Neural Computing and Applications (2024) 36:2653–2664
https://doi.org/10.1007/s00521-023-09163-y

ORIGINAL ARTICLE

Dementia classification from magnetic resonance images by machine learning

Georgina Waldo-Benítez[1] · Luis Carlos Padierna[1] · Pablo Ceron[2] · Modesto A. Sosa[1]

Received: 1 March 2023 / Accepted: 20 October 2023 / Published online: 22 November 2023

Abstract

Dementia is a threatening condition that affects communication, thinking, and memory skills, being Alzheimer its most common type. The early detection of this disease allows for better care of the patient. Recently, Machine Learning (ML) methods have been developed to support the finding and forecast of Alzheimer's disease through the analysis of Magnetic Resonance Images (MRI). Existing ML methods present some limitations: (i) require an expert to extract relevant features from MRI, (ii) depend on multistep image preprocessing, or (iii) need complex architectures and several images to train them. To surpass these limitations, in the present work, we analyze different Convolutional Neural Networks (CNNs) for Alzheimer's classification, formulated to learn from a set of representative MRI sagittal images available in the Open Access Series of Imaging Studies (OASIS-2, 72 non-demented and 64 demented subjects, with ages from 60 to 96 years) and the Alzheimer's Disease Neuroimaging Initiative (ADNI, 200 early Alzheimer and 200 control patients, with ages from 55 to 90 years) datasets. All CNNs were compared with state-of-the-art ML methods, being the VGG-16 variant the best performed architecture with an average validation accuracy of 56% $\pm$ 4%, evaluated with a bootstrapping strategy to measure the variability on independent runs. This result confirms the best performance reported so far ($< 60\%$) with different ML methods. The low accuracy evidences the hardness of the problem and contrasts with the higher accuracy levels (up to 97%) reached with preprocessed and well-characterized MRI axial images from the OASIS-1 or ADNI-2 datasets. Thus, opening an interesting discussion about what MRI plane should be considered when training CNNs for Alzheimer's classification, and leaving a wide room for improvement on the performance of CNNs trained with sagittal MRI images. The resulting model implemented in software and experimental data are publicly available.

Keywords Convolutional neural network · Alzheimer classification · Machine learning · Brain MRI · OASIS-2

1 Introduction

Dementia is a dangerous condition characterized by affecting communication, thinking, and memory skills. It affects over 50 million people around the world, with nearly 60% living in developing nations. Ten million new cases arise each year and there is no treatment to cure this disorder or to modify its progression [1]. Magnetic Resonance Imaging (MRI) data allow us to understand the inner functioning of the human brain. In MRI studies, radio waves along with a magnet are employed to generate brain images. Normal and diseased tissue can be discriminated from these images [2, 3]; thus, MRI data can support the finding and forecast of dementia. The early detection of dementia allows us to identify the initial state of the disease and to act for better care of the patient [4]. The screening process is a burden to neuroradiologists and depends on their experience, which is why there is a need for computer-assisted tools that increase the reliability of the diagnosis. Moreover, cognitive tests, evaluate patients who present dementia symptoms, while MRI can provide structural markers such as the reduction of the hippocampus or brain atrophy before these symptoms appear [5].

✉ Luis Carlos Padierna
lc.padierna@ugto.mx

[1] División de Ciencias de Ingenierías, Universidad de Guanajuato, 103 Lomas del Campestre Street, 37150 León, Mexico

[2] División de Ciencias, Ingeniería y Tecnología, Universidad de Quintana Roo, 77019 Chetumal, Mexico

The problem of classifying dementia from MRI data has been addressed since 2008 with the Open Access Series of Imaging Studies (OASIS) project, which provides neuroimaging public data for scientific purposes [6]. This dataset has been used for the 3D modeling of some regions of the brain [7], subjective age estimation based on brain age [8], long-term chronological change in disease progression [9], and prediction of brain connectivity from a single timepoint [10]. ADNI is a consortium of universities and medical centers in the United States and Canada established to develop standardized imaging techniques and biomarker procedures in normal subjects, subjects with mild Alzheimer's Disease (AD), and subjects with AD [11]. Recently developed computer-aided tools based on Machine Learning (ML) algorithms have reduced neuroradiologists' workload and improved diagnosis accuracy by doing an automated classification, using only MRI data without any other patient information [12–22]. These algorithms, including both classical ML and Deep Learning (DL) methods, have addressed the Alzheimer's classification problem based on the OASIS dataset, mainly axial plane and demographic data, with accuracies ranging from 68 to 98%. However, classical ML algorithms need training data with features selected by a specialist, and current DL architectures are complex, overspecialized, and require intricate image preprocessing. Furthermore, the sagittal plane of MRI images has been scarcely studied, missing the opportunity of finding patterns of neurological atrophy, such as the reduction of the hippocampus, which is not observable from the axial plane.

The main objective of this work is to find the best Convolutional Neural Network (CNN) architecture for Alzheimer's classification from sagittal MRI images. This architecture must fulfill the following requirements: i) be formulated to learn from a set of representative MRI images, ii) do not require geographic data, iii) do not need feature selection performed by specialists and iv) minimize both image preprocessing and the number images for training. If these requirements are met, a worthy computer-aid diagnosis tool would be available for practitioners, complementing the early detection of the disease.

2 Methods

2.1 Datasets, subjects and MRI acquisitions

The OASIS datasets provide open access to neurological images for clinical and cognitive research [23]. There are four datasets, namely, OASIS-1, OASIS-2, OASIS-3, and OASIS-4. The datasets OASIS-1 (axial) and OASIS-2 (sagittal) have been employed for segmentation algorithms, hypothesis-based analyzes, and beyond. OASIS-3 is useful as a longitudinal neuroimaging and cognitive dataset for the study of Alzheimer's and other diseases. OASIS-4 is the latest release in the OASIS project, it contains MR, clinical, cognitive, and biomarker data for individuals that present memory complaints.

