse filtering are described in [23].

SoftMax regression

In neural networks, Softmax regression is often implemented in the final layer for multi-class classification [19]. Its implementation is easy and quickly calculated. Suppose we have a training set of $\{x^i\}^M i = 1$ along with a labeled set of $\{y^i\}^M i = 1$, where, $x^i \in R^{N \times 1}$ and $y_i \in \{1, 2, \ldots K\}$. For each input sample x_i, the model tries to estimate the probability of $p(y_i = k|x_i)$ for each of the labels of $k = 1, 2, \ldots, K$.

Hence, the assumption of the SoftMax regression produces a vector that gives the K-estimated probability of the input sample of xi belonging to each label. The assumption of h_θ is as follows:

$$
h_\theta(x_i) = \begin{bmatrix} p(y_i = 1|x_i; \theta) \\ p(y_i = 2|x_i; \theta) \\ \ldots \\ p(y_i = k|x_i; \theta) \end{bmatrix} \frac{1}{\sum_{k=1}^{K} e^{\theta_k^T x^i}} \begin{bmatrix} e^{\theta_1^T x^i} \\ e^{\theta_k^T x^i} \\ \ldots \\ e^{\theta_k^T x^i} \end{bmatrix}
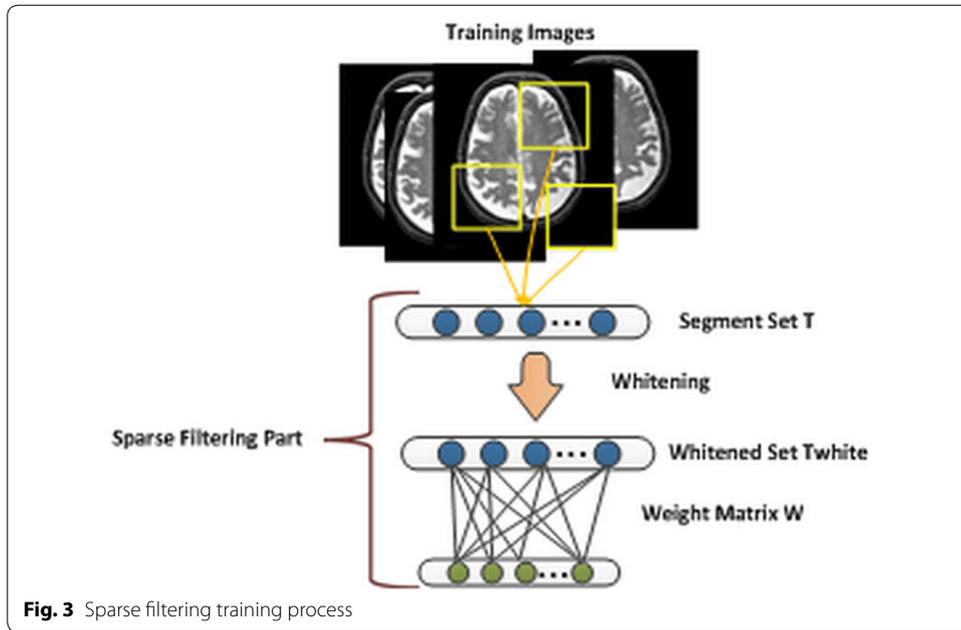\tag{6}
$$

where, $\theta = [\theta_1, \theta_2, \ldots, \theta_n]^T$ is the parameter of the Softmax regression model. It should be noted that $\sum_k = 1^K e^{\theta_k^T x^i}$ normalizes the distribution, so that the sum of the components of the assumption equals 1. According to the assumption, the model is done by minimizing the cost function of $J(\theta)$.

$$
J(\theta) = -\frac{1}{M} \left[\sum_{i=1}^{M} \sum_{k=1}^{K} 1\left\{y^i = k\right\} \log \frac{e^{\theta_k^T x^i}}{\sum_{k=1}^{K} e^{\theta_k^T x^i}} \right] + \sum_{k=1}^{K} \sum_{l=1}^{l} \theta_{kl}^2
\tag{7}
$$

where, 1{.} represents the function of the marker which, if it is true, it returns the condition 1; otherwise, λ is the term of weight loss. The term of weight loss forces some Softmax regression parameters to adopt values close to zero, while allowing other parameters to maintain their relatively large values, thereby improves generalize. With the term of weight loss (for every $\lambda > 0$), the cost function of $J(\theta)$ strictly become convex, and the Softmax regression model ensures that theoretically has a unique solution. In addition, Softmax regression is a special solution to the issue of classification, which assumes that the linear combination of features of sample can be used to determine the probability that a sample will belong to any health status label. For example, Softmax regression provides a possible classification.

