Computers in Biology and Medicine xxx (xxxx) xxx

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Deep learning based pipelines for Alzheimer's disease diagnosis: A comparative study and a novel deep-ensemble method

Andrea Loddo *, Sara Buttau, Cecilia Di Ruberto

Department of Mathematics and Computer Science, University of Cagliari, via Ospedale 72, 09124, Cagliari, Italy

ARTICLE INFO

Keywords:
Alzheimer's disease
Computer-aided diagnosis
Convolutional neural networks
Deep learning
MRI image Analysis
Image classification

ABSTRACT

Background: Alzheimer's disease is a chronic neurodegenerative disease that destroys brain cells, causing irreversible degeneration of cognitive functions and dementia. Its causes are not yet fully understood, and there is no curative treatment. However, neuroimaging tools currently offer help in clinical diagnosis, and, recently, deep learning methods have rapidly become a key methodology applied to these tools. The reason is that they require little or no image preprocessing and can automatically infer an optimal representation of the data from raw images without requiring prior feature selection, resulting in a more objective and less biased process. However, training a reliable model is challenging due to the significant differences in brain image types.

Methods: We aim to contribute to the research and study of Alzheimer's disease through computer-aided diagnosis (CAD) by comparing different deep learning models. In this work, there are three main objectives: i) to present a fully automated deep-ensemble approach for dementia-level classification from brain images, ii) to compare different deep learning architectures to obtain the most suitable one for the task, and (iii) evaluate the robustness of the proposed strategy in a deep learning framework to detect Alzheimer's disease and recognise different levels of dementia. The proposed approach is specifically designed to be potential support for clinical care based on patients' brain images.

Results: Our strategy was developed and tested on three MRI and one fMRI public datasets with heterogeneous characteristics. By performing a comprehensive analysis of binary classification (Alzheimer's disease status or not) and multiclass classification (recognising different levels of dementia), the proposed approach can exceed state of the art in both tasks, reaching an accuracy of 98.51% in the binary case, and 98.67% in the multiclass case averaged over the four different data sets.

Conclusion: We strongly believe that integrating the proposed deep-ensemble approach will result in robust and reliable CAD systems, considering the numerous cross-dataset experiments performed. Being tested on MRIs and fMRIs, our strategy can be easily extended to other imaging techniques. In conclusion, we found that our deep-ensemble strategy could be efficiently applied for this task with a considerable potential benefit for patient management.

1. Introduction

Alzheimer's disease (AD) is an irreversible and chronic neurodegenerative disease and is the leading cause of dementia among the elderly [1]. It is estimated that 131 million people worldwide will suffer from AD and other dementias by 2050, presenting a significant health challenge in the 21st century [2]. In other words, 1 in 85 people will be diagnosed with Alzheimer's disease. Slowing down the course of the illness by even one year could decrease eleven million cases worldwide, thus significantly mitigating its impact on the world. People suffering from AD will gradually lose cognitive functions, such as remembering or thinking, and will eventually lose the ability to perform daily activities. In the context of AD development, mild cognitive impairment (MCI) represents a slight decline in mental skills along the continuum from normal cognition to AD, while more than 33% of individuals with MCI will progress to AD within five or more years [2]. Typically, there are two subtypes of MCI: stable MCI (sMCI), which will not develop to AD, and progressive MCI (pMCI), which will progress to AD.

Unfortunately, the cause and mechanism of AD are still not fully understood, and there is no curative treatment. However, the disease

* Corresponding author.
E-mail address: andrea.loddo@unica.it (A. Loddo).

https://doi.org/10.1016/j.compbiomed.2021.105032
Received 5 August 2021; Received in revised form 23 October 2021; Accepted 10 November 2021
Available online 21 November 2021

progression can be slowed down through medication, exercise, and memory training [3]. On this subject, the early detection of AD and the accurate diagnosis of MCI are crucial to delay disease progression and improve the patient's quality of life [4].

Since various neuroimaging tools, such as structural magnetic resonance imaging (sMRI) [5], resting-state functional magnetic resonance imaging (rs-fMRI) [6], and positron emission tomography (PET) [7] can differentiate neuropathological changes associated with these diseases, they have been increasingly used for the clinical diagnosis of AD and MCI [8].

Computer-aided diagnosis (CAD) systems based on neuroimaging tools have opened up new avenues in Alzheimer's research [9,10]. Neuroimaging has acquired a crucial role in diagnosing primary neurodegenerative diseases with magnetic resonance imaging (MRI), PET or Diffusion Tensor Imaging (DTI). These two techniques are used as biomarkers of the pathology and progression of Alzheimer's disease (AD) and differentiate AD from other neurodegenerative disorders. In general, they can give essential insights into the study of brain science, providing enormous information on global and local brain features, which can assess disease status, identify crucial brain regions of AD, and reveal the mechanism of AD. Therefore, they can be used as the foundation of CAD systems [9,10].

A CAD system is based on image analysis techniques, which we can distinguish between traditional and deep learning-based. The former uses a four-step pipeline: preprocessing, segmentation, feature extraction and classification. Image preprocessing prepares the image before analysing it to eliminate possible distortions or unnecessary data or highlight and enhance important features for further processing. Next, the segmentation step divides the significant regions into groups of pixels with shared characteristics such as colour, intensity, or texture extracted in the subsequent step. The purpose of segmentation is to simplify and change the image representation into something more meaningful and easier to analyse. The last step is classification, which consists of assigning a label to objects using supervised or unsupervised machine learning approaches. However, in recent years, deep learning workflows have emerged since the proposal of the AlexNet convolutional neural network (CNN) in 2012 [11]. CNNs do not follow the typical image analysis workflow because they can extract features independently without the need for feature descriptors or specific feature extraction techniques.

Deep learning algorithms differ from conventional machine learning methods. They require little or no image preprocessing and can automatically infer an optimal data representation from raw images without requiring prior feature selection, resulting in a more objective and less biased process. Therefore, deep learning algorithms are better suited for detecting fine and diffuse anatomical abnormalities. Moreover, they achieved optimal results in many domains such as speech recognition tasks, computer vision and natural language understanding and, more recently, medical analysis, such as MRI [12], microscopy [13], CT [14], ultrasound [15], X-ray [16] and mammography [17]. These models showed notable results for organ and substructure segmentation, disease detection and classification in pathology, brain, lung, abdomen, breast, bone, and retina.

Motivated by these properties and important results, we devise a contribution to the research and study of AD. In this work, we propose a comprehensive investigation on the problem of binary and multiclass classification of Alzheimer's Disease from MR images from different perspectives:

i). we trained several off-the-shelf CNN architectures on brain's MRI to find the most suitable one in the analysis of Alzheimer's patients;
ii). we performed a binary classification task to detect if patients are healthy or have dementia on four public data sets: three composed of MRIs and one of fMRI;
iii). we repeated the investigation on the same data sets to distinguish the different stages of dementia in a multiclass classification;
iv). we explored the possibility of combining a deep learning approach with a machine learning one and proposed a deep-ensemble based solution;
v). we investigated the robustness of the methods performing some cross-data sets experiments and evaluating the performance of the different systems.

The overall aim of the work is to investigate either the behaviour of the main existing off-the-shelf CNNs and a deep ensemble-based strategy aimed at the realisation of a comprehensive CAD framework based on patient MRIs and fMRIs. For this reason, we have also realised a preliminary methodology that seeks to offer a possible solution to the problem.

We verified the robustness of the solution on the Open Access Series of Imaging Studies (OASIS), the Alzheimer's Disease Neuroimaging Initiative (ADNI), and the Alzheimer-MRI (KAGGLE) public data sets. Our proposed approach achieves excellent results in identifying AD and exhibits promising performance in evaluating disease status.

The structure of the article is as follows. Sec. 1.1 presents a review of the machine and deep learning approaches for Alzheimer's disease, while Sec. 2 describes the data sets, the CNNs and the ensemble used in our experiments and a detailed outline of our methodological study. The experimental results are illustrated in Sec. 3. In Sec. 4 we analyse and discuss the experimental results and, finally, we give the conclusions and the future works in Sec. 5.

1.1. Related work

Although research is still evolving, the automatic classification of Alzheimer's disease has recently gained considerable attention. In this context, both traditional [18–20] and deep learning [21] approaches have been exploited. As the development of deep learning technology for neuroimaging data provides powerful tools to compute and analyse the brain network, many studies have exploited deep learning models to obtain AD-related features [22–29,29–34]. Existing works can be broadly divided into those oriented towards segmentation of brain parts or classification tasks [21]. We mainly focused this study on deep learning-based classification methods for detailed analysis of MRI tissue structures.

1.1.1. Traditional machine learning methods

Among the traditional machine learning approaches, Grey et al. [18] realised a multimodal classification method, based on imaging and biological information from the ADNI study, using the similarity measure generated by the random forest classifier. Zhang et al. [19] proposed a multi-view classifier to take advantage of multiple views of data. Their experiments explored the correlation between the features and the label by constructing a latent representation. A feature selection method for joint regression and classification via discriminative sparse learning and relational regularisation were realised by Lei et al. [20] to predict clinical scoring and use multimodal features to classify AD stages.

A traditional machine learning approach also requires a traditional image analysis procedure that uses a pipeline of four steps: preprocessing, segmentation, feature extraction and classification. Nevertheless, deep learning workflows have emerged since the proposal of the convolutional neural network AlexNet in 2012 [11]. CNNs do not follow the typical image analysis workflow because they can extract features independently without the need for feature descriptors or specific feature extraction techniques. In specific tasks, including the subject of this study, the traditional methods are computationally intensive and depend mainly on handcrafted features, which are difficult to obtain.

1.1.2. Deep learning methods

Among the state-of-the-art methods in the field of AD classification,

most authors have analysed the entire brain content [22–33], while others have focused on the grey matter or hippocampus [35–38]. Data representation is realised by various biomarkers: MRI, PET, fMRI, DTI. Some works have also made use of a combination of these. Key methods employed are auto-encoder [22,23], CNN [24–33], Deep Belief Networks (DBP) [34] and Fully Connected Networks (FCN) [29].

The setting of whole-brain MRI images analysed with CNN strategies is the same as the one addressed in this work. In this particular setting, Lee et al. [30] adopted an AlexNet-based method, with a data permutation scheme and an outlier removal approach, to perform both a binary classification of AD vs NC and a three-class classification between AD, NC, and MCI. An FCN was employed by Lian et al. [29], which achieved significant results, particularly in the pMCI vs sMCI classification. On the other hand, Hosseini et al. [39] and Sarraf et al. [40] used full brain fMRI images. The former employed a pre-trained 3D-CNN with a 3D convolutional autoencoder on the MRI data, producing several binary classifications (AD + MCI vs NC, AD vs MCI, MCI vs NC) and a three-class classification (AD vs MCI vs NC). The latter classified AD data from normal control using the LeNet architecture.

Some works [41–44] still employed a CNN strategy but based on a combination of whole-brain MRI and PET images. In this context, Feng et al. [43] implemented a deep learning network based on a 3D-CNN and Fully Stacked Bidirectional Long Short-Term Memory (FSBi-LSTM). Specifically, the image of each MRI or PET is transferred to the 3D-CNN network to extract features. In addition, FSBi-LSTM extracts high-level semantic and spatial information instead of the traditional FC layer. The aim is to use the multimodal data obtained from PET and MRI for AD diagnosis to address the binary classification problem of AD vs NC.

Regarding autoencoder-based approaches, Siqi et al. [22] performed classifications between AD and NC and MCI vs NC using an auto-encoder followed by a soft-max classifier. In contrast, Shi et al. [23] used both MRI and PET to implement a multimodal stacked DPN (MM-SDPN) algorithm for both binaries (AD vs NC, MCI-C vs MCI-NC, MCI vs NC) and multiclass (AD vs MCI-C vs MCI-NC vs NC) classification.

Finally, the approach taken by Andres et al. [34] is significantly different from the previously described. They worked on a combination of grey and white matter areas taken from MRI and PET and implemented a DBN that accepts 3D patches as input, further classified by an SVM. They aim to perform AD vs NC classification.