In this work, we focus on the OASIS-2 dataset, which is made up of a collection of 150 right-handed patients (72 non-demented, 14 converted, and 64 demented) with ages ranging from 60–96 years. Each patient was explored on at least two visits, with one year or more between visits, obtaining 373 imaging scans. All data were acquired using the same procedure and scanner. Three to four T1-weighted individual images were acquired for each patient by a 1.5 T Vision scanner. Only the first T1-weighted image was used for experiments. A thermoplastic face mask and cushioning were employed to minimize head movement. The type of scanner and the number of images was maintained during the whole study. A sample of the available MRI images is illustrated in Fig. 1. ADNI-1 dataset is similar to OASIS-2, both of them provide the sagittal plane of MRI images. ADNI consists of 400 subjects diagnosed with mild cognitive impairment (MCI), 200 subjects with early AD, and 200 elderly control subjects with ages ranging from 55 to 90 years [11].

2.2 Preprocessing and preliminary analysis of the OASIS-2 clinical records

Feature Selection: based on correlation analysis, seven features were selected as input for the classical ML models: (1) age, (2) sex, (3) education level, (4) Clinical Dementia Rating (CDR), (5) Mini-mental State Examination score (MMSE), (6) normalized whole brain volume (nWBV) and (7) Atlas Scaling Factor (ASF). These features were also considered in previous works [15], but ASF was used instead of the Estimated Total Intracranial Volume (eTIV). Delay, visit, and Socio-Economic Status (SES) were eliminated for not being relevant characteristics for the model. The eTIV was omitted because it is strongly correlated with the ASF.

2.3 Classical machine learning methods

To have a basis for comparison with our proposal, various classic ML algorithms were implemented, and are briefly defined below.

(a) Support vector machines (SVMs). It is an ML model useful for classification that finds an optimal hyperplane to discriminate between two classes by transforming the classification problem into a quadratic optimization one as follows [24, 25]: Starting from a set of n sample points $D = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$, and considering $\mathbf{x}_i \in \mathbb{R}^d$ as the i-th training

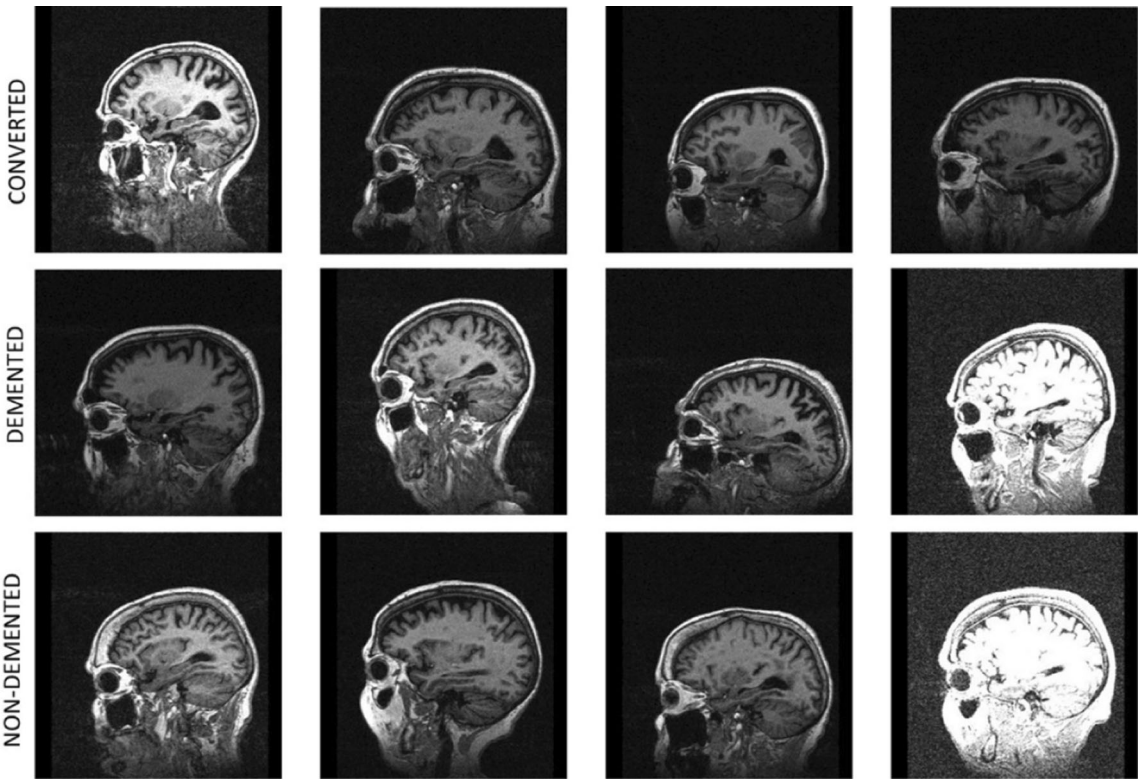

Fig. 1 Representative images of the OASIS-2 dataset for Alzheimer's classification, showing examples of the three classes: demented, non-demented and converted. The ADNI dataset only involves demented and non-demented classes

vector along with $y_i \in \{-1, +1\}$ as its category label; a SVM algorithm is stated as:

$$MaxL(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \mathrm{K}(\mathbf{x}_i, \mathbf{z}_j) \tag{1}$$

$$s.t. \quad C \geq \alpha_i \geq 0 \quad \forall i = 1, \ldots, n \sum_{i=1}^{n} \alpha_i y_i = 0$$

where $\boldsymbol{\alpha}$ are the Lagrange multipliers, C is a penalty hyperparameter, $\mathbf{x}_i, \mathbf{z}_j \in \mathbb{R}^d$ are two given training vectors; and $\mathrm{K}(\mathbf{x}, \mathbf{z})$ in $\mathbb{R}^d \times \mathbb{R}^d$ is a kernel function $\mathrm{K}(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$ [26].