Proposed method

This section describes the two-stage learning method presented to diagnose Alzheimer's disease. At the first stage of learning, sparse filtering was used to extract the distinct features of the crude brain images, and the learned features of the pixels were obtained by averaging these local features [23]. At the second stage, SoftMax regression was applied to classify the health status of individuals using the learned features. Brain images have been obtained under various health conditions [24]. These images have constructed a

Fig. 3 Sparse filtering training process

training set of $\{x_i, y_i\}^M i = 1$ where, $x^i \in R^{N \times 1}$ of the *ith* sample consists of N points of data, and y_i is the label of health status.

The first learning stage has three stages as shown in Fig. 3. Firstly, the sparse filtering was trained and the weight matrix of W was obtained. Then, the learned sparse filtering was used to get the local features from each sample. Finally, these local features were averaged to obtain the features learned from each sample.
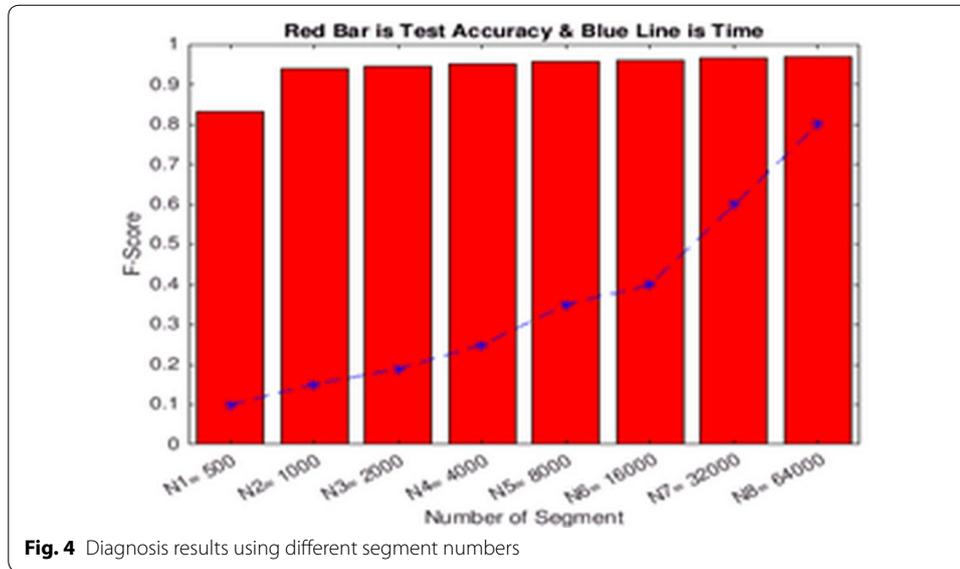
Suppose that the input dimension of the sparse filtering is N_{in}, and the output dimension of the sparse filtering is N_{out}. When we are training the sparse filtering model, we randomly sample N_s pieces from the training samples. This means that random segments are obtained by the overlapping method. As shown in Fig. 4, these segments are composed of an unsupervised training set of $\{s^j\}^{Ns} j = 1$, where, $s^j \in R^{N_{in} \times 1}$ is the *jth* segment containing N_{in} points of the data. The set of $\{s^j\}^{Ns} j = 1$ is rewritten as a matrix of $S \in R^{N_{in} \times N_s}$ and preprocessed by whitening. The goal of whitening is to reduce the correlation between the segments and accelerate the convergence of sparse filtering training. Whitening uses a special amount of covariance matrix.