2. Material and methods

2.1. Data sets

We now describe the different data sets used in our study. These data sets make biomarkers such as neuroimaging modalities, genetic and blood information, and clinical and cognitive assessments publicly available. OASIS [45,46] is a project aimed at freely distributing brain MRI data, including two comprehensive data sets. The sagittal data set includes MRI data of 416 subjects (young, middle-aged, non-demented, and demented older adults) aged 18 to 96. The longitudinal data set includes MRI data of 150 subjects (non-demented and demented older adults) aged 60 to 96. The second data set used is Alzheimer-MRI, available from the Kaggle online community [47]. The data set will henceforth be referred to as the KAGGLE data set. The ADNI [48] is distinguished by being a longitudinal and multicentre study. It is the most common data set in the literature, used in about 90% of studies alone or combination with others. More details on each data set are provided below.

2.1.1. OASIS data set

OASIS is composed of 416 MRIs of patients. For each patient, there are three or four T1-weighted MRIs. All subjects are right-handed and belong to both sexes. One hundred of the included subjects over 60 years of age were diagnosed with very mild to mild AD. The data set structure is quite complex, and the sections provided are both sagittal and axial.

Table 1
Description of the OASIS data set. VM-AD indicates very mild AD, Mi-AD refers to mild AD, while Mo-AD indicates moderate AD patients.

	NC	VM-AD	Mi-AD	Mo-AD
Number	316	70	28	2
Ages (years)	75.9 ± 9.0	76.4 ± 7.0	77.2 ± 7.5	82 ± 5.7
Sex (M/F)	119/197	39/31	19/9	1/1

Each patient has several anatomical measurements from scans and the Clinical Dementia Rating (CDR), which indicates the level of dementia. All healthy patients (NC) have zero CDR, while patients with dementia (CDR > 0) are diagnosed with probable Alzheimer's disease (CDR = 0.5 for very mild dementia, CDR = 1 for mild dementia, and CDR = 2 for moderate dementia). The images have a resolution of 176 × 208. Further details are provided in Table 1.

The following is a list of the images used in our experiment:

- 1772 sagittal images, of which 1385 images are diagnosed as healthy and 387 with dementia (109 with mild dementia, 270 with very mild dementia and 8 with moderate dementia).
- 457 axial images, of which 357 healthy and 100 with dementia (28 with mild dementia, 70 with very mild dementia and 2 with moderate dementia).
- 2229 sagittal and axial images, of which 1742 healthy and 487 with dementia (137 with mild dementia, 340 with very mild dementia and 10 with moderate dementia.

Fig. 1 shows some examples of sagittal images.

2.1.2. KAGGLE data set

The KAGGLE data set consists of a total of 5121 axial images. Each image is labelled with the corresponding level of dementia: no dementia, very mild dementia, mild dementia and moderate dementia. The age of the patients is unknown, and no other data about them are provided. The data set includes 2560 healthy subjects and 2561 with dementia (1792 with very mild dementia, 717 with mild dementia, 52 with moderate dementia). The images have a resolution of 176 × 208. Examples are shown in Fig. 2. No details on patient status were provided for this data set.

2.1.3. ADNI data set

ADNI is a project underway since 2004 to follow AD's progress through its biomarkers to diagnose the disease in its early stages. Currently, ADNI is divided into three phases: ADNI1, ADNI GO/2 and ADNI3. ADNI registers participants aged 55–90 years among 57 sites in the United States and Canada. After giving consent, participants undergo several initial tests repeated over time, such as clinical assessment, neuropsychological and genetic testing, lumbar puncture, MRI, and PET scans. For MRI scans, the scans are 1.5T and 3T. For our analysis, we chose ADNI-1 MRI and ADNI-2 functional MRI data sets, as following described.

2.1.3.1. ADNI-1 MRI data set. Precisely, we exploited the *ADNI1: Complete 3Yr.3T* data set. Patients underwent screenings every six months for the first two years and another screening for the third year. The data set structure is as follows: 61 patients and 349 images, 127 healthy and 222 with dementia (145 with mild dementia and 77 with Alzheimer's). Other information is recorded for each patient, such as ID, screening date, gender, age, number of visits, and diagnosis (NC for Normal Control, i.e., no dementia; MCI for Mild Cognitive Impairment; and AD for Alzheimer's disease). The images have a resolution of 256 × 256. Some examples are shown in Fig. 3, and several details regarding the patients' distribution are presented in Table 2.

2.1.3.2. ADNI-2 fMRI data set. Regarding the functional MRI, we

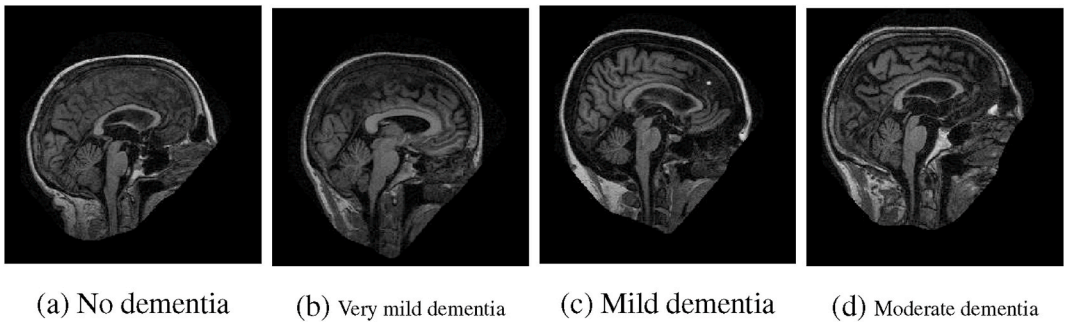

Fig. 1. Examples of sagittal images in the OASIS data set: from left to right, MRI of patients with no dementia, very mild dementia, mild dementia and moderate dementia.

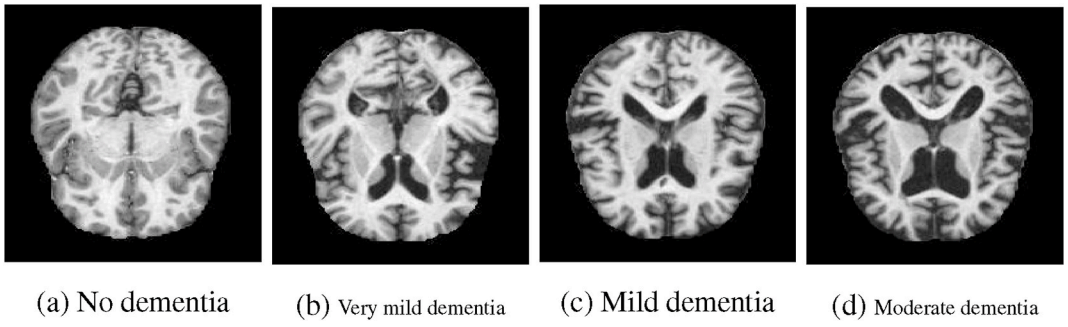

Fig. 2. Examples of axial images in the KAGGLE data set: from left to right, MRI of patients with no dementia, very mild dementia, mild dementia and moderate dementia.

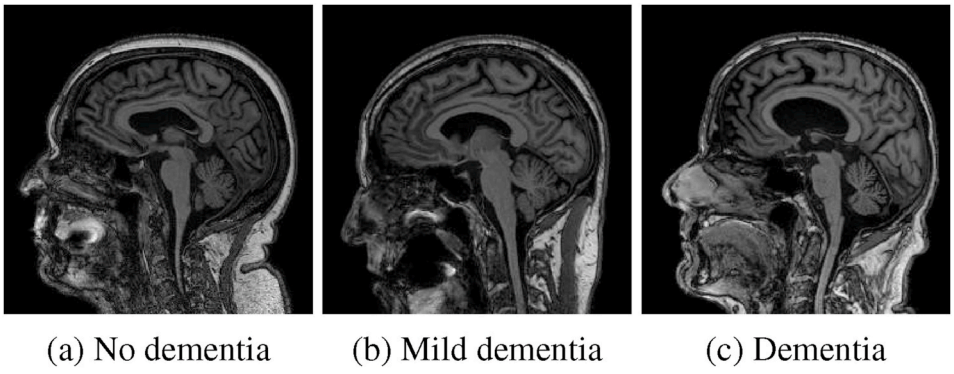

Fig. 3. Examples of images in the ADNI data set: from left to right, MRI of patients without dementia, with mild cognitive impairment and with dementia.

Table 2
Description of the ADNI MRI data set. Data are provided by ADNI.

	NC	sMCI	pMCI	AD
Number	213	90	126	130
Ages (years)	75.7 ± 5.0	74.9 ± 7.5	73.7 ± 7.0	74.1 ± 7.7
Sex (M/F)	108/105	58/32	68/58	64/66

Table 3
Description of the ADNI-2 fMRI data set data set. Data are provided by ADNI.

Class	Number	Ages (years)	Sex (M/F)
NC	433	75.49 ± 20.5	215/218
EMCI	431	71.94 ± 19.06	261/170
LMCI	354	72.47 ± 16.53	157/196
MCI	50	78.89 ± 12.11	37/13
SMC	68	72.35 ± 19.65	23/45
AD	198	74.88 ± 14.11	119/79

exploited the *rs-fMRI ADNI2* data set. In particular, it is composed of 1534 patients and a total amount of 402 446 resting-state functional MRIs (rs-fMRI), of which we describe the patients' distribution and status information in Table 3. Even in this case, the authors recorded additional information for each patient, such as ID, screening date, gender, age, number of visits and their diagnosis. The latter, in detail, are divided into the following six: NC for normal clinically; SMC for subjective memory concerns; EMCI for early mild cognitive impairment; LMCI for late mild cognitive impairment, and, finally, AD for mild Alzheimer's disease dementia. The images have a resolution of 64 × 64. Some image samples are shown in Fig. 4.

2.2. Convolutional neural networks

This work is oriented towards the exploitation of deep learning approaches. In particular, we aimed to study several off-the-shelf convolutional neural networks both as classifiers and feature extractors, embedded in an ensemble context, to produce a baseline on several case

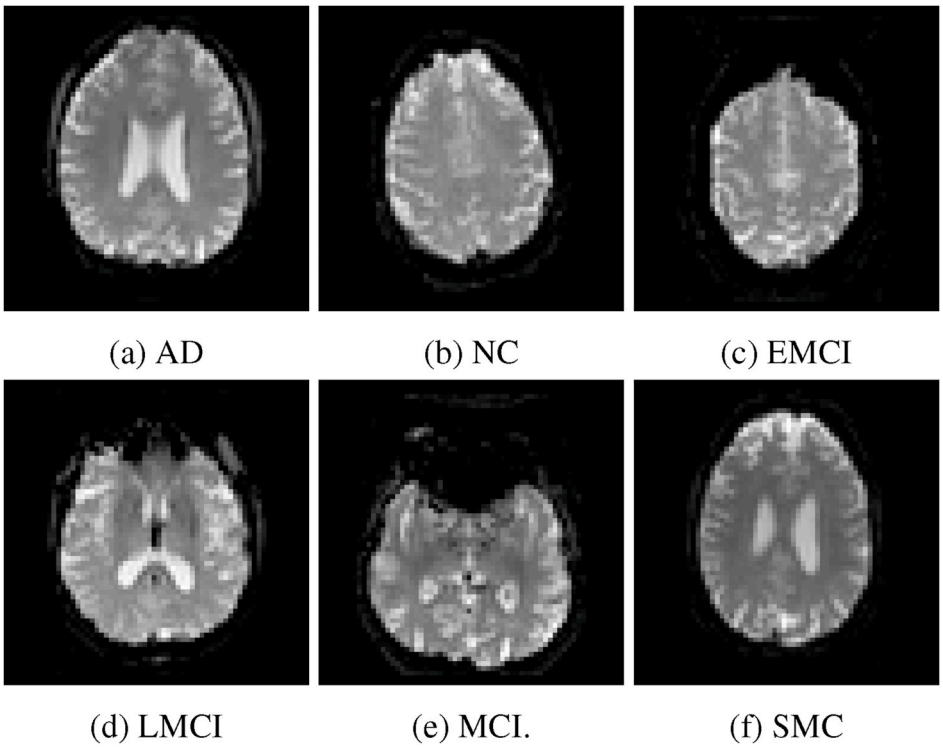

Fig. 4. Examples of images in the ADNI-2 fMRI data set.

studies concerning MRI and fMRI biomarkers. We evaluated the following architectures: AlexNet, ResNet-50, ResNet-101, GoogLeNet and Inception-ResNet-v2. They were all pre-trained on a well-known natural image data set (ImageNet [49]) and adapted to medical imaging tasks, following an established procedure for transfer learning and CNN models [50] fine-tuning. AlexNet [11] is a simple eight layers architecture. Nevertheless, it is frequently used for transfer learning and fine-tuning [50], since it offered excellent performance in many classification tasks [11]. The ResNet architectures are slightly more complex but, based on residual learning, they are easier to optimise even when the depth increases considerably [51]. They present 50 and 101 layers for ResNet-50 and ResNet-101, respectively. GoogLeNet [52] and Inception-ResNet-v2 [53] are both based on the Inception layer; indeed, Inception-ResNet-v2 is a variant of GoogLeNet. They differ in the number of layers, 100 and 164 for GoogLeNet and Inception-ResNet-v2, respectively. Regarding the transfer learning strategy, we followed the approach described in Ref. [50], keeping all CNN layers except the last fully connected one. We replaced it with a new layer, which was just initialised and set up to accommodate the new object categories.