(b) Naïve Bayes (NB). This classifier predicts a label $y \in \{0, 1\}$ based on a feature vector $\mathbf{x} = (x_1, \ldots, x_d)$, where each $x_i \in \{0, 1\}$, the learning algorithm determines the probability of being part of one class taking the maximum frequency of attribute identified in the training dataset [27].

(c) k -Nearest Neighbors (k-NN). It is a distance-based classifier that identifies the k closest points (neighbors) to the instance being evaluated, where the labels of these neighbors decide the predicted label [28]. The k points in D that are closest to the evaluated $\mathbf{x}$ are selected. This estimator counts the number of points corresponding to each class (c) that are in the closest group and provides the estimate shown in Eq. 2, where $N_k(\mathbf{x}, D)$ correspond to indices of the k closest points to $\mathbf{x} \in D$, $y \in \mathbb{N}$, and $\mathbb{I}(e)$ is a function indicator that returns 1 when e is positive or 0 when e is negative.

$$p(y = c|\mathbf{x}, D, k) = \frac{1}{k} \sum_{i \in N_k(\mathbf{x}, D)} \mathbb{I}(y_i = c) \tag{2}$$

(d) Decision Tree (DT). As a classifier, the goal is to predict the label corresponding to the test $\mathbf{x}$ by processing the tree from the root to a leaf, where each intermediate node states a threshold to discriminate the instances according to a feature value [29]. This model can be written as:

$$f(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] = \sum_{m=1}^{M} \mathbf{w_m} \mathbb{I}(\mathbf{x} \in R_m) = \sum_{m=1}^{M} w_m \phi(\mathbf{x}; \mathbf{v_m}) \tag{3}$$

where R_m is defined as the m'th region, $\mathbf{w_m}$ is a class label distribution for each leaf and $\mathbf{v_m}$ represents the variable to be split. Finally, ϕ is defined as $\phi(\mathbf{x}) = [\mathrm{K}(\mathbf{x}, \mu_1), \ldots, \mathrm{K}(\mathbf{x}, \mu_N)]$ where μ_k are the complete training data or a given subset [30].

(e) Random Forest (RF): A combination of decision trees is used to build this classifier aiming to decorrelate individual trees randomly choosing a subset of input variables. The prediction of the RF is based on a voting scheme that counts the individual predictions of each tree

[27]. It is possible to take M distinct trees for training, belonging to different subgroups, so that these are selected randomly and with replacement, to calculate Eq. (4), where f_{m} is the m'th tree [30].

$$f(\mathbf{x}) = \sum_{m=1}^{M} \frac{1}{M} f_{m}(\mathbf{x}) \tag{4}$$

2.4 Convolutional neural networks

CNN is a feed-forward neural network widely used in image processing. The CNN input vectors are images structured as hypermatrices arrays from which, each part of the input image is extracted in the so-called receptive field. CNN uses a mathematical operation called convolution, denoted as $s(t) = (x * w)(t)$, where $x(t)$ and $w(t)$ are two functions. In ML applications, data are discrete, and the following discrete convolution is employed:

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a) \tag{5}$$

where x is the Input and w is the Kernel in the convolutional network terminology. In practice, both functions are zero everywhere, but the points for which the function values are stored. When working with a two-dimensional image as input I and assuming a two-dimensional kernel K is applied, (5) can be conveniently implemented as [31]:

$$S(i,j) = (I * K)(i,j) = \sum_{m} \sum_{n} I(i+m, j+n)K(m,n) \tag{6}$$

where K is a multidimensional array of parameters arranged on a regular grid with a variable number of axes [31, 32]. VGG-16 [33] and RestNet-50 [34], are representative CNN architectures, that were used to classify the ImageNet dataset in 2014 and 2015, respectively. Through the Transfer Learning strategy, it is possible to reuse the learned filters of these two architectures in new classification tasks.

2.5 Validation methods and performance metrics

To improve performance, it is necessary to find suitable hyperparameters for the algorithms. In addition, to ensure the performance reliability of the algorithms under examination, standardized methods and metrics are required.

Grid search is a process of hyperparameter tuning to decide the optimal values for a given model. It uses a loop through predefined hyperparameters and fits the model on the training set, selecting the best parameters [35].

The k-fold cross-validation strategy consists in dividing a set of training points into k subsets (folds), of the same number of elements. The first subset is used for validation, and the remaining k – 1, is used to train the method [36].

Bootstrapping refers to any test employing a random sample using replacement. It assigns accuracy values to each sample. The distribution of any statistic can be estimated by this technique [35].

Accuracy is the ratio between right predictions and the total of cases examined, Sensitivity is the number of positive cases predicted correctly, and Specificity is how well our model predict the negative cases, as stated in Eq. (7,8,9) [37]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\textit{Specificity} = \frac{TN}{TN + FP} \tag{9}$$

where TN = True negative, TP = True positive, FN = False negative and, FP = False positive.

2.6 Experimental design

The whole experimental setup is illustrated in Fig. 2. The concepts, definitions, and methods indicated in each stage were provided in previous sections. Hyperparameters and experimental strategies are described below.

a. Selection of the more relevant features of the OASIS-2 geographic information using a correlation heatmap. The resulting features were already described in Sect. 2.2.
b. A hyperparameter tuning scheme by using cross-validation and Grid-search strategy, was used to find the best hyperparameters of both classical ML classifiers and CNN architectures (excluding VGG-16 and ResNet-50, in which predefined settings and frozen convolutional layers were employed, using the implementations from Keras [38] and Tensorflow [39]. All datasets (ADNI-1 MRI images, OASIS-2 MRI images and OASIS-2 demographic data) were partitioned with a 75/25 scheme. 75% of data for training and the remaining 25% for test. This partitioning was resampled 10 times to measure variability between independent runs. Care was taken to ensure that each partition include different patients.
c. Classical ML methods (SVM, RF, DT, NB, and k-NN) and their corresponding experimental settings, were implemented for a direct comparison against both the proposed and Transfer Learning CNNs. For these ML classical methods, the best hyperparameters were used,