$$cov(S) = EDE^T \tag{8}$$

where, E is the orthogonal matrix of the special vector of the covariance matrix ($cov(S)$), and D is a special vector diagonal matrix. Thus, the set of the whitening training segment (T_{white}) can be obtained as follows:

$$T_{white} = ED^{-1/2}E^T S \tag{9}$$

The sparse filtering model is trained using T_{white} and the trained weight matrix of W is used to compute the local features of the training samples. Alternately, the training samples are divided into J segments, where, J is an integer equal to N/N_{in}, for example, x^i is

Fig. 4 Diagnosis results using different segment numbers

divided into the set of segments of $\{x_{ij}\}^J j = 1$, where, $x_j^i \in R^{N_{in} \times 1}$. For each x_j^i, the local feature of $f_j^i \in R^{1 \times N_{out}}$ can be obtained by the trained sparse filtering. Beforehand, the learned features of f^i of x^i with combining these local features of f_j^i used to be obtained by the sum method. In other words, local features are connected to a feature vector as the learned features

$$f^i = [f_1^i, f_1^i, \ldots, f_1^i]^T. \tag{10}$$

In this paper, the middle method is used instead of the mass method, and the learned features are obtained as follows

$$f^i = \left(\frac{1}{J} \sum_{j=1}^{J} f_j^i \right)^T. \tag{11}$$

Averaging is used due to the fact that the distinctive features that segments are shared with each other improves and prevents accidental effects caused by noises. As soon as the learned set of $\{f^j\}^M j = 1$ was obtained, it was trained by the label set of $\{y^j\}^M j = 1$ for training the Softmax regression. The Softmax regression model computes the probability that the fi feature has the health status labels of yi or not, as shown in Eq. (6). The sum of probabilities for all class labels of 1 ensures that the right (equation) in Eq. (6) determines the normalized distribution. After being trained, the former probability in $h_\theta(xi)$ determines which health condition labels feature belonging to it.

Data set

The data set used in this paper is from the ADNI[1] standard data sets that are considered in experiments with the cerebrospinal fluid (SCF) and MRI data sets [25]. Meanwhile, the number of AD patients was 51 and the number of MCI patients was 99, and it should be noted that 43 of these patients had the capability to suffer from AD and 56 of these

---

[1] https://adni.loni.ucla.edu.

Razavi *et al. J Big Data*   (2019) 6:32

Page 9 of 16

Table 1 Clinical information of the images

Name	AD N = 51	MCI Converter (N = 43)	MCI Non-converter (N = 56)	HC (N = 52)
Female/male	18/33	15/39	17/39	18/34
Age (mean ± SD)	75.2 ± 7.4 [59–88]	75.7 ± 6.9 [58–88]	75.0 ± 7.1 [55–89]	75.3 ± 5.2 [62–85]
Education (mean ± SD)	14.7 ± 3.6 [4–20]	15.4 ± 2.7 [10–20]	14.9 ± 3.3 [8–20]	15.8 ± 3.2 [8–20]
MMSE (mean ± SD)	23.8 ± 2.0 [20–26]	26.9 ± 2.7 [20–30]	27.0 ± 3.2 [18–30]	29 ± 1.2 [25–30]
CDR (mean ± SD)	0.7 ± 0.3 [0.5–1]	0.5 ± 0.0 [0.5–0.5]	0.5 ± 0.0 [0.5–0.5]	0.0 ± 0.0 [0–0]

Table 2 Clinical criteria for patients

Criteria	Normal-person	MCI-person	AD-person
MMSE score	24–30	24–30	20–26
CDR score	0	0	0.5–1.0
Depressed	No	–	–
MCI	No	Yes	Yes
Dementia	No	Yes	Yes
Memory complaint	–	Yes	–

patients did not have the capability to suffer from the disease. The rest of the population was HC.

In this work, the ADNI data set available on the site was used. In particular, only MRI, FDG-PET, and CSF data sets were considered, which included 51 patients with AD and 99 patients with MCI (43 of MCI patients had the capability to suffer from AD and 56 of these patients did not have the capability to suffer from the disease). The rest of the population (52 people) was HC. The detailed information associated with this data is shown in Table 1, so that the results are visible on the full available information.

In this work, two clinical scores were considered for the MMSE and ADAS-Cog data. The criteria for a healthy and unhealthy person are shown in Table 2.