CNNs can also be used to replace traditional feature extractors, as they have a robust ability to extract complex features that describe the image in detail [54–57]. Therefore, we exploited them for classification and feature extraction. In particular, we extracted features from the penultimate fully connected layer (FC7) on AlexNet and the last (only one) on the ResNet and Inception architectures to produce the most refined features for the proposed ensemble strategy.

2.3. Ensemble classifiers

The ensemble strategies are broadly categorised into bagging, boosting, and stacking. In particular, the main idea of bagging [58], also known as bootstrap aggregation, is to generate a set of independent observations with the same size and distribution as the original data. The set of observations generates an ensemble predictor better than the single predictor generated on the original data. Bagging increases two steps in the original models: first, generating bagging samples and passing each bag of samples to the base models; second, combining the predictions of multiple predictors. The bagging samples can be generated with or without replacement. The combination of the output of the base predictors may vary as the majority voting is used for classification problems. In contrast, the averaging strategy is used in regression problems to generate the ensemble output.

2.4. Evaluation metrics

The measures used to quantify the performance of each classification model are accuracy, sensitivity, specificity and F-score. In detail, Accuracy (Acc) is defined as the ratio of correctly labelled instances over the entire pool of cases; sensitivity (Sen), or true positive rate, or recall, is defined as the ratio of positives correctly identified by the prediction; specificity (Spec) measures the proportion of negative results that are correctly identified (also called the true negative rate); F-score is defined as the harmonic mean of precision and recall. Finally, since we are dealing with a multiclass imbalance problem, we also applied three of the most common global metrics for learning multiclass imbalance to evaluate the performance of the network [59]. The measures used are the macro geometric mean (MAvG), defined as the geometric mean of the partial accuracy of each class, the F-measure mean (MFM) and the macro arithmetic mean (MAvA) defined as the average of the partial accuracies of each class.

2.5. Our methodology

In this section, we describe our study. Specifically, it is a deep learning approach based on the transfer learning technique applied to several CNN architectures pre-trained on *Imagenet* [49]. We chose the networks based on their previous uses in the AD analysis domain, and their relative size/precision ratio. They are as follows: i). AlexNet [11]; ii). Inception-ResNet-v2 [53]; iii). the Residual Networks [51] ResNet-50 and ResNet-101; iv). GoogLeNet [52].

Finally, an ensemble of the best three trained networks has been proposed. Each of the most performing networks, after a fine-tuning

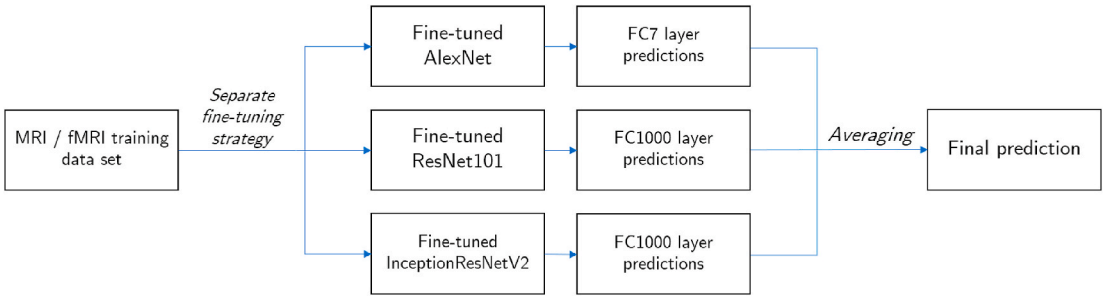

Fig. 5. Schematic representation of our proposed ensemble model.

procedure, has been used to extract a feature vector we combined and classified by an ensemble bagged trees model, with an average strategy, without any feature selection procedure. Specifically, fine-tuned AlexNet, ResNet-101 and Inception-ResNet-v2 have been used in our ensemble classifier. The illustration of our proposed ensemble strategy is shown in Fig. 5.

We conducted all the experiments on a single machine with the following configuration: Intel(R) Core(TM) i9-8950HK @ 2.90 GHz CPU with 32 GB RAM and NVIDIA GTX1050 Ti 4 GB GPU. The environment was MATLAB (ver. 2020b). The code, the models and the experiments are publicly available[1]. Each MRI data set has been divided into three subsets: one for training (training set), one for validation (validation set) and one for testing (test set), according to the following proportions: 80%, 10%, 10%. Regarding the fMRI data set and considering its sheer number of images, we divided it with the following proportions: 30% for training, 20% for validation, and 50% for testing. In addition, to further facilitate reproducibility, the subdivisions were randomly selected with a fixed seed. To ensure the heterogeneity of the training set, we trained the ensemble classifier with stratified 10-fold cross-validation to ensure that every fold contained a representative ratio of each class. We selected the model with the largest area under the ROC curve (AUC) for each case.

2.5.1. Settings and experiments on OASIS data set

We used sagittal images from the RAW directory and axial images from the *PROCESSED>MPRAGE>T88_111* directory. We performed two types of classification, a binary one, i.e., NC vs AD, and a multiclass one (NC, very mild AD, mild AD, moderate AD). The binary classification was done only on the sagittal sections, while the multiclass on both sagittal and axial sections. According to the CDR value, we produced the corresponding labels for each patient's image. Patients without CDR were considered healthy. Regarding augmentation, the options for the training set and the validation set are as follows:

- Random rotation between $-35°$ and $35°$;
- Random x-scale between 0.5 and 4;
- Random y-scale between 0.5 and 1;
- Grey-scale preprocessing.

The augmentation step applied to the training and validation set permit to realise a balanced data set.

We tested all the networks mentioned above. The images have been resized for each CNN. The training options of the networks are not all fixed. After empirical evaluation, we chose ADAM as the solver in most experiments because its accuracy was better than that obtained with SGDM and RMSPROP. The only fixed parameters are the initial learning rate set to 0.000 1 and the validation frequency set to 30. The same values were also used for the other two data sets. The following section presents the experimental results for the three best-performing networks: AlexNet, ResNet-101 and Inception-ResNet-v2. The maximum number of epochs for the binary classification is 300 for AlexNet and ResNet-101 and 200 for Inception-ResNet-v2. The minibatch size is 128 for AlexNet and 10 for the others. The maximum number of epochs for multiclass classification is 200 for sagittal sections, 70 for axial sections and 100 for both sections and all CNNs. The minibatch size is 128 for AlexNet, 16 for ResNet-101 and 10 for Inception-ResNet-v2 and all section types. Details on the data splits adopted are provided in Table 4.

Table 4
Training and testing set characteristics of the patients employed for the experiments performed on OASIS data set. NC indicates normal control, VM-AD indicates very mild AD, Mi-AD refers to mild AD, while Mo-AD indicates moderate AD patients.

Set	NC	VM-AD	Mi-AD	Mo-AD
Data set	316	70	28	2
Training set	252	56	22	0
Validation set	32	7	3	0
Test set	32	7	3	2

2.5.2. Settings and experiments on KAGGLE data set

We performed both binary classification on the KAGGLE data set, i.e., NC vs AD, and multiclass classification (NC, very mild AD, mild AD, moderate AD) using the AlexNet, ResNet-101 Inception-ResNet-v2 networks. For both tasks, the maximum number of epochs is 200 for all networks. Regarding the minibatch size and the solver, they differ for each network. In fact, ADAM was chosen for AlexNet and Inception ResNetV2, while SGDM was chosen for ResNet-101; a minibatch size of 128 was set for AlexNet, 10 for Inception-ResNet-v2 and 8 for ResNet-101.

Regarding the multiclass classification, we followed two different approaches: one for AlexNet, which consists of the average accuracy of five trainings on different test sets with the same training options, and one for ResNet-101 and Inception-ResNet-v2, to test the relationship between the accuracy value and the minibatch size and the number of epochs, as well as to find the best combination of the two parameters leading to the best performance.

We trained AlexNet for 200 epochs. Its accuracy no longer increased after then, while Inception-ResNet-v2 had a fixed minibatch size of 10 and improved the results by progressively increasing the number of

Table 5
Training and testing set characteristics of the patients employed for the experiments performed on the KAGGLE data set. ND, vmD, miD, and moD indicates no dementia, very mild dementia, mild dementia and moderate dementia, respectively.

Set	NC	vmD	miD	moD
Data set	2560	1792	717	52
Training set	2048	1434	573	42
Validation set	256	179	72	5
Test set	256	179	72	5

[1] https://github.com/andrealoddo/AD_classification.

Table 6
Training and testing set characteristics of the patients employed for the experiments performed on the ADNI MRI data set.

Set	NC	sMCI	pMCI	AD
Data set	213	90	126	130
Training set	170	72	100	104
Validation set	21	9	13	13
Test set	22	9	13	13

Table 7
Training and testing set characteristics of the patients employed for the experiments performed on the ADNI fMRI data set.

Set	NC	MCI	AD
Data set	27	68	20
Training set	8	20	6
Validation set	6	14	4
Test set	13	34	10

epochs: 30, 50, 70, 100. On the other hand, for ResNet-101, we used different experiments: we fixed the number of epochs and progressively reduced the minibatch size. The minibatch sizes used with the ResNet-101 were: 16, 8, 4, 2, while the number of epochs was 200, when the accuracy stopped improving. Finally, the augmentation options on the KAGGLE data set were the same as those on the OASIS data set, described in Sec. 2.5.1. Even in this case, the augmentation step permits to realise a balanced data set. Further details regarding the patients' distributions of the different sets are provided in Table 5.

2.5.3. Settings and experiments on ADNI data set

Also, on the ADNI data set, we performed both binary classification, i.e., NC vs AD, and multiclass classification (NC, MCI, AD). We used five networks for this experiment: AlexNet, ResNet-50, ResNet-101, Inception-ResNet-v2, and GoogLeNet. For both tasks, the maximum number of epochs is 200 for all networks. The minibatch is 128 for AlexNet and GoogLeNet, 10 for ResNet-101 and ResNetInceptionV2, and 30 for ResNet-50. Given the small size of the data set, we applied an augmentation to double the number of images, as described below:

- Random rotation between $-35°$ and $35°$;
- Random x-scale between 0.5 and 4;
- Random y-scale between 0.5 and 1;
- Grey-scale preprocessing.

The patients' data sets details produced, without the augmentation step, are described in Table 6.