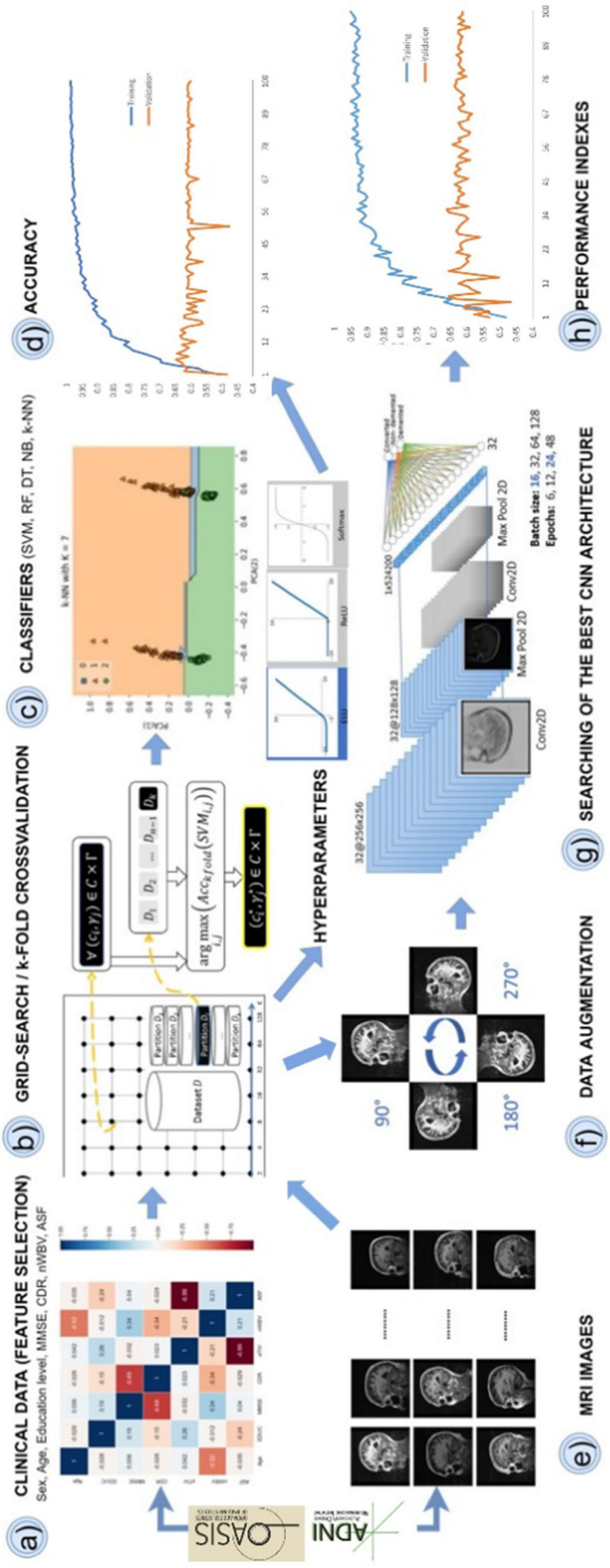

Fig. 2 Experimental setup for the search for the best CNN architecture and comparison against classical ML methods

Table 1 CNN-2D architecture and hyperparameters

Hyperparameter/component	Values
Convolution	3 × 3 Kernel
Pooling	Max pooling
CNN activation function	{ELU, ReLU, Leaky ReLU, PReLU, Softmax, Linear}
Dense layer	Linear activation function
Dense layer	SoftMax function
Batch size	{16,32,64,128}
Epochs	{6,12,24,48,100,1000}
Initialization rate	0.001
Optimizer	AdaGrad
Dropout	0.5

a SVM with radial basis function (RBF) and $\gamma = 0.4$ kernel; a RF with a max depth of 10; a DT with a max depth of 2; a gaussian NB; and a k-NN, which achieve the best accuracy, was trained by calculating the Manhattan distance and k = 7 nearest neighbors.

d. The performance of the different designed CNNs was compared in terms of accuracy against the framework of classical ML methods and against the state-of-the-art DL methods. Sensitivity and Specificity metrics were also used to evaluate all tested CNNs.

e. From the first T1 study of each patient, the 90th slice of the MRI sequence for OASIS-2 and a similar slice between 90th and 115th from ADNI-1 was selected because it shows a neat image of the brain, and also because previous studies concluded that center slices contain the most relevant information for diagnosis purposes [18, 19]. In [18] a total of 10 slices (88 ± 5) were extracted, in [19] 30 slices from 71st to 100th were selected and in [22]the 80th slice from OASIS-1 was used.

f. An augmentation technique was implemented, in which the 373 images selected were rotated at 90°, 180°, and 270°, obtaining at the final a total of 1492 images to train the different CNNs. This strategy strengthens the CNN architecture because it promotes invariance to rotation. Noise injection, gamma correction, and other augmentation methods were not considered for two reasons: i) Because the raw image in the original dataset already includes electronic noise, motion artifacts, distortions from dental work and limited image contrast presented in real environments, ii) to keep a reduced sample size.

g. In this study, we experiment with different CNN architectures, looking for the most simple and accurate one, capable of solving the Alzheimer's classification problem without the feature selection step. CNNs were designed by varying their architecture and hyperparameters, as shown in Table 1. To measure the CNN variability between independent runs, a bootstrapping method was used ten times, each time resampling with the replacement of the whole cross-validation folds. The same strategy is followed by [22].

For all the images and codes used in this study to be available, a GitHub repository was created, which can be accessed following the link https://github.com/GinaWaldo/OASIS2-CNN.git [40].

3 Results

Three groups of ML algorithms were employed to solve the MRI Alzheimer's classification problem. The first group includes classical ML classifiers. The second group consists of Deep Neural Networks with one or two convolutional layers as feature extractors and dense layers as classifiers, which were designed by varying the components and hyperparameters specified in Table 1. The last group involves pretrained CNN bases from the VGG-16 and Resnet-50 architectures as feature extractors and the same shape of dense layers as those used in the second group as classifiers. The results of these three groups of algorithms are reported in Table 2 and described below.