According to the data available on the website linked to the ADNI data set, all MRI data were obtained using a 1.5T scanners. The MRI data from the ADNI website has been downloaded in NITI format. The FDC-PET images were obtained 30 to 60 minimum after injection of the ampoule. The MRI images were pre-processed using the conventional AC-PC correction, skull cleavage and cerebellar removal methods. Typically, the MIPAV software was used to correct AC-PC. Images were resized to $256 \times 256 \times 256$. A precision and increase in skull cleavage were performed, and the cerebellum was then removed. Then, FAST of the FSL package was used to segregate the structure of the MR image into three tissues of gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). Finally, they were applied to 93 areas with Kabani wrapping. In this paper, GM was used to classify AD-MCI, because GM has a high relative relationship with AD-MCI as compared to the other. For each ROI, the GM tissue volume of the MRI was

used, and the average intensity of FDG-PET is as a feature that is widely used in AD diagnosis.

Results and discussion

Firstly, the parameters in this problem should be considered, which are the parameters of the selected method, and they include the N_{in} input dimension and the N_{out} output dimension of sparse filtering, and the term "weight loss" of λ of SoftMax regression. Initially, the N_{in} input dimension of sparse filtering was examined. Randomly, 10% of the training samples of the proposed method were selected, in which 20,000 segments of the learning samples were sampled to train the sparse filtering at the first stage and the remaining samples were used for testing. The output dimension is equal to $N_{out} = N_{in}/2$ and the term weight loss is equal to $\lambda = 1E-5$.

In this section, the effect of the proposed method was investigated for extracting the feature in the deep learning using the spars filtering technique by considering the four binary categories: (AD vs. HC), (MCI vs. HC), (AD vs. MCI), and (MCI-C vs. MCI-NC).

For each classification issue, the data sets were randomly divided into 10 subsets, each containing 10% of the data sets, and then; 9 of the 10 subsets were used for training and one for the test.

This process was repeated 10 times for evaluation are averaged the training and testing accuracy seen over 98.7%, 97.9%. Indeed, proposed method could classify data set of Alzheimer's disease with high accuracies that applied various input dimensions of sparse filtering. However, the input dimension which is larger causing to spend the more time for the method.

To show the validity of our proposed method, the results of the proposed method were compared for those that had low-level capabilities, using similar strategies to feature selection and SoftMax regression training. It is important to note that the same sample of training and testing was used to compare the existing methods.

Table 3 shows the optimal performance of sparse filtering. According to the structure of a sparse filtering model, three hidden layers were considered for MRI, FDG-PET, and CAT11, and two hidden layers were considered for CSF, taking into account the dimensions of low-level features in each method. To determine the number of hidden units, classification of sparse filtering was implemented with the search for network. Due to the likelihood of more connections with a few training samples, the first set up was repeated with a few target samples. For example, in the AD and HC classification, the best accuracy was obtained from 87.7% with MRI in the SAE classification of 500-50-10 (from bottom to top) in the supervised hidden units.

The effect of the number of data for training

In this section, the effect of the number of training data was studied; that is, the percentage of the samples was examined to train the proposed method and the number of sampled segments from the training samples for the sparse filtering that is being trained. Figure 4 shows the accuracy of the diagnosis using the number of different time parts when 10% of the samples are randomly selected for the training. It can be observed that when the number of the segments is high, the accuracy of testing is high and the standard deviation is smaller. It should be taken into consideration that the data segments are

Table 3 Classification accuracy derived from the SP classification and the relevant structure as the number of hidden units