2.5.4. Settings and experiments on ADNI fMRI data set

Further experimentation was conducted on a data set composed of resting-state functional MRIs. The original data set is composed of six different classes, as described in Section2.1.3: NC, EMCI, LMCI, MCI, SMC, and AD. In this experiment, we grouped all the classes related to the cognitive impairment status, making it a three-class classification between NC, MCI, and AD. We used five networks for this experiment: AlexNet, ResNet-50, ResNet-101, Inception-ResNet-v2, and GoogLeNet. Unlike the previous, we adopted a maximum of 30 epochs due to the huge number of images available. For the same reason, the split adopted were the following: 30% of the images for training, 20% for validation, and 50% for testing. However, to maintain fairness and consistency concerning the previous experiments, we kept only the images with a single scan and removed all the others. More specifically, we removed all the *motion correcting series* and the *perfusion Weighted* because they contain more than one scan per image. Finally, we converted all the images from DCM to JPG format. The final setting of the image composition is described in Table 7.

2.5.5. Cross-data set settings and experiment

The models obtained by transfer learning on the KAGGLE data set were also used for another experiment: a refined fine-tuning on OASIS and a test on OASIS itself. This experiment aimed to see how networks that had already learned features from brain MRIs would behave on new MRIs belonging to a different data set. The objective was first to improve the overall classification accuracy and, second, to show the robustness and validity of our approach in a cross-data set scenario.

As mentioned in Sec. 2.5.2, the KAGGLE data set only provides axial sections, whereas OASIS provides both sagittal and axial images. Therefore, models trained on KAGGLE should perform better on the axial sections of OASIS, as they are prepared with the most similar images. However, experimentation was also carried out on sagittal sections, thus using all the images of the OASIS data set. Binary classification, however, was only done on the sagittal views of the data set. The networks are again AlexNet, ResNet-101 and Inception-ResNet-v2. Regarding the training parameters, the maximum number of epochs for the binary task is 300 for all networks. The minibatch size is 128 for AlexNet and 10 for the others.

For multiclass classification, the maximum number of epochs is 200 for sagittal sections, 70 for axial sections and 100 for both sections and all CNNs. The minibatch size is 128 for AlexNet, 16 for ResNet-101 and 10 for Inception-ResNet-v2 and all section types.

3. Results and analysis

This section aims to present the experimental results and compare them with some relevant studies available in the literature. We order the different approaches for the detection of AD considering the main aspects on which they are based: in terms of input, type of biomarkers used, some details about the data set, which deep learning technique was employed, and, finally, which performance measures were calculated for the evaluation.

Considering all the studies presented, we can group the approaches to input data management into four different categories, depending on the type of features extracted: voxel-based, slice-based, patch-based, and ROI- based. Concerning biomarkers, the most prevalent type of neuroimaging modality used is MRI. However, several studies considered PET and fMRI as equally discriminating. Some studies have considered other aspects such as age, gender, education level, speech pattern, EEG, retinal abnormalities, postural kinematic analysis, cerebrospinal fluid biomarkers, and neuropsychological measures as possibly relevant for AD detection.

The main algorithms of deep learning techniques are AEs, DNNs, DBNs, and 2D/3D CNNs, as introduced in Sec. 1.1.

One of the main problems encountered when comparing our study with others concerns the availability of clear data sets and training procedures. We selected work in which the authors employed public data sets, i.e., OASIS or ADNI. To the best of our knowledge, there are no publicly presented results yet regarding the KAGGLE data set. However, most of the studies in the literature did not submit their source code to any hosting platform for software development or to online competition. Moreover, they did not specify some essential aspects in the experiments: which type of section they chose, whether axial or sagittal, which images of the data set they selected, details about the training procedures, i.e., the values of the training parameters. Therefore, it is not easy to compare studies impartially with each other. Consequently, we can consider the performance measures reported for each approach, usually accuracy and, in some cases, specificity and sensitivity. Even for studies on the same data set and with the same number of subjects, the results may still not be comparable because different subjects may be used as training sets and test sets. Many studies address the NC vs AD problem because it helps other classification tasks, especially in understanding early signs of AD. But the most important and main challenge in AD assessment is determining whether someone has MCI or not and predicting whether an MCI patient will develop the disease. Therefore, we

Table 8
Results on OASIS data set on NC vs AD classification.

Work	Input	Data set details	Method	Acc	Sen	Spe	F-score
[60]	Slice-based (MRI)	416 subjects. Train 200. (sagittal)	2D CNN based on the VGG16	74.12	–	–	–
[60]	Slice-based (MRI)	”	InceptionV4	96.25	–	–	–
[60]	Slice-based (MRI)	”	VGG16	92.30	–	–	–
[61]	Voxel-based (MRI)	98 subjects including 49 AD and 49 NC.	Voxel-based DBN	92.16	90.59	93.36	–
[62]	Slice-based (MRI)	95 subjects including 51 AD and 44 NC (only right-handed subjects)	Stacked sparse AEs and a softmax layer with fine tuning	91.60	98.09	84.09	–
[63]	Slice-based (MRI)	80 subjects including 40 AD and 40 NC. Several scans for each subject.	2D CNN	85.00	90.00	80.00	–
[64]	Slice-based (MRI)	196 subjects including 98 AD and 98 NC from OASIS + local data	2D CNN	97.65	97.96	97.35	–
[65]	Slice-based (3D MRI)	382 subjects including 167 NC, 87 very mild AD, 105 mild AD, 23 AD.	AlexNet	89.66	100.0	82.0	–
This	Slice-based (MRI)	1772 sagittal images including 387 AD and 1385 NC	AlexNet	99.44	98.75	99.64	99.19
This	Slice-based (MRI)	”	ResNet-101	97.74	95.97	97.63	96.80
This	Slice-based (MRI)	”	Inception-ResNet-v2	98.31	97.14	97.99	97.56
This	Slice-based (MRI)	”	Deep-Ensemble	98.51	97.57	98.42	97.85

Table 9
Results on OASIS data set and multiclass classification (NC, very mild AD, mild AD, moderate AD.).

Work	Input	Data set details	Method	Acc	Sen	Spe	MAvG	MFM	MAvA
[66]	Slice based (MRI)	416 subjects with axial scans	2D CNN inspired by Inception-V4	73.75	–	–	–	–	–
[67]	Slice based (MRI)	416 subjects with axial scans	Ensemble of three DenseNet networks	93.18	93.00	–	–	–	–
[68]	Slice based (MRI)	416 subjects with axial scans	2D CNN model for each view with the majority voting strategy (Inception V4 -ResNet)	93.18	93.00	–	–	–	–
[65]	Slice based (MRI)	382 subjects including 167 NC, 87 very mild AD, 105 mild AD and 23 AD	AlexNet	92.85	92.85	74.27	–	–	–
[69]	Slice based (MRI)	382 subjects including 167 NC, 87 very mild AD, 105 mild AD and 23 AD	Siamese convolutional neural network inspired by VGG16	99.05	–	–	–	–	–
This	Slice based (MRI)	1772 sagittal images including 1385 NC, 270 very mild AD, 109 mild AD and 8 moderate AD	AlexNet	100	100	100	100	100	100
This	Slice based (MRI)	”	ResNet-101	97.74	96.77	99.28	99.27	97.91	99.28
This	Slice based (MRI)	”	Inception-ResNet-v2	88.70	82.07	84.26	86.78	80.64	84.26
This	Slice based (MRI)	457 axial images including 357 NC, 70 very mild AD, 28 mild AD and 2 moderate AD	AlexNet	78.26	48.68	88.89	78.26	78.26	48.68
This	Slice based (MRI)	”	ResNet-101	82.61	68.43	77.25	72.49	65.77	77.25
This	Slice based (MRI)	”	Inception-ResNet-v2	91.30	87.94	82.28	81.42	83.59	82.28
This	Slice based (MRI)	2229 sagittal and axial images including 1742 NC. 340 very mild AD. 137 mild AD and 10 moderate AD	AlexNet	94.17	87.33	88.59	91.13	87.92	88.59
This	Slice based (MRI)	”	ResNet-101	90.58	83.60	90.15	89.90	86.37	90.15
This	Slice based (MRI)	”	Inception-ResNet-v2	92.83	92.09	89.95	89.22	89.95	89.95
This	Slice based (MRI)	”	Deep-Ensemble	98.24	93.05	97.31	94.24	96.38	96.14

also considered studies working on AD, MCI and NC. Our search includes all the images present in each data set considered. If several sections are available, we analyse them all, either separately or by combining them.

3.1. Results for OASIS data set

In Table 8 and Table 9, we present numerical results for the binary and multiclassification task on the OASIS data set, respectively. The tables include our results and those of the literature. For each approach, we indicate the reference number, the type of input and biomarkers, some details about the data set, the deep learning technique and the values of the performance measures, where reported.

Overall, all three networks gave satisfactory results. In particular, AlexNet performed best in the sagittal imaging scenario, achieving 99.44% and 100% accuracy in binary and multiclass classification, respectively. It also outperformed the state-of-the-art in the remaining metrics, particularly in the multiclass experiments. Inception-ResNet-v2 achieved the highest accuracy for axial images, i.e., 91.30%, outperforming the other two networks tested for all metrics. However, this classification seems to be the most critical image configuration because all metrics are critically lower than the sagittal image scenario, e.g., AlexNet achieved a minimum sensitivity of 48.68%. The combination of axial and sagittal images did not produce any improvement in results. In particular, Table 9 shows that no network managed to achieve a maximum score for all metrics, in contrast to the two previous case studies. In this case, the deep-ensemble strategy is preferable considering that it obtained 98.24% accuracy, outperforming the remaining networks. Multiclass classification on axial sections achieved relatively

Table 10
Results on the KAGGLE data set for NC vs AD classification.

Work	Input	Data set details	Method	Acc	Sen	Spe	F-score
This	Slice-based (MRI)	5121 axial images including 2560 NC and 2561 AD	AlexNet	89.65	89.79	89.65	89.72
This	Slice-based (MRI)	”	ResNet-101	96.09	96.11	96.09	96.10
This	Slice-based (MRI)	”	Inception-ResNet-v2	91.21	91.44	91.21	91.32
This	Slice-based (MRI)	”	Deep-Ensemble	96.57	96.57	98.28	96.57

Table 11
Results on the KAGGLE data set for multiclass classification (NC vs very mild AD vs mild AD vs moderate AD).

Work	Input	Data set details	Method	Acc	Sen	Spe	MAvG	MFM	MAvA
This	Slice-based (MRI)	5121 axial images including 2560 NC, 1792 very mild AD, 717 mild AD, and 52 moderate AD	AlexNet	89.26	90.58	81.66	80.54	84.83	81.66
This	Slice-based (MRI)	”	ResNet-101	96.48	97.78	96.78	96.74	97.26	96.78
This	Slice-based (MRI)	”	Inception-ResNet-v2	89.65	90.11	85.64	88.45	87.39	85.64
This	Slice-based (MRI)	”	Deep-Ensemble	97.71	96.67	98.22	96.41	95.98	96.40

Table 12
Results on the ADNI data set for NC vs AD classification and comparison with the state-of-the-art.