The group of classical ML algorithms was considered to establish a reference framework of comparison with both the designed and pretrained CNN architectures. These algorithms were trained and tested with the OASIS-2 geographic data. The best results obtained both in our experiments and in previous works are reported in Table 2. The highest average accuracy of 92.13% ± 3.48 was obtained with a k-NN algorithm.

Regarding the group of designed CNNs, multiple architectures with one or two convolutional layers were tested. The best CNN found has the architecture illustrated in Fig. 3, which consists of a single convolutional layer with a 3 × 3 kernel and a max pooling layer, also includes an ELU activation function, a flatten layer, and a dense layer of three outputs corresponding to each of the

Table 2 Performance of ML algorithms employed for Alzheimer's classification

Classical ML algorithms	Accuracy (mean ± std. dev) our experiments	Best reported [12, 13, 15, 17]
k-NN	92.13% ± 3.48	90.74%
RF	92.0% ± 1.8	96.66%[(i)], 93.56%[(ii)]
DT	90.0% ± 4.3	99.28%[(iii)]
SVM	89.9% ± 3.2	92.57%
NB	85.3% ± 3.5	87.29%

Designed CNNs	Acc	Sensitivity (mean ± std. dev)	Specificity	Best hyperparameters
CNN-1 (OASIS-2)	50.1% ± 4.1	50.1% ± 4.1	48.8% ± 10.0	Batch size = 16 Epochs = 100 AF = ELU
CNN-2 (OASIS-2)	50.5% ± 5.5	49.1% ± 3.7	49.0% ± 5.8	Batch size = 16 Epochs = 100 AF = ELU

Pretrained CNNs	Acc	Sensitivity (mean ± std. dev)	Specificity	Best hyperparameters and best reported accuracies [18, 21, 22]
VGG-16 (OASIS-2)	55.6% ± 4.2	55.6% ± 4.2	55.6% ± 4.2	Batch size = 16 Epochs = 100 AF = ELU
ResNet-50 (OASIS-2)	54.4% ± 5.8	54.4% ± 5.8	54.4% ± 5.8	Batch size = 16 Epochs = 1000 AF = ELU
BrainNet2D (OASIS-1 and 2)				88%[(iv)] AF = ReLU
ResNet-18 (OASIS-1 and 2)				89%[(iv)] AF = ReLU
ADVIAN (OASIS-1)				97.76% ± 1.13 AF = ReLU
CNN-BND (OASIS-1)				97.19% ± 0.89 Batch size = 256 AF = ReLU

Pretrained CNNs	Acc	Sensitivity (mean ± std. dev)	Specificity	Best reported accuracy mean [41]
VGG-16	56% ± 4.0	56% ± 4.0	56% ± 4.0	< 60% with different classical ML methods k-NN, RF, XGBoost, SVM, MLP and others[(v)]

All algorithms used a train/test data partition of 75/25, except for [18] and [19] that used an 80/20 (fivefold cross-validation partition). Mean ± standard deviation is computed from a bootstrapping strategy of 10 independent runs. In the hyperparameters column, AF stands for the Activation Function of the dense layer in CNNs

[(i)] The accuracy reported for RF by Shanmuga et al. [15] is the maximum obtained in their experiments. No confidence interval is given

[(ii)] Before the classification with RF, a feature selection was performed by employing the Particle Swarm Optimization (PSO) algorithm [17]

[(iii)] The accuracy reported for DT by Bansal et al. [12] is doubtful and this will be explained in the discussion section

[(iv)] No standard deviation for the accuracy, batch size nor epochs were reported by Saratxaga et al. [18] . Authors do not describe how images partitions were made. They mention that *"Only the horizontal (transverse) plane images have been used for training and testing the models."* And that "*OASIS−2 has been used for validation purposes of the proposed approach."* Therefore, the high accuracies reported are associated to the training with OASIS−1

[(v)] The work of Balansundaram et al. 2023 [41] , is the only work found in the literature that experimented with MRI images of OASIS−2. No average accuracy was reported, the performance of the algorithms was presented in Figs. 6 and 7, Sects. 5.1.1 and 5.1.2 in that work. [5]