Models	Hidden units	Accuracy (ACC)
AD vs. HC		
MRI	500-50-10	0.876 ± 0.017
PET	1000-50-30	0.879 ± 0.023
CSF	50-3	0.845 ± 0.014
CONTACT	500-100-20	0.902 ± 0.019
MCI vs. HC		
MRI	100-100-20	4 ± 0.018
PET	300-50-10	0.694 ± 0.014
CSF	10-3	0.704 ± 0.019
CONTACT	100-50-20	0.752 ± . 0.017
AD vs. MCI		
MRI	1000-50-30	0.662 ± 0.021
PET	100-50-10	0.678 ± 0.020
CSF	10-1	0.687 ± 0.011
CONTACT	100-100-20	0.712 ± 0.019
MCI-C vs. MCI-NC		
MRI	100-100-10	0.561 ± 0.030
PET	100-100-10	0.614 ± 0.028
CSF	30-2	0.601 ± 0.021
CONTACT	500-50-20	0.625 ± 0.025

unlabeled and are much easier to obtain than the labeled data, which means that the proposed method can use the advantage of unsupervised learning and improve the accuracy of its diagnosis. However, the meantime increases exponentially by increasing the number of the segments. To create a balance between the elapsed time and the accuracy of the diagnosis, 20,000 segments for the data sets were used.
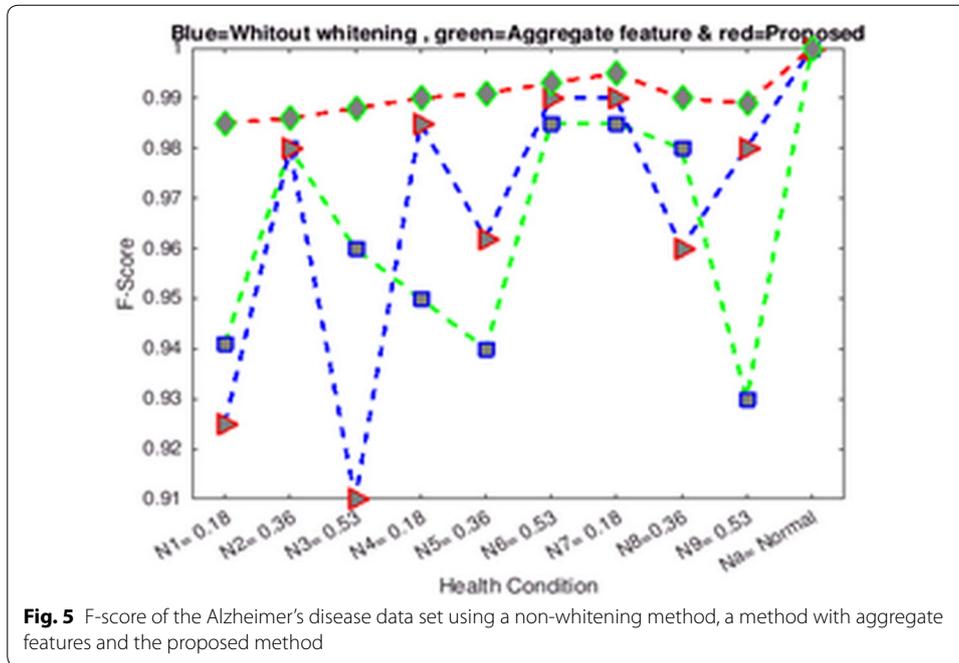
Necessity of whitening

At the two-stage learning method, it was observed that whitening has been used in the sparse filtering training process and the features learned by local features instead of collecting were averaged. Here, the necessity of whitening is studied.

The Alzheimer's disease data set was calculated using a non-whitening method, and a method called aggregate features using Eq. 10.

Performance comparison

To compare performance with the proposed method, F-score was used, which is the criterion for measuring the efficiency of the classification method that is commonly used, F-score reaches its best value at 1 and worst at 0, besides This method takes the sensitivity, and sensitivity results that in Eqs. (12) to (15), represent the classification accuracy (ACC), sensitivity (SEN) and specificity (SPE), F-score is selected as evaluation indices.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

Fig. 5 F-score of the Alzheimer's disease data set using a non-whitening method, a method with aggregate features and the proposed method

$$Recall, sensitivity = \frac{TP}{TP + FN} \tag{13}$$

$$specificity = \frac{TN}{TN + FP} \tag{14}$$

$$F - score = \frac{2TP}{2TP + FP + FN} \tag{15}$$

In Fig. 5, the F-score of the Alzheimer's disease data set is shown by a method, in which each F-score is averaged by 20 tests. It can be observed that F-score uses the range of the proposed method in 0.9993 to 1, while F-score without whitening range is in 0.907 to one range and the use of a method over the range of aggregated features is in 0.929 to 1. For most health conditions, the proposed method obtains a higher F-score than the other two methods, which means that the proposed method works better. Therefore, both whitening and averaging feature are necessary for the proposed method.