Work	Input	Data set details	Method	Acc	Sen	Spe	F-score
[70]	ROI-based (MRI)	311 subjects including 65 AD, 67 MCIc, 102 MCInc, and 77 NC	Stacked sparse AEs and a softmax layer	88.16	88.57	87.22	–
[41]	Voxel and Patch-based (MRI, PET)	398 subjects including 93 AD, 204 MCI and 101 NC	Multi-modal DBM with SVM	95.35	95.05	95.22	–
[71]	ROI-based (MRI, PET)	311 subjects including 65 AD, 67 MCIc, 102 MCInc and 77 NC	Stacked sparse AEs and a softmax layer	91.4	92.32	90.42	–
[72]	Voxel-based (MRI)	1728 subjects including 346 AD, 358 LMCI, 450 MCI, and 574 NC	3D CNN based on ResNet with a lower number of layers	94.00	–	–	–
[23]	ROI-based (MRI, PET)	202 subjects including 51 AD, 43 MCIc, 56 MCInc, and 52 NC	Multi-modal stacked DPN and a linear kernel SVM	97.13	96.33	98.53	–
[73]	Voxel-based (MRI)	825 subjects including 407 NC and 418 AD	3D CNN	99.20	98.90	99.50	–
[74]	ROI-based (MRI)	818 subjects including 188 AD. 229 NC. 401 MCI	2.5D CNN	79.90	84.00	74.80	–
[40]	Slice-based (functional MRI)	43 subjects including 28 AD and 15 NC	LeNet-5	96.85	–	–	–
This	Slice-based (MRI)	349 subjects including 77 AD. 145 MCI. 127 NC	AlexNet	85.71	84.87	87.06	85.95
This	”	”	ResNet-101	97.14	97.83	96.15	96.98
This	”	”	Inception-ResNet-v2	97.14	96.43	97.73	97.07
This	”	”	ResNet-50	99.34	98.73	99.40	99.03
This	”	”	GoogLeNet	94.29	93.88	93.88	93.88
This	”	”	Deep-Ensemble	99.74	99.36	99.89	99.55

low sensitivity values, especially with AlexNet and ResNet-101, which classified healthy cases more accurately. Finally, the sensitivity and specificity values obtained are generally high (mainly in the range of 90–100%), indicating that the models produced are equally effective in recognising *positive* and *negative* cases, except in the axial imaging scenario.

3.2. Results for KAGGLE data set

In Tables 10 and 11, we present the numerical results for the binary and multiclassification task on the KAGGLE data set, respectively.

We achieved the best results with the ResNet-101 network for binary classification, with an accuracy of 96.09%, which also shows high specificity and sensitivity. In contrast, the other networks appear less stable concerning the experiments on the OASIS data set. About multi-class classification, ResNet-101 confirmed and improved on the results obtained in the binary case. It outperformed both AlexNet and Inception-ResNet-v2 in every single metric.

However, the deep-ensemble strategy is the most performing in both cases, obtaining 96.57% and 97.71% accuracy in binary and multiclass cases.

3.3. Results for ADNI data set

In Tables 12 and 13, we present the numerical results for the binary and the multiclassification task on the ADNI data set, respectively.

Regarding the binary classification, we obtained values of 100% for each metric using ResNet-50. In general, every network tested showed good classification results, except for AlexNet, which could not even reach the 90% threshold for any metric. ResNet-101 again showed the best results for multiclass classification on sagittal sections, with 100% in every metric, followed by GoogLeNet (97.14% overall accuracy) and AlexNet (94.29%). In contrast to the binary classification, ResNet-50 was the only network with an accuracy of less than 90%. Its performance was generally lower than the binary classification and the rest of the networks used in the multiclass experiment.

Even in this case, our deep-ensemble strategy showed the highest results in every metric, obtaining 99.74% and 99.22% accuracy in the two cases.

3.4. Results for ADNI fMRI data set

In this work, we aimed at finding a method that could be adaptable for heterogeneous biomarkers. Therefore, we exploited the off-the-shelf CNNs that produced the best performance in the MRI cases for the ADNI

Table 13
Results on the ADNI data set for multiclass classification (NC vs MCI vs AD).

Work	Input	Data set details	Method	Acc	Sen	Spe	MAvG	MFM	MAvA
[71]	ROI-based (MRI)	311 subjects including 65 AD, 169 MCI and 77 NC	Stacked sparse AEs and a softmax regression layer	59.19	51.38	84.36	–	–	–
[39]	Voxel-based (MRI)	CADDementia + ADNI, 210 subjects including 70 AD, 70 MCI, and 70 NC	3D CNN pre-trained with stacked 3D convolutional AEs	89.10	–	–	–	–	–
[75]	Slice-based (MRI)	900 subjects including 300 AD, 300 MCI and 300 NC	2D CNN based on the VGGNet-16	92.25	–	–	–	–	–
[39]	Voxel- based (MRI)	CADDementia + ADNI, CADDementia: 30 subjects for pre-training. ADNI: 210 subjects including 70 AD, 70 MCI and 70 NC	3D CNN pre-trained with stacked 3D convolutional AEs	94.80	–	–	–	–	–
[76]	Slice-based (MRI)	660 images including 188 AD, 243 MCI and 229 NC	A 2D CNN based on ResNet-18	56.80	–	–	–	–	–
[77]	ROI-based (MRI, PET, Genetic Data)	805 subjects including 190 AD, 389 MCI and 226 NC subjects. All the subjects have MRI data, while only 736 subjects have genetic data and 360 subjects have PET data.	Novel three-stage deep feature learning and fusion framework using DNN	64.40	–	–	–	–	–
[78]	ROI-based (MRI, PET, Genetic Data)	805 subjects including 186 AD, 393 MCI, and 226 NC	Multi-task deep neural network with a softmax layer	65.80	–	–	–	–	–
[79]	Slice-based (MRI)	504 subjects including 101 AD, 234 MCI, and 169 NC	2D CNN	96.00	96.00	98.00	–	–	–
[80]	ROI-based (MRI)	ADNI + CADDementia - 504 subjects including 101 AD, 232 MCI, and 171 NC	Stacked AEs and a softmax layer	56.80	–	–	–	–	–
[81]	ROI-based (MRI)	694 subjects (~2 scans per subject) including 272 AD, 726 MCI, and 379 NC	Deep supervised feature extraction approach using General Stochastic Networks	79.40	–	–	–	–	–
[72]	Voxel-based (MRI)	1728 subjects including 346 AD. 358 LMCI. 450 MCI, and 574 NC	3D CNN based on ResNet with a lower number of layers	87.00	–	–	–	–	–
[82]	Voxel-based (MRI, FDG, PET)	615 images including 193 AD, 215 MCI, and 207 NC	3D VGGNet-16 pre-trained by an AE	91.13	–	–	–	–	–
[83]	Voxel-based (MRI, age, gender)	841 subjects including 200 AD, 411 MCI, and 230 NC	3D CNN with transfer learning	61.10	63.00	–	–	–	–
[36]	ROI-based (MRI, DTI)	531 subjects including 53 AD, 228 MCI, and 250 NC	3D Inception-based CNN for each region	68.90	–	–	–	–	–
[84]	Slice-based (MRI-GM)	321 subjects including 150 AD, 129 MCI, and 112 NC. Total of 3744 scans	3D CNN based on VGGNet	91.32	–	–	–	–	–
[85]	Voxel-based (MRI)	ADNI + AIBL (2464 subjects with 20 060 scans)	3D CNN with 3 different filter size in its first convolutional layer	60.20	–	–	–	–	–
[86]	ROI-based (MRI, PET, Genetic Data)	805 subjects including 190 AD, 389 MCI, and 226 NC subjects. All the subjects have MRI data. while only 736 subjects have genetic data. and 360 subjects have PET data.	Novel three-stage deep feature learning and fusion framework using DNN	64.40	–	–	–	–	–
[87]	Slice-based (MRI)	150 subjects including 50 AD, 50 MCI, and 50 NC	2D inspired by VGGNet-16	95.73	–	–	–	–	–
[88]	Voxel- based (MRI)	833 subjects including 221 AD, 297 MCI, and 315 NC	Ensemble of 3D DenseNets	97.52	–	–	–	–	–
This	Slice-based (MRI)	349 subjects including 77 AD, 145 MCI, 127 NC	AlexNet	94.29	95.24	93.45	93.31	94.16	93.45
This	”	”	ResNet-101	100	100	100	100	100	100
This	”	”	Inception-ResNet-v2	91.43	94.12	87.5	85.5	89.08	87.5
This	”	”	ResNet-50	85.71	87.39	84.52	83.84	84.89	84.52
This	”	”	GoogLeNet	97.14	96.3	97.44	97.37	96.71	97.44
This	”	”	Deep-Ensemble	99.22	97.53	99.20	98.36	98.81	98.33

fMRI data set. In general, regarding the fMRI biomarkers and as reported in Tables 14 and 15, the tested off-the-shelf networks produced lower performance than MRI data set experimentations. Specifically, in the binary case, no network outperforms the state-of-the-art. At the same time, our ensemble proposal can reach 98.3% accuracy, exceeding both Sarraf et al.works [40,89] which, in any case, exploited only a portion of the same ADNI fMRI data set considered in this work. Also, as reported in Fig. 6, the AD class reached 97.0% accuracy, with a reduced error margin. In this case, the MCI class is undoubtedly the most difficult to classify, having 89.4% accuracy. This problem is certainly due to the heterogeneity of the four different MCI classes considered as grouped inside this class. Finally, as it can be seen from Fig. 7, the ROC curve performance of the deep-ensemble model reached 99%.

3.5. Cross-data sets results

In the second experiment on the KAGGLE data set, we fine-tuned the produced KAGGLE models, as described in Sec. 2.5.5. In detail, we used them to perform a fine-tuning strategy on a 10% portion of the OASIS data set. The validation set was another 10%, while we used the remaining 80% portion as the test set. The results for the binary task on the OASIS sagittal images are given in Table 16, while Table 17 gives the results for the multiclass task.

Concerning binary classification for sagittal images, the results obtained were no better than the previous ones, although generally good and above 90%. The best performing network was AlexNet, with an accuracy of 98.87%, a sensitivity of 99.29% and an F-score of 98.35%. However, Inception-ResNet-v2 was found to have the best specificity at 98.56%. For multiclass classification on sagittal sections, ResNet-101

Table 14
Results on the ADNI fMRI data set for the binary classification (AD vs NC).

Work	Input	Data set details	Method	Acc	Sen	Spe	F-score
[40]	Slice-based (fMRI)	43 subjects including 28 AD and 15 NC	LeNet-5	96.85	–	–	–
[89]	Slice-based (fMRI)	43 subjects including 28 AD and 15 NC	LeNet-5	96.85	–	–	–
This	Slice-based (fMRI)	47 subjects including 20 AD and 27 NC	AlexNet	80.8	61.5	100.0	76.2
This	Slice-based (fMRI)	”	ResNet-101	76.9	79.5	75.6	77.5
This	Slice-based (fMRI)	”	Inception-ResNet-v2	82.0	66.7	96.3	78.8
This	Slice-based (fMRI)	”	Deep-Ensemble	98.3	96.0	95.9	95.9

achieved higher overall accuracy and sensitivity than standard ResNet-101, with 98.31% and 97.46%, respectively. However, AlexNet achieved the best results in these image settings, even for multiclass metrics, showing excellent classification ability for each class. For axial sections, the accuracy exceeded standard transfer learning with AlexNet and ResNet-101 with SGDM solver, while they were lower for ResNet-101 with ADAM solver and KAGGLE-trained Inception-ResNet-v2. Overall, AlexNet and Inception-ResNet-v2 achieved the best accuracy, sensitivity, F-score and MAvG, and specificity, MFM and MAvA, respectively. It becomes more evident that axial section images alone are not the most appropriate for the classification tasks in this scenario. When considering the combination of sagittal and axial sections, we obtained outstanding results with both AlexNet and Inception-ResNet-v2, with some differences on each side. In detail, AlexNet showed higher sensitivity and F-score values, while Inception-ResNet-v2 showed a higher overall accuracy of 95.07% and the highest multiclass specificity and accuracy values.

Despite the promising results produced by the fine-tuned CNNs, our deep-ensemble strategy outperforms the remaining methods even in this case, achieving an accuracy of 99.29% in the binary case and 96.02% in the multiclass one. Also, as reported in Table 9, the AD class reached 99.4% accuracy, with a low error margin. As shown in Fig. 8, the MCI class appeared the most difficult to classify in this scenario, considering its 92.7% accuracy, probably due to the differences in the images of the two data sets. Finally, from Fig. 9, we can see that the ROC curve performance of the presented deep-ensemble strategy reached 99%.