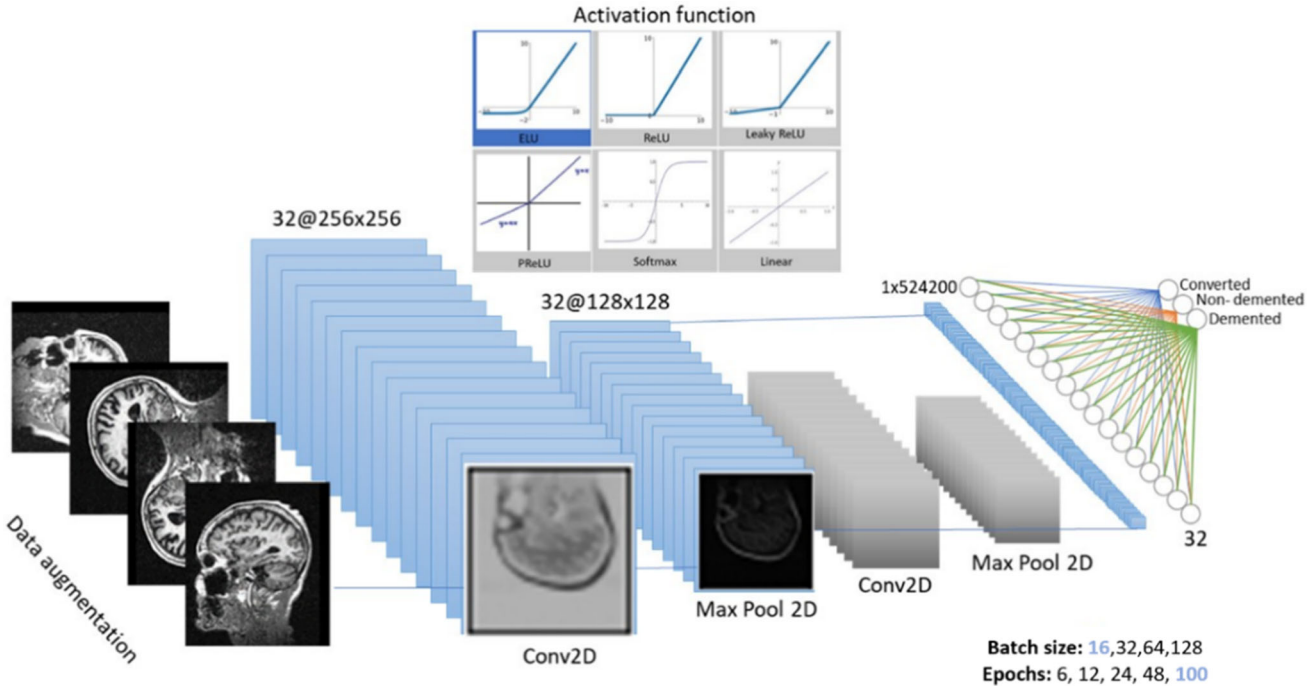

Fig. 3 CNN-2D architectures designed for dementia MRI classification problem. The elements highlighted in blue show the best architecture found: Conv2D-ELU-MaxPooling2D-Dropout-Flatten-Dense-ELU-Dropout-Dense. The best hyperparameters were batch size = 16, epochs = 100

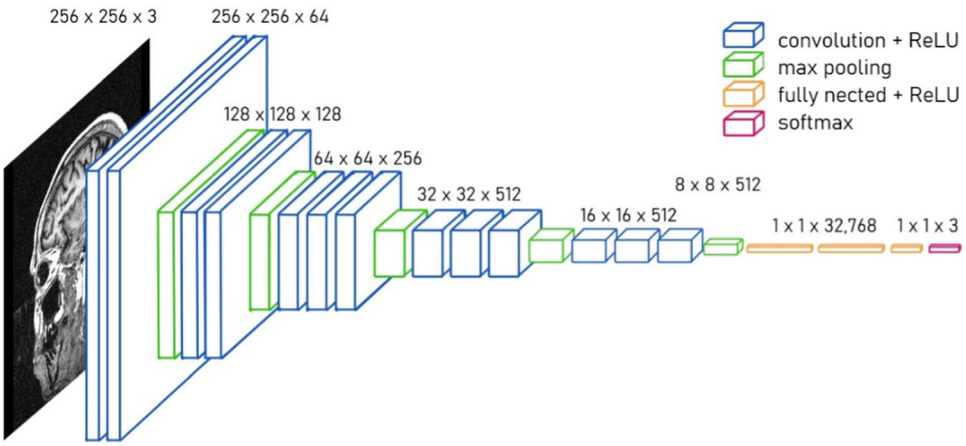

Fig. 4 VGG-16 architecture processing MRI dementia images. The pretrained convolutional base (*blue* and *green* blocks) is used as a feature extractor. The classifier (*orange* and *red* blocks) was trained with the OASIS-2 dataset

categories in the OASIS-2 classification problem. The convergence to the highest accuracy reached by the best CNN found is presented in Fig. 5, graphs a) and b).

Concerning the third group of classifiers, VGG-16 and ResNet-50 architectures were tested. The best CNN was VGG-16 with the configuration illustrated in Fig. 4, whose convolutional base is already pretrained with the ImageNet dataset. The convergence to the highest accuracy reached by both VGG-16 and ResNet-50 is presented in Fig. 5, graphs c) and d). It is worth mentioning that ResNet-50 required more than 100 epochs to converge, and because of this, it was extended to 1000 epochs. This and other relevant results and findings are discussed in the next section.

4 Discussion

Two groups of algorithms were studied in this work, classical ML methods and DL CNN architectures. Regarding classical ML methods for Alzheimer's classification, previous studies do not report a variability measure, which can be a sign of bias in their experiments and no option for reproducibility. Therefore, these classical ML