The results of classification

According to the selection of the parameter, it was found that this method based on classification showed a better performance than the Lasso-based group [26]. The acronyms LLF and SPF were used for low-level learning features and sparse filtering features, respectively.

Table 4 shows the average accuracy of the methods in the AD and HC classification. Although, the proposed method of LLF + SPF has a better performance than the LLF-based method, for example, 89% (LLF) versus 88.2% (LLF + SPF) with MRI, 93.7% (LLF) versus 93.5% (LLF + SPF) with CONCAT, in general, they showed the

Table 4 The average accuracy of the methods in the AD and HC classification

Images	Ref. [19]	Proposed method
MRI	0.882 ± 0.019	0.901 ± 0.020
PET	0.850 ± 0.018	0.872 ± 0.023
CSF	0.801 ± 0.012	0.819 ± 0.021
CONCAT	0.935 ± 0.012	0.953 ± 0.017
MK-SVM	0.979 ± 0.007	0.983 ± 0.009

Table 5 The average accuracy of the methods in the MCI and HC classification

Images	Ref. [19]	Proposed method
MRI	0.80 ± 0.016	0.901 ± 0.020
PET	0.745 ± 0.018	0.872 ± 0.023
CSF	0.679 ± 0.022	0.819 ± 0.021
CONCAT	0.836 ± 0.005	0.857 ± 0.009
MK-SVM	0.888 ± 0.012	0.912 ± 0.008

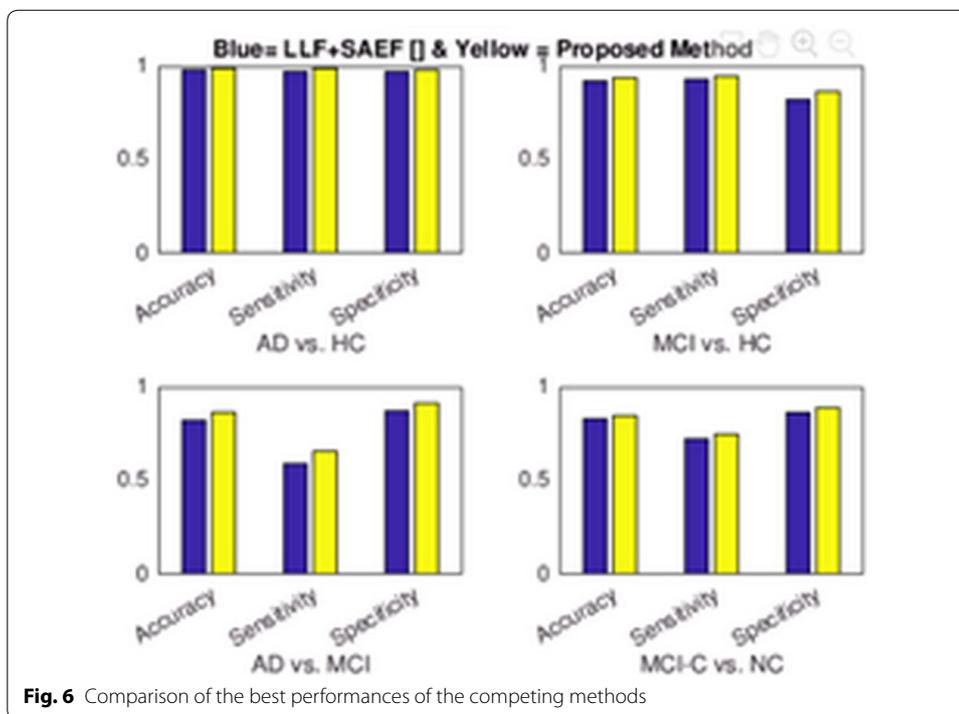**Table 6 The average accuracy of the methods in the AD and MCI classification**

Images	Ref. [19]	Proposed method
MRI	0.704 ± 0.026	0.901 ± 0.020
PET	0.711 ± 0.025	0.872 ± 0.023
CSF	0.655 ± 0.009	0.819 ± 0.021
CONCAT	0.752 ± 0.030	0.655 ± 0.009
MK-SVM	0.827 ± 0.025	0.843 ± 0.021

best accuracy up to 98.3% by some methods using the SoftMax regression. As compared to the accuracy of 97% in the LLF-A-based method in [13], the accuracy of the proposed method improved up to 1.3%.