4. Discussion

This work aimed to explore the use of deep learning techniques to diagnose AD. More specifically, to investigate the possibility of using convolutional neural networks to detect AD disease and differentiate different degrees of dementia from MRI or fMRI images. We used four public data sets for our study and compared our results with state of art. In particular, the first objective was to train the networks for binary classification, i.e., to distinguish between healthy and dementia images. We best achieved this goal with AlexNet, ResNet-101, and ResNet-50 architectures, which reached an accuracy of 99.44% on the OASIS data set; 96.09% on the KAGGLE, and 100% on the ADNI, respectively. A second objective was to identify a network capable of classifying the stages of dementia. This goal was also achieved, with 100% accuracy by AlexNet for the four-class classification task on the sagittal sections of the OASIS data set; by ResNet-101, for the three-class classification on the ADNI data set with 100% accuracy; and the four-class classification on the KAGGLE data set with 96.48% accuracy.

Considering the results achieved by the off-the-shelf CNNs, we

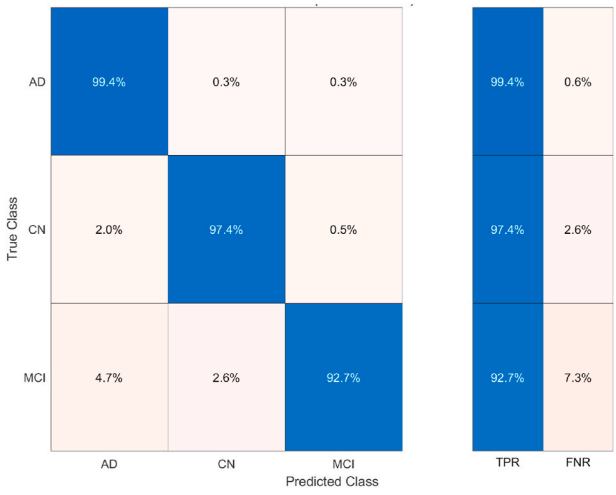

Fig. 6. Multiclass classification confusion matrix on the ADNI fMRI data set (NC vs MCI vs AD).

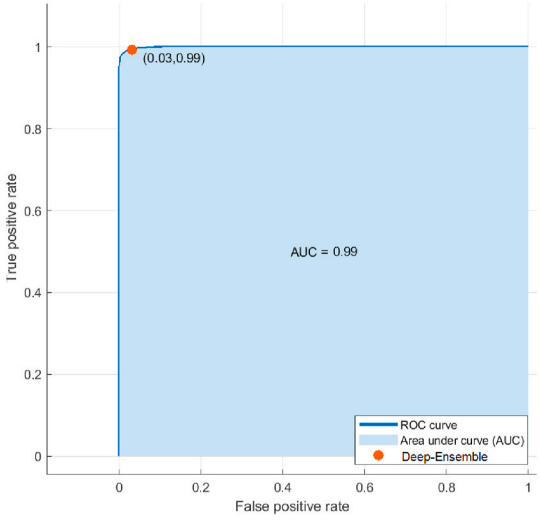

Fig. 7. ROC curve for our deep-ensemble strategy on the cross-data set multiclass experiment.

Table 15
Results on the ADNI fMRI data set for the multiclass classification (NC vs MCI vs AD).

Work	Input	Data set details	Method	Acc	Sen	Spe	MAvG	MFM	MAvA
This	Slice-based (fMRI)	115 subjects including 27 NC, 68 MCI and 20 AD	AlexNet	78.26	81.67	82.50	81.32	80.39	81.67
This	Slice-based (fMRI)	”	ResNet-101	82.61	87.08	81.28	88.78	86.14	87.09
This	Slice-based (fMRI)	”	Inception-ResNet-v2	86.96	91.67	85.56	84.97	90.67	91.67
This	Slice-based (fMRI)	”	Deep-Ensemble	98.16	96.49	98.62	97.52	96.97	97.53

Table 16
Cross-data set (training on KAGGLE, fine-tuning and testing on OASIS) results for NC vs AD classification.

Work	Input	Data set details	Method	Acc	Sen	Spe	F-score
This	Slice-based (MRI)	OASIS 1772 sagittal images including 1385 NC and 387 AD	AlexNet	98.87	99.29	97.44	98.35
This	Slice-based (MRI)	”	ResNet-101	98.31	97.96	97.07	97.52
This	Slice-based (MRI)	”	Inception-ResNet-v2	98.31	97.37	98.56	97.96
This	Slice-based (MRI)	”	Deep-Ensemble	99.29	98.34	99.55	98.94

Table 17
Cross-data set (training on KAGGLE, fine-tuning and testing on OASIS) results for multiclass classification.

Input	Data set details	Method	Acc	Sen	Spe	F-score	MAvG	MFM	MAvA
Slice-based (MRI)	1772 OASIS sagittal images including 1385 NC, 270 very mild AD, 109 mild AD, and 8 moderate AD	AlexNet	97.18	95.31	97.8	96.54	97.79	96.47	97.8
Slice-based (MRI)	”	ResNet-101	98.31	97.46	91.67	94.48	90.86	93.93	91.67
Slice-based (MRI)	”	Inception-ResNet-v2	77.97	78.41	78.41	78.41	88.32	78.41	33.33
Slice-based (MRI)	457 OASIS axial images including 357 NC. 70 very mild AD. 28 mild AD and 2 moderate AD	AlexNet	84.78	84.78	51.46	68.12	82.21	51.21	51.46
Slice-based (MRI)	”	ResNet-101	80.44	72.22	48.28	57.87	35.91	53.25	48.28
Slice-based (MRI)	”	Inception-ResNet-v2	80.43	63.73	69.97	66.70	68.97	66.35	69.97
Slice-based (MRI)	2229 OASIS sagittal and axial images including 1742 NC. 340 very mild AD. 137 mild AD and 10 moderate AD	AlexNet	94.17	90.76	89.50	90.13	89.37	90.09	89.50
Slice-based (MRI)	”	ResNet-101	84.75	75.7	65.63	70.31	57.72	65.24	65.63
Slice-based (MRI)	”	Inception-ResNet-v2	95.07	86.15	93.16	89.52	92.83	88.02	93.16
Slice-based (MRI)	”	Deep-Ensemble	96.02	87.01	94.09	90.04	93.76	88.90	94.09

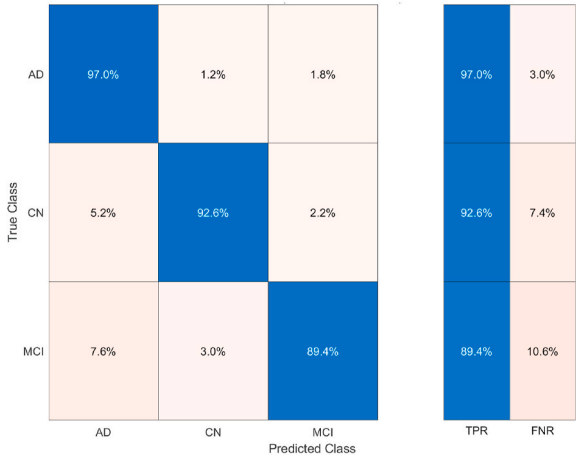

Fig. 8. Multiclass classification confusion matrix on the cross-data set experiment.

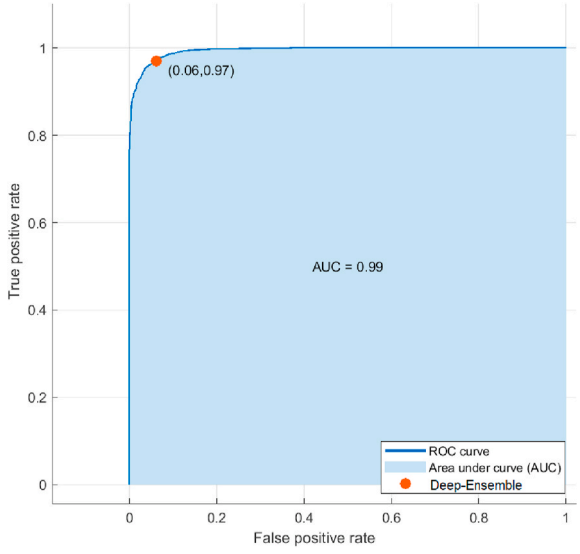

Fig. 9. ROC curve for our deep-ensemble strategy on the cross-data set multiclass experiment.

explored the possibility of combining deep learning features with a machine learning approach. Therefore, we proposed a deep-ensemble based solution, which reached outstanding performance in the previous three experiments, outperforming the networks in all cases, except for OASIS one.

Moreover, we carried out cross-data sets experiments, which allowed us to further validate our results. On the one hand, it was possible to compare the performance of the models trained on KAGGLE and OASIS. On the other hand, it was possible to validate the results of the OASIS models by comparing them with the KAGGLE models on the same data set.

The models trained on the KAGGLE data set performed equally well on OASIS, achieving excellent accuracies, such as 98.7% (AlexNet, for binary classification), 98.31% (ResNet-101, for multiclass classification - sagittal section), 95.07% (Inception-ResNet-v2, for multiclass classification - sagittal and axial sections combined). In general, the networks trained on OASIS were better than the networks trained on KAGGLE and tested on OASIS. However, the accuracies achieved by both models are mostly above 90%. Even in this context, our deep-ensemble proposal obtained the highest results.

Thus, the validation of the results showed the validity of the trained models and the deep-ensemble by applying them to data sets other than the training data set and resulted in extremely high accuracies.

Concerning sensitivity and specificity, overall, most values are between 90 and 100%.

Models trained on OASIS were more accurate in classifying healthy cases than dementia cases, while some models trained on KAGGLE had higher sensitivity values than specificity values. The KAGGLE models trained on OASIS in axial sections were the least reliable in classifying positive cases, with sensitivity values between 33% and 66%. In contrast, models trained on ADNI were the most accurate in classifying positive and negative cases, with sensitivity and specificity values equal to 100%.

As far as state of the art is concerned, this work has succeeded in making an important contribution to it. For binary classification on OASIS, the best result obtained in terms of accuracy was 99.44%, which is higher than the best-reported result of Hon et al. [60] (96.25%). Regarding multiclass classification on OASIS, the best result was obtained by Mehmood et al. [69] (99.05%). The comparison, in this case, is not straightforward as the portion of the data set used is not specified. The present report made use of a substantially more significant number of images. However, in the case of the sagittal section, the best result obtained here was 100%; in the case of the frontal section, 91.30%; in both sections combined 94.17% was obtained. For the ADNI data set, Wang et al. [88] performed a multiclass classification on 355 images, obtaining a maximum accuracy of 98.88%. Our study achieved the classification on an initial set of 349 images and got the best results with ResNet-101 (100%) and GoogLeNet (97.14%).

Finally, regarding the cross experiments, our deep-ensemble solution reached the highest performance either in the cross-data set one, with a 96.02% accuracy, and in the fMRI case, with a 98.16 accuracy.

5. Conclusions

In this work, we investigated different deep learning techniques for AD diagnosis, employing transfer learning strategies to detect AD disease and differentiate different degrees of dementia, using distinct and heterogeneous data sets. The experimental results have allowed us to identify the most promising and performing networks for the addressed problems. Among the CNNs considered, also supported by comparisons with state of the art, we can select the ResNet-50 and ResNet-101 models as the most suitable solution to be used by transfer learning, both for binary and multiclass tasks without having to design ad hoc networks. However, considering the general scenario, we demonstrated that on three different MRI data sets, AlexNet, ResNet-101, and Inception-ResNet-v2 are suitable for the addressed issue. Considering their high performance, indeed, we exploited their features in combination with an ensemble bagging classifier, bringing significant improvements in every single experiment. Moreover, this performance motivates us to explore the behaviour of our proposed solution firstly in a cross-data set scenario and secondly in a data set composed of fMRIs, different biomarkers than MRIs. The cross-data set experiment permits us to demonstrate the robustness of the proposed method when the target domain is different from the source domain. At the same time, the investigation with fMRI biomarkers showed the method's adaptability with different data sources. In both cases, the AD class reached high accuracy, with a reduced error margin. However, the MCI class is undoubtedly the most difficult to classify, with lower accuracy than the AD class. This problem is certainly due to the heterogeneity of the images representing the class in the two data sets exploited in the cross-data set experiments. At the same time, the heterogeneity of the four different MCI classes considered as grouped inside MCI, in the case of fMRIs, motivates this result.