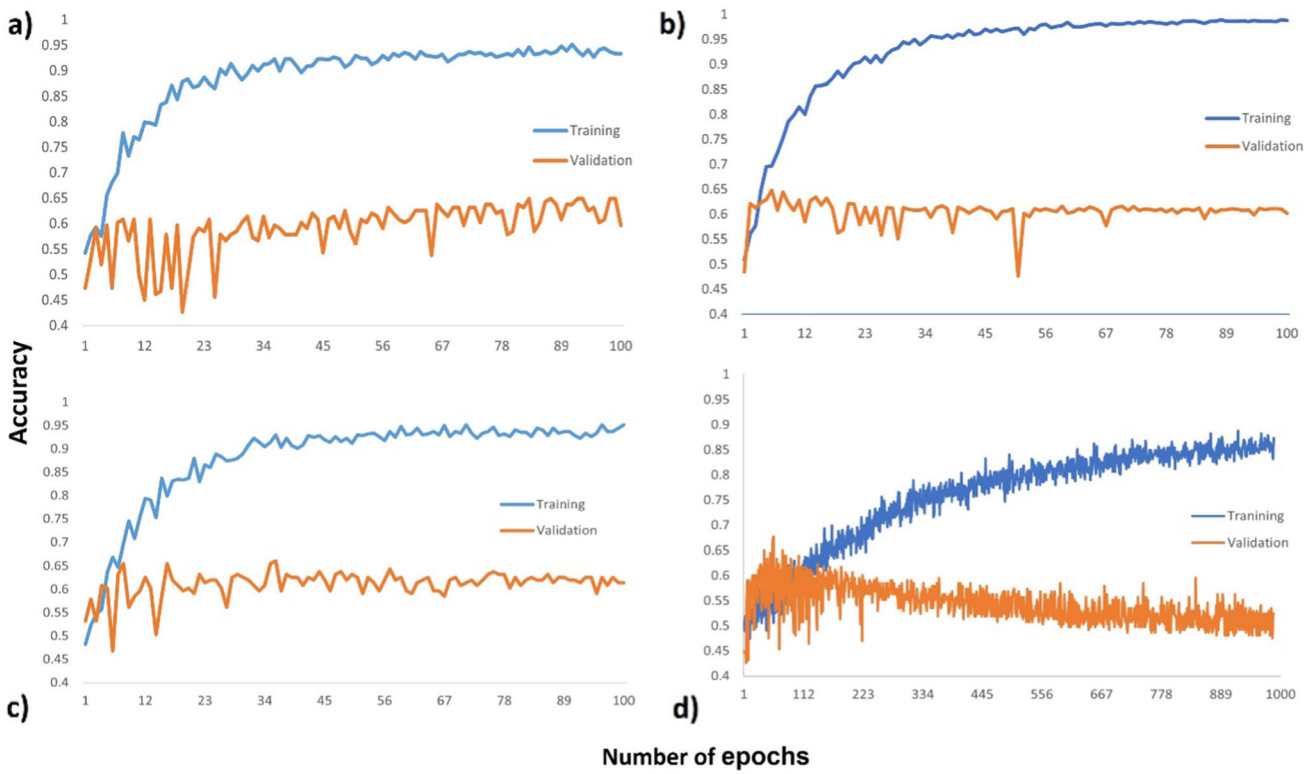

Fig. 5 Convergence of the designed **a** CNN-1, **b** CNN-2 and pretrained **c** VGG-16, **d** ResNet-50 convolutional networks. Maximum, average, and minimum of 10 independent runs are shown for each architecture

methods were reproduced to have a reliable basis for comparison with our proposed DL CNN architectures. Performance metrics of classical ML methods found in our experiments agree with the results of previous works [10–15] as observed in Table 2. Two notable exceptions are the case of the 96.66% accuracy obtained with the RF algorithm reported in [15], where the maximum instead of the average performance is reported; and the case of the DT algorithm reported in [12] which claim to reach an accuracy of 99.28%. This result lies outside the range of accuracies reported in all previous studies, including state-of-the-art DL approaches reported in [20]. After a thorough revision of the DT exception, two methodological omissions were found: (i) the authors do not explain the number of instances used for the accuracy computation, and (ii) a validation scheme like k-fold or bootstrapping was missing. Therefore, this accuracy level is doubtful and should be considered carefully. In [14] authors obtained a 68.75% accuracy by using a SVM with RBF kernel on the geographic data of OASIS-2. This low performance might be due to the hyperparameter tuning scheme used by the authors in which only low and high values for C and gamma were considered. In our experiments, a grid search scheme obtained 89.9% ± 3.2, which agrees with the result found in [15] of 92.57%.

In contrast to previous studies that employ classical ML algorithms and are dependent on the features proposed by neuroradiologists, the DL CNN architectures offer a complementary interpretation of the MRI images because these find non-evident features for the specialists, thus increasing available information for the Alzheimer's diagnostic.

The MRI studies available in the OASIS-2 and ADNI-1 datasets are sequences of longitudinal images that form a complete brain scan. Our experiments proved that a representative image (90th slice in OASIS-2 and a similar slice from ADNI-1 90th -115th) of each of these sequences is enough for the extraction of relevant features that lead to a classification of Alzheimer, up to 56 ± 4% according to the results presented in Table 2, confirming state-of-the-art results reported by [41].

Deep Learning approaches have been recently implemented (since 2017) for Alzheimer's classification, and these are reviewed in [20]. The most recent and best-performed DL methods found in the literature are CNN-BND (2020) and ADVIAN (2021) with accuracies of 97.19% ± 0.89 and 97.76% ± 1.13, respectively. These two works are close to our proposal; however, both were trained with the images of OASIS-1 (axial plane), which are different from OASIS-2 (sagittal plane), and therefore not directly comparable, but complementary since these are

the two most common types of MRI images available in clinical practice.

One of the closest studies to our proposal that trained the CNNs, BrainNet2D, and ResNet-18, in 2021 obtained accuracies of 88 and 89%, respectively [18]; however, CNNs were trained with OASIS-1 and validated with OASIS-2 impeding a direct comparison. Another recent related work trained different classic ML algorithms with the MRI sagittal images from OASIS-2 [41], this is the only study found in literature that allows a comparison against our proposal. This comparison is presented in Table 2 and the conclusion is that both classic ML algorithms our proposed CNNs reached similar performance on OASIS-2 dataset. Leaving a wide room for improvement in the performance of Deep Learning algorithms trained with MRI sagittal images.

The VGG-16 architecture is simpler than ResNet-18, ResNet-50, and similar to BrainNet2D, concerning the number of convolutional layers. This behavior was also observed for ADVIAN (a variant of VGG-16 with convolutional blocks attention modules) and CNN-BND (eight layers), regarding the better performance compared to more complex CNNs like ResNet-50 or Inception V4. An explanation of this behavior is offered in terms of overspecialization. According to theory [38], the first layers of a convolutional network learn generic filters that are useful for solving tasks with input images different from those used for training. In contrast, deeper layer filters learn very specialized patterns and are therefore not useful in other tasks. This does not mean that complex architectures are useless, but that due to their complexity, they probably require several images for complete retraining or, as shown for ResNet-50 in Fig. 5 d), many more epochs to converge.

Although image preprocessing of brain MRI, involving skull removal, motion correction, atlas registration, and gamma correction, among other strategies are commonly used in previous studies [18, 21, 22], there is no standard preprocessing workflow. In addition, no performance was reported by these studies about pretrained DL methods handling images without explicit preprocessing. It is possible to carry out an implicit preprocessing because pretrained CNNs have several filters available to recognize borders, textures, and other patterns that allow the extraction of relevant features from raw images independently of their location and scale. This is known as Transfer Learning and has shown promising results for medical image analysis in recent years as reported in the review provided by [42]. Following this latter scheme, we applied the convolutional base of the VGG-16 pretrained with the ImageNet dataset and the results shown in Table 2 confirm that is possible to reach reasonable accuracy rates for Alzheimer's classification using transfer learning without explicit image preprocessing.