In the MCI and HC classifications, shown in Table 5, the proposed method for classification showed the best accuracy from 91.2. Performance was improved by 2.3% when compared with the classification accuracy of 88.8% based on the LLF-4 method in [13].

In the AD and MCI classifications, shown in Table 6, the proposed method has the best classification accuracy of 84.3. The classification accuracy by 2.3% can be slightly increased, as compared to the LLF-based method [13], which increases accuracy by 82.7%.

Regardless of the model training program, for four binary classification problems with their accuracy, sensitivity, and specificity are as shown in Fig. 6. As shown in this figure, the proposed method is better than the competitive ones. It is noteworthy to note that there is a tendency to increase AD vs. HC, AD vs. MCI, MCI vs. HC, and MCI-C vs. MCI-NC.

Fig. 6 Comparison of the best performances of the competing methods

Comparing with the existing methods in literatures

In addition, performance of the proposed method was compared with a specific latent feature representation [13]. To make an acceptable comparison, the same training and sample test was used for M3T. As compared to the accuracy of M3T, which was 94.5 ± 0.8 and 84 ± 1.1, the accuracies were 78.8 ± 1.8 and 71.8 ± 2.6 for AD vs. HC, AD vs. MCI, MCI vs. HC, and MCI-C vs. MCI-NC, respectively. The performance improvement of 3.4, 4.8 and 3.9 was obtained for the proposed method with LLF + SPF, and our method achieve better accuracy, sensitivity, specificity in most scenarios.

Conclusion and future work

The main purpose of our work is that there may be the inherent, hidden high-level information in the main low-level features, such as relations of intermediate features, which can be useful for a stronger diagnosis. To this end, in this paper, we suggested the use of deep learning with SP for a hidden feature representing data for the diagnosis of AD/MCI. While, SP is a neural network in the structural model, thanks to the two-stage training program before training and setting for deep learning, we can reduce the risk of falling into local optima, which is the main limitation of the typical neural network. We believe that deep learning can be a new way to analyze the imaging data, and we presented the application of this method for diagnosing Alzheimer's disease for the first time. Through a case study of Alzheimer's disease Images, it has been suggested that this proposed method adaptively learns the features of raw signals for different diagnostic problems and is superior to the methods available in diagnosing Alzheimer's images. The proposed method is able to utilize the advantages of unsupervised learning and

improve the accuracy of its diagnosis, along with an increase in the number of unlabeled data. In future work, the features of neural network weights are studied in depth through the unsupervised feature learning, so that we can fill the gap between manual extraction by using signal processing and feature learning using artificial intelligence techniques. In addition, the application of neural networks in the field of control of diagnostic systems is an attractive subject; therefore, the study of the use of uncontrolled neural network in this area is an attractive subject for future research.

Abbreviations
MRI: magnetic resonance imaging; PET: positron emission tomography; GSG: geometric sensitivity gating; AD/MCI: Alzheimer's disease/mild cognitive impairment; ADNI: Alzheimer's Disease Neuroimaging Initiative; SP: sparse filtering; 3D-CNN: three-dimensional conventional neural network; DBN: deep belief network; LLF: low-level learning; SPF: sparse filtering features; M3T: multi-modal multi-task learning.

Authors' contributions
FR has contributed for acquisition of data, analysis and interpretation of data, drafting of the manuscript. MJT has served as advisor in study conception, and for critical revision. MA has critically reviewed the study proposal and for design. All authors read and approved the final manuscript.