As a general rule, when deep and machine learning models are combined, both binary and multiclass classification can generally benefit, and in some cases, remarkably.

Although the results obtained from our extensive experimentation and state-of-art comparisons on using deep learning techniques for AD diagnosis are more than satisfactory, we believe that research needs further developments before testing the models on real-world diagnoses.

As a future direction, we aim to further improve the results obtained by investigating the possibility of combining the HC and CNN features, particularly to overcome the difficulties in recognising some classes and a feature selection step to reduce the dimensionality of the features.

These could include a further refinement of the models, achieved by improving sensitivity and specificity where accuracy is high or testing them on new brain sections, such as the coronal.

Moreover, an in-depth analysis of the cross-transfer learning work could be done by fine-tuning on a different data set than the first training. As an example, a possible cross could be between OASIS and ADNI in both directions.

Also, new pre-trained models for standard transfer learning could be tested, as well as new biomarkers (e.g., PET), motivated by the excellent results obtained with fMRI ones.

Finally, we think it is important to address the problem of unbalanced classes. There is a high disproportionate ratio of observations in some data sets in each category: in the problem managed, there are many more healthy cases than AD cases. Consequently, this issue could produce models with poor predictive performance for the minority classes. This motivated us to compute several appropriate metrics for class imbalance problems in the multiclass scenarios presented on the three MRI data sets and avoid the class imbalance by oversampling the poor classes thanks to augmentation strategies.

In conclusion, our trial confirms that deep learning and transfer learning techniques can be valuable tools to detect AD from medical images. In particular, the deep-ensemble strategy we proposed can provide important indications in cross-data sets environments and with different biomarkers. However, there is still a long way to go before deep learning techniques can be used without medical supervision to accurately detect AD. Although the available computer-aided systems can still not entirely replace a medical expert, they can already provide supporting information to improve the accuracy of clinical decisions, with a considerable potential benefit for patient management.

Acknowledgments

The Regione Autonoma della Sardegna partially supported the research in this paper with the research project "Algorithms and Models for Imaging Science [AMIS], grant id: RASSR57257" (finanziato con risorse FSC 2014–2020 Patto per lo Sviluppo della Regione Sardegna).

List of Acronyms

Acc	Accuracy
AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
AE	Auto-Encoder
CDR	Clinical Dementia Rating
CNN	Convolutional Neural Network
DTI	Diffusion Tensor Imaging
MAvA	Macro Average Arithmetic
MAvG	Macro Average Geometric
MCI	Mild Cognitive Impairment
MFM	Mean F-measure
MRI	Magnetic Resonance Imaging
NC	Normal Control
OASIS	Open Access Series of Imaging Studies
PET	Positron Emission Tomography
Pre	Precision
Sen	Sensitivity
Spe	Specificity

References

[1] M. Dadar, T.A. Pascoal, S. Manitsirikul, K. Misquitta, V.S. Fonov, M.C. Tartaglia, J. Breitner, P. Rosa-Neto, O.T. Carmichael, C. Decarli, et al., Validation of a

regression technique for segmentation of white matter hyperintensities in alzheimer's disease, IEEE Trans. Med. Imag. 36 (8) (2017) 1758–1768.
[2] G. Livingston, A. Sommerlad, V. Orgeta, S.G. Costafreda, J. Huntley, D. Ames, C. Ballard, S. Banerjee, A. Burns, J. Cohen-Mansfield, et al., Dementia prevention, intervention, and care, Lancet 390 (10113) (2017) 2673–2734.
[3] A. Association, 2020 alzheimer's disease facts and figures, Alzheimer's Dementia 16 (3) (2020) 391–460.
[4] M. Liu, F. Li, H. Yan, K. Wang, Y. Ma, L. Shen, M. Xu, A.D.N. Initiative, et al., A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in alzheimer's disease, Neuroimage 208 (2020) 116459.
[5] M. Symms, H.R. Jäger, K. Schmierer, T.A. Yousry, A review of structural magnetic resonance neuroimaging, J. Neurol. Neurosurg. Psychiatr. 75 (9) (2004) 1235–1244, https://doi.org/10.1136/jnnp.2003.032714.
[6] S.A. Huettel, A.W. Song, G. McCarthy, et al., Functional Magnetic Resonance Imaging, vol. 1, Sinauer Associates Sunderland, MA, 2004.
[7] A. Nordberg, J.O. Rinne, A. Kadir, B. Långström, The use of pet in alzheimer disease, Nat. Rev. Neurol. 6 (2) (2010) 78–87.
[8] G.B. Frisoni, M. Boccardi, F. Barkhof, K. Blennow, S. Cappa, K. Chiotis, J.-F. Démonet, V. Garibotto, P. Giannakopoulos, A. Gietl, et al., Strategic roadmap for an early diagnosis of alzheimer's disease based on biomarkers, Lancet Neurol. 16 (8) (2017) 661–676.
[9] P. Padilla, M. Lopez, J.M. Gorriz, J. Ramirez, D. Salas-Gonzalez, I. Alvarez, Nmf-svm based cad tool applied to functional brain images for the diagnosis of alzheimer's disease, IEEE Trans. Med. Imag. 31 (2) (2012) 207–216.
[10] A. Ortiz, J.M. Górriz, J. Ramírez, F. Martínez-Murcia, Lvq-svm based cad tool applied to structural mri for the diagnosis of the alzheimer's disease, Pattern Recogn. Lett. 34 (14) (2013) 1725–1733 (innovative Knowledge Based Techniques in Pattern Recognition).
[11] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: P.L. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a Meeting Held December, vols. 3–6, 2012, pp. 1106–1114, 2012, Lake Tahoe, Nevada, United States.
[12] A.S. Lundervold, A. Lundervold, An overview of deep learning in medical imaging focusing on mri, Z. Med. Phys. 29 (2) (2019) 102–127 (special Issue: Deep Learning in Medical Physics).
[13] C. Di Ruberto, A. Loddo, G. Puglisi, Blob detection and deep learning for leukemic blood image analysis, Appl. Sci. 10 (3) (2020) 1176.
[14] K. Hu, Y. Huang, W. Huang, H. Tan, Z. Chen, Z. Zhong, X. Li, Y. Zhang, X. Gao, Deep supervised learning using self-adaptive auxiliary loss for covid-19 diagnosis from imbalanced ct images, Neurocomputing 458 (2021) 232–245.
[15] R.J.G. van Sloun, R. Cohen, Y.C. Eldar, Deep learning in ultrasound imaging, Proc. IEEE 108 (1) (2020) 11–29.
[16] E. Çallı, E. Sogancioglu, B. van Ginneken, K.G. van Leeuwen, K. Murphy, Deep learning for chest x-ray analysis: a survey, Med. Image Anal. 72 (2021) 102125.
[17] L. Shen, L.R. Margolies, J.H. Rothstein, E. Fluder, R. McBride, W. Sieh, Deep learning to improve breast cancer detection on screening mammography, Sci. Rep. 9 (1) (2019) 1–12.
[18] K.R. Gray, P. Aljabar, R.A. Heckemann, A. Hammers, D. Rueckert, Random forest-based similarity measures for multi-modal classification of alzheimer's disease, Neuroimage 65 (2013) 167–175.
[19] C. Zhang, E. Adeli, T. Zhou, X. Chen, D. Shen, Multi-layer multi-view classification for alzheimer's disease diagnosis, in: S.A. McIlraith, K.Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, AAAI Press, Louisiana, USA, 2018, pp. 4406–4413. February 2-7, 2018.
[20] B. Lei, P. Yang, T. Wang, S. Chen, D. Ni, Relational-regularized discriminative sparse learning for alzheimer's disease diagnosis, IEEE Transac. Cybernet. 47 (4) (2017) 1102–1113.
[21] N. Yamanakkanavar, J. Y. Choi, B. Lee, Mri segmentation and classification of human brain using deep learning for diagnosis of alzheimer's disease: a survey, Sensors 20 (11).
[22] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, D. Feng, Early diagnosis of alzheimer's disease with deep learning, in: IEEE 11th International Symposium on Biomedical Imaging, vol. 2014, ISBI), 2014, pp. 1015–1018, https://doi.org/10.1109/ISBI.2014.6868045.
[23] J. Shi, X. Zheng, Y. Li, Q. Zhang, S. Ying, Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of alzheimer's disease, IEEE J. Biomed. Health Inf. 22 (1) (2018) 173–183, https://doi.org/10.1109/JBHI.2017.2655720.
[24] A. Payan, G. Montana, Predicting Alzheimer's Disease: a Neuroimaging Study with 3d Convolutional Neural Networks, 2015 arXiv:1502.02506.
[25] M. Liu, J. Zhang, E. Adeli, D. Shen, Landmark-based deep multi-instance learning for brain disease diagnosis, Med. Image Anal. 43 (2018) 157–168, https://doi.org/10.1016/j.media.2017.10.005. URL, https://www.sciencedirect.com/science/article/pii/S1361841517301524.
[26] S. Korolev, A. Safiullin, M. Belyaev, Y. Dodonova, Residual and plain convolutional neural networks for 3d brain MRI classification, in: 14th IEEE International Symposium on Biomedical Imaging, ISBI 2017, Melbourne, Australia, April 18-21, 2017, IEEE, vol. 2017, pp. 835–838. doi:10.1109/ISBI.2017.7950647.
[27] J. Islam, Y. Zhang, Brain mri analysis for alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks, Brain Informatics 5 (2). doi:10.1186/s40708-018-0080-3.
[28] D. Lu, K. Popuri, G. W. Ding, R. Balachandar, M. F. Beg, Multimodal and multiscale deep neural networks for the early diagnosis of alzheimer's disease using structural mr and fdg-pet images, Sci. Rep.:10.1038/s41598-018-22871-z.
[29] C. Lian, M. Liu, J. Zhang, D. Shen, Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri, IEEE Trans. Pattern Anal. Mach. Intell. 42 (4) (2020) 880–893, https://doi.org/10.1109/TPAMI.2018.2889096.
[30] B. Lee, W. Ellahi, J.Y. Choi, Using deep cnn with data permutation scheme for classification of alzheimer's disease in structural magnetic resonance imaging (smri), IEICE Transac. Inf. Syst. (7) (2019) 1384–1395, https://doi.org/10.1587/transinf.2018EDP7393.
[31] N.M. Khan, N. Abraham, M. Hon, Transfer learning with intelligent training data selection for prediction of alzheimer's disease, IEEE Access 7 (2019) 72726–72735, https://doi.org/10.1109/ACCESS.2019.2920448.
[32] S. Ahmed, K.Y. Choi, J.J. Lee, B.C. Kim, G. Kwon, K.H. Lee, H.Y. Jung, Ensembles of patch-based classifiers for diagnosis of alzheimer diseases, IEEE Access 7 (2019) 73373–73383, https://doi.org/10.1109/ACCESS.2019.2920011.
[33] T.E. Kam, H. Zhang, Z. Jiao, D. Shen, Deep learning of static and dynamic brain functional networks for early mci detection, IEEE Trans. Med. Imag. 39 (2) (2020) 478–487, https://doi.org/10.1109/TMI.2019.2928790.
[34] A. Ortiz, J. Munilla, J.M. Górriz, J. Ramírez, Ensembles of deep learning architectures for the early diagnosis of the alzheimer's disease, Int. J. Neural Syst. 26 (2016) 1650025, https://doi.org/10.1142/S0129065716500258, 07.
[35] K. Aderghal, J. Benois-Pineau, K. Afdel, C. Gwenaëlle, Fuseme: classification of smri images by fusion of deep cnns in 2d+ϵ projections, in: Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing, No. 34 in CBMI '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1–7.
[36] A. Khvostikov, K. Aderghal, J. Benois-Pineau, A. Krylov, G. Catheline, 3d Cnn-Based Classification Using Smri and Md-Dti Images for Alzheimer Disease Studies, 2018 arXiv:1801.05968.
[37] K. Aderghal, A. Khvostikov, A. Krylov, J. Benois-Pineau, K. Afdel, G. Catheline, Classification of alzheimer disease on imaging modalities with deep cnns using cross-modal transfer learning, in: IEEE 31st International Symposium on Computer-Based Medical Systems, vol. 2018, CBMS), 2018, pp. 345–350, https://doi.org/10.1109/CBMS.2018.00067.
[38] R. Cui, M. Liu, Hippocampus analysis by combination of 3-d densenet and shapes for alzheimer's disease diagnosis, IEEE J. Biomed. Health Inf. 23 (5) (2019) 2099–2107, https://doi.org/10.1109/JBHI.2018.2882392.
[39] E. Hosseini-Asl, R. Keynton, A. El-Baz, Alzheimer's disease diagnostics by adaptation of 3d convolutional network, in: IEEE International Conference on Image Processing, vol. 2016, ICIP), 2016, pp. 126–130.
[40] S. Sarraf, G. Tofighi, Classification of Alzheimer's Disease Using Fmri Data and Deep Learning Convolutional Neural Networks, 2016 arXiv:1603.08631.
[41] H.-I. Suk, S.-W. Lee, D. Shen, Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis, Neuroimage 101 (2014) 569–582, https://doi.org/10.1016/j.neuroimage.2014.06.077. URL, https://www.sciencedirect.com/science/article/pii/S1053811914005540.
[42] M. Liu, D. Cheng, K. Wang, Y. Wang, Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis, Neuroinformatics:10.1007/s12021-018-9370-4.
[43] C. Feng, A. Elazab, P. Yang, T. Wang, F. Zhou, H. Hu, X. Xiao, B. Lei, Deep learning framework for alzheimer's disease diagnosis via 3d-cnn and fsbi-lstm, IEEE Access 7 (2019) 63605–63618, https://doi.org/10.1109/ACCESS.2019.2913847.
[44] X. Fang, Z. Liu, M. Xu, Ensemble of deep convolutional neural networks based multi-modality images for alzheimer's disease diagnosis, IET Image Process. 14 (2) (2020) 318–326, https://doi.org/10.1049/iet-ipr.2019.0617.
[45] A. Abdullah, Alzheimer MRI Dataset, 2020, https://doi.org/10.1162/jocn.2007.19.9.1498. URL, https://www.oasis-brains.org/.
[46] D.S. Marcus, T.H. Wang, J. Parker, J.G. Csernansky, J.C. Morris, R.L. Buckner, Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults, J. Cognit. Neurosci. 19 (9) (2007) 1498–1507.
[47] A. Abdullah, Alzheimer MRI dataset, URL, https://www.kaggle.com/legendahmed/alzheimermridataset/metadata, 2020.
[48] U. L. of, Neuro imaging, ADNI alzheimer's disease neuroimaging initiative, URL, http://adni.loni.usc.edu, 2017.
[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2009, Ieee, 2009, pp. 248–255.
[50] H. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R. M. Summers, Deep convolutional neural networks for computer-aided detection: cnn architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imag. 35 (5) (2016) 1285–1298, https://doi.org/10.1109/TMI.2016.2528162.
[51] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, vol. 2016, IEEE Computer Society, NV, USA, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90. June 27-30, 2016.
[52] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S.E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, IEEE Computer Society, Boston, MA, USA, 2015, pp. 1–9. June 7-12, 2015.
[53] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the Thirty-First

AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, pp. 4278–4284.
[54] M. Baygin, S. Dogan, T. Tuncer, P. Datta Barua, O. Faust, N. Arunkumar, E. W. Abdulhay, E. Emma Palmer, U. Rajendra Acharya, Automated asd detection using hybrid deep lightweight features extracted from eeg signals, Comput. Biol. Med. 134 (2021) 104548.
[55] S. Alinsaif, J. Lang, 3d shearlet-based descriptors combined with deep features for the classification of alzheimer's disease based on mri data, Comput. Biol. Med. 138 (2021) 104879.
[56] H. Li, X. Wang, C. Liu, P. Li, Y. Jiao, Integrating multi-domain deep features of electrocardiogram and phonocardiogram for coronary artery disease detection, Comput. Biol. Med. 138 (2021) 104914.
[57] A. Narin, Accurate detection of covid-19 using deep features based on x-ray images and feature selection methods, Comput. Biol. Med. 137 (2021) 104771.
[58] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[59] R. Alejo, J.A. Antonio, R.M. Valdovinos, J.H. Pacheco-Sánchez, Assessments metrics for multi-class imbalance learning: a preliminary study, in: J.A. Carrasco-Ochoa, J.F. Martínez-Trinidad, J.S. Rodríguez, G.S. di Baja (Eds.), Pattern Recognition, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 335–343.
[60] M. Hon, N.M. Khan, Towards alzheimer's disease classification through transfer learning, in: IEEE International Conference on Bioinformatics and Biomedicine, vol. 2017, BIBM), 2017, pp. 1166–1169, https://doi.org/10.1109/BIBM.2017.8217822.
[61] M. Faturrahman, I. Wasito, N. Hanifah, R. Mufidah, Structural mri classification for alzheimer's disease detection using deep belief network, in: 2017 11th International Conference on Information Communication Technology and System, ICTS), 2017, pp. 37–42, https://doi.org/10.1109/ICTS.2017.8265643.
[62] D. Jha, G.-R. Kwon, Alzheimer's disease detection using sparse autoencoder, scale conjugate gradient and softmax output layer with fine tuning, Int. J. Machine Learn. Comput. 7 (1) (2017) 13–17, https://doi.org/10.18178/ijmlc.2017.7.1.612. URL, http://www.ijmlc.org/index.php?m=content&c=index&a=show&catid=69&id=705.
[63] J.M. Ortiz-Suárez, R. Ramos-Pollán, E. Romero, Exploring Alzheimer's anatomical patterns through convolutional networks, in: 12th International Symposium on Medical Information Processing and Analysis, vol. 10160, International Society for Optics and Photonics, 2017, p. 101600Z.
[64] S.H. Wang, P. Phillips, Y. Sui, B. Liu, M. Yang, H. Cheng, Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling, J. Med. Syst. 42 (5) (2018) 85.
[65] M. Maqsood, F. Nazir, U. Khan, F. Aadil, H. Jamal, I. Mehmood, O. Y. Song, Transfer learning assisted classification and detection of alzheimer's disease stages using 3D MRI scans, Sensorsdoi:10.3390/s19112645.
[66] J. Islam, Y. Zhang, A novel deep learning based multi-class classification method for alzheimer's disease detection using brain mri data, in: Y. Zeng, Y. He, J. H. Kotaleski, M. Martone, B. Xu, H. Peng, Q. Luo (Eds.), Brain Informatics=, Springer International Publishing, Cham, 2017, pp. 213–222.
[67] J. Islam, Y. Zhang, An Ensemble of Deep Convolutional Neural Networks for Alzheimer's Disease Detection and Classification, 2017 arXiv:1712.01675.
[68] J. Islam, Y. Zhang, Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks, Brain Inf. 10.1186/s40708-018-0080-3.
[69] A. Mehmood, M. Maqsood, M. Bashir, Y. Shuyuan, A deep siamese convolution neural network for multi-class classification of alzheimer disease, Brain Sci.: 10.3390/brainsci10020084.
[70] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, D. Feng, Early diagnosis of alzheimer's disease with deep learning, in: IEEE 11th International Symposium on Biomedical Imaging, vol. 2014, ISBI), 2014, pp. 1015–1018, https://doi.org/10.1109/ISBI.2014.6868045.
[71] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M. J. Fulham, Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer's disease, IEEE Trans. Biomed. Eng.:10.1109/TBME.2014.2372011.
[72] H. Karasawa, C.L. Liu, H. Ohwada, Deep 3D convolutional neural network architectures for alzheimer's disease diagnosis, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, pp. 287–296.
[73] S. Basaia, F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo, M. Filippi, Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks, Neuroimage: Clinic 21 (2019) 101645.
[74] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu, T.A. D.N. Initiative, Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment, Front. Neurosci. 12 (2018) 777, https://doi.org/10.3389/fnins.2018.00777. URL, https://www.frontiersin.org/article/10.3389/fnins.2018.00777.
[75] C.D. Billones, O.J.L.D. Demetria, D.E.D. Hostallero, P.C. Naval, Demnet, A convolutional neural network for the detection of alzheimer's disease and mild cognitive impairment, in: IEEE Region 10 Conference, vol. 2016, TENCON), 2016, pp. 3724–3727, https://doi.org/10.1109/TENCON.2016.7848755.
[76] A. Valliani, A. Soni, Deep residual nets for improved Alzheimer's diagnosis, in: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2017, 615–615.
[77] T. Zhou, K.H. Thung, X. Zhu, D. Shen, Feature learning and fusion of multimodality neuroimaging and genetic data for multi-status dementia diagnosis, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017, pp. 132–140.
[78] K.H. Thung, P.T. Yap, D. Shen, Multi-stage diagnosis of Alzheimer's disease with incomplete multimodal data via multi-task deep learning, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017, pp. 160–168.
[79] K.A. Gunawardena, R.N. Rajapakse, N.D. Kodikara, Applying convolutional neural networks for pre-detection of Alzheimer's disease from structural MRI data, in: 2017 24th International Conference on Mechatronics and Machine Vision in Practice vol. 2017, M2VIP, 2017, pp. 1–7.
[80] C.V. Dolph, M. Alam, Z. Shboul, M.D. Samad, K.M. Iftekharuddin, Deep learning of texture and structural features for multiclass Alzheimer's disease classification, in: Proceedings of the International Joint Conference on Neural Networks, 2017, pp. 2259–2266.
[81] D. Collazos-Huertas, A. Tobar-Rodriguez, D. Cárdenas-Peña, G. Castellanos-Dominguez, Mri-based feature extraction using supervised general stochastic networks in dementia diagnosis, in: Natural and Artificial Computation for Biomedicine and Neuroscience, Springer International Publishing, Cham, 2017, pp. 363–373.
[82] T.D. Vu, N.H. Ho, H.J. Yang, J. Kim, H.C. Song, Non-white matter tissue extraction and deep convolutional neural network for Alzheimer's disease detection, Soft Computing 22 (20) (2018) 6825–6833.
[83] S. Esmaeilzadeh, D.I. Belivanis, K.M. Pohl, E. Adeli, End-to-end alzheimer's disease diagnosis and biomarker identification, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, pp. 337–345.
[84] H. Tang, E. Yao, G. Tan, X. Guo, A fast and accurate 3D fine-tuning convolutional neural network for alzheimer's disease diagnosis, in: Communications in Computer and Information Science, 2018, pp. 115–126.
[85] V. Wegmayr, S. Aitharaju, J. Buhmann, Classification of brain MRI with big data and deep 3D convolutional neural networks, in: Medical Imaging 2018: Computer-Aided Diagnosis, vol. 10575, 2018, p. 105751S.
[86] T. Zhou, K. H. Thung, X. Zhu, D. Shen, Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis, Hum. Brain Mapp.:10.1002/hbm.24428.
[87] R. Jain, N. Jain, A. Aggarwal, D. J. Hemanth, Convolutional neural network based Alzheimer's disease classification from magnetic resonance brain images, Cognit. Syst. Res.:10.1016/j.cogsys.2018.12.015.
[88] H. Wang, Y. Shen, S. Wang, T. Xiao, L. Deng, X. Wang, X. Zhao, Ensemble of 3D Densely Connected Convolutional Network for Diagnosis of Mild Cognitive Impairment and Alzheimer's Disease, Neurocomputing.
[89] S. Sarraf, G. Tofighi, Deep learning-based pipeline to recognize alzheimer's disease using fmri data, in: Proceedings of the Future Technologies Conference, FTC), 2016, pp. 816–820.