In previous studies and our experiments, was observed that CNNs architectures with few layers suffice to obtain a high performance in Alzheimer's classification. Following this observation, the question of how many layers we need emerges. To answer this question, we tried several configurations of CNNs. As illustrated in Fig. 3, the architecture with one convolutional layer (VGG-16) outperformed the classic two-layer architecture (ResNet50). This is an indicator that generic filters obtained in the first convolutional step are enough to provide a good characterization of the input image. More specialized filters obtained in the second convolutional layer reduce the effectiveness of the classifier. Not only the accuracy is better but also the convergence is reached in fewer epochs with VGG-16. Figure 5 a) and b) shows that both CNN-1 and CNN-2 convergence in around 50 epochs for training accuracy, but the validation accuracy is unstable without a sign of convergence during the 100 epochs. Similarly, Fig. 5 c) shows that the VGG-16 convergence in less than 100 epochs, while d) ResNet-50 requires more than 1000 epochs.

Figure 4 shows the configuration of the pretrained convolutional base VGG-16 that outperformed ResNet-50. VGG-16 is a simple architecture that may be considered as a baseline for MRI Alzheimer's classification using the MRI sagittal slices, for axial plane images other options such as ADVIAN and CNN-BND are available. To ease the reproducibility of the discussed experimental results, all our codes and images are publicly available at [40].

4.1 Limitations

There are two main limitations of the proposed CNNs. First, as the number of subjects used to train the CNN might be not representative of the overall population of dementia patients, care must be taken not to blindly trust the tool when applied to other populations of subjects with dementia. Also, for the converted group, classifications are not accurate in most cases and might tend to underestimate the change in the patients from one category to another. The MRI images employed are in NIfTI-1 format and we cannot assure that the CNN works with other formats. With the accuracy levels reached so far by both classic ML methods and CNNs trained with MRI sagittal images, the proposed approach is not yet suitable to implement on the clinical practice and opens the opportunity of testing more preprocessing strategies, CNN architectures, MRI slices, and datasets, looking for improvements on the performance of methods for Alzheimer's classification.

5 Conclusions

The VGG-16 variant proposed for Alzheimer's classification from the OASIS-2 and ADNI-1 datasets of MRI images was as accurate as previous classical ML methods but inferior to the training with the OASIS-1 dataset. The highest accuracy in Alzheimer's classification obtained by both ML and DL methods is close to 98% for OASIS-1, less than 60% for OASIS-2 or ADNI-1 considering all the algorithms found in the literature. All the state-of-the-art CNNs achieved a higher accuracy than the best classical ML algorithms previously reported. While classical methods start from a characterization conducted by specialists, CNNs start from raw images without preprocessing; with the exception of OASIS-1, that is already preprocessed. Therefore, it is concluded and confirmed in our experiments that CNNs have the advantage of being simpler to train while maintaining or surpassing the accuracy of previous classical ML methods.

Simple CNNs, with few layers, suffice to automatically learn filters to extract features that characterized and lead to accurate discrimination between the presence or absence of Alzheimer. The designed and pretrained CNNs analyzed, learned from one representative MRI image, and classified Alzheimer without the feature selection step, where geographic data is provided by neuroradiologists nor explicit preprocessing.

Our methods and the rest of the state-of-the-art CNNs have been studied on benchmark datasets like OASIS-1, OASIS-2, ADNI, and others. Since these images were obtained from real cases, they can be considered as a reference for the study of structural damage in Alzheimer's patients and until now, OASIS-1 has demonstrated to be the best available option.

Acknowledgements Georgina Waldo-Benítez would like to thank the National Council for Science and Technology of Mexico (CONACYT) for the financial support through grant number 786231.

Author's contributions All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by GWB, LCP, PC, and MAS. The first draft of the manuscript was written by GWB and LCP, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the National Council for Science and Technology of Mexico (CONACYT) (Grant number 786231).

Data availability The datasets generated during and/or analyzed during the current study are available in the GitHub repository, [https://github.com/GinaWaldo/OASIS2-CNN].

Declarations

Conflict of interests Authors Georgina Waldo-Benitez, Luis Carlos Padierna, Pablo Ceron, and Modesto A. Sosa declare they have no financial interests. The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Delgado-Saborit J, Guercio V, Gowers A, Shaddick G, Fox N, Love S (2021) A critical review of the epidemiological evidence of effects of air pollution on dementia, cognitive function and cognitive decline in adult population. Sci Total Environ 757:143734
2. Reda R, Zanza A, Mazzoni A, Cicconetti A, Testarelli L, Di Nardo D (2021) An update of the possible applications of magnetic resonance imaging (MRI) in dentistry: a literature review. J Imaging 7(5):75
3. Westbrook C (2014) Handbook of MRI Technique. Wiley-Blackwell, England.
4. Custodio N, Duque L, Montesinos R, Alva-Diaz C, Mellado M, Slachevsky A (2020) Systematic review of the diagnostic validity of brief cognitive screenings for early dementia detection in spanish-speaking adults in Latin America. Front Aging Neurosci 4(12):270
5. Frisoni G, Fox NC, Jack CR Jr, Scheitens P, Thompson PM (2010) The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol 6:67–77
6. Marcus DF, Fotenos AF, Csernansky JG, Morris JC, Buckner RL (2010) Open access series of imaging studies (OASIS): longitudinal MRI data in nondemented and demented older adults. J Cogn Neurosci 22(12):2677–2684
7. Acabchuk R, Sun Y, Wolferz R J et al. (2015) 3D Modeling of the lateral ventricles and histological characterization of periventricular tissue in humans and mouse. J Vis Exp 19(99)
8. Kwak S, Kim H, Chey J, Youm Y (2018) Feeling how old i am: subjective age is associated with estimated brain age. Front Aging Neurosci 7(10):168
9. Ishida T, Tokuda K, Hisaka A et al (2019) A novel method to estimate long