Author details
[1] Department of Information Technology Management, Islamic Azad University, Science and Research Branch of Tehran, Tehran, Iran. [2] Department of Industrial Engineering, K.N. Toosi University of Technology, Tehran, Iran.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
1. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics. 2015;8(1):33.
2. Chen M, Mao S, Liu Y. Big data: a survey. Mob Netw Appl. 2014;19(2):171–209.
3. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. Biomed Inform Insights. 2016;8:1.
4. Siuly S, Zhang Y. Medical big data: neurological diseases diagnosis through medical data analysis. Data Sci Eng. 2016;1(2):54–64.
5. Poldrack RA, Gorgolewski KJ. Making big data open: data sharing in neuroimaging. Nat Neurosci. 2014;17(11):1510–7.
6. Glenner GG. Alzheimer's disease. In: Biomedical advances in aging. Springer. 1990:51–62.
7. Baum LW, Chow HLA, Cheng KK. Nanoparticle contrast agent for early diagnosis of alzheimer's disease by magnetic resonance imaging (mri). ed: Google Patents. 2016.
8. Sabri O, et al. Florbetaben PET imaging to detect amyloid beta plaques in Alzheimer's disease: phase 3 study. Alzheimer's Dement. 2015;11(8):964–74.
9. Li R et al. Deep learning based imaging data completion for improved brain disease diagnosis. In: International conference on medical image computing and computer-assisted intervention. Springer; 2014. pp. 305–12.
10. Socher R. Recursive deep learning for natural language processing and computer vision. Citeseer. 2014.
11. Yu D, Deng L. Automatic speech recognition: a deep learning approach. Berlin: Springer; 2014.
12. Bhatkoti P, Paul M. Early diagnosis of Alzheimer's disease: a multi-class deep learning framework with modified k-sparse autoencoder classification. In: Image and vision computing New Zealand (IVCNZ), 2016 international conference on, IEEE. 2016. pp. 1–5.

13. Hu C, Ju R, Shen Y, Zhou P, Li Q. Clinical decision support for Alzheimer's disease based on deep learning and brain network. In: Communications (ICC), 2016 IEEE international conference on, IEEE. 2016. pp. 1–6.
14. Sarraf S, Tofighi G. Classification of Alzheimer's disease using fmri data and deep learning convolutional neural networks. arXiv preprint arXiv:1603.08631. 2016.
15. Shi J, Zheng X, Li Y, Zhang Q, Ying S. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE J Biomed Health Inform. 2017;22:173–83.
16. Suk HI, Lee SW, Shen D, A. S. D. N. Initiative. Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. Brain Struct Funct. 2016;221(5):2569–87.
17. Tao S, Zhang T, Yang J, Wang X, Lu W. Bearing fault diagnosis method based on stacked autoencoder and softmax regression. In: Control conference (CCC), 2015 34th Chinese, IEEE. 2015. pp. 6331–5.
18. Suk H-I, Lee S-W, Shen D, A. S. D. N. Initiative. Deep ensemble learning of sparse regression models for brain disease diagnosis. Med Image Anal. 2017;37:101–13.
19. Suk H-I, Lee S-W, Shen D, A. S. D. N. Initiative. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct Funct. 2015;220(2):841–59.
20. Sarraf S, Tofighi G. Classification of Alzheimer's disease structural MRI data by deep learning convolutional neural networks. arXiv preprint arXiv:1607.06583. 2016.
21. Hosseini-Asl E, Gimel'farb G, El-Baz A. Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network. arXiv preprint arXiv:1607.00556. 2016.
22. Brosch T, Tam R, A. s. D. N. Initiative. Manifold learning of brain MRIs by deep learning. In: International conference on medical image computing and computer-assisted intervention. Springer. 2013, pp. 633–40.
23. Ngiam J, Chen Z, Bhaskar SA, Koh PW, Ng AY. Sparse filtering. In: Advances in neural information processing systems. 2011. pp. 1125–33.
24. Held E, Cape J, Tintle N. Comparing machine learning and logistic regression methods for predicting hypertension using a combination of gene expression and next-generation sequencing data. In: BMC proceedings. BioMed Central, vol. 10, no. 7. 2016. p. 34.
25. Risacher S, et al. Alzheimer's disease neuroimaging initiative (ADNI). Neurobiol Aging. 2010;31:1401–18.
26. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B (Stat Methodol). 2006;68(1):49–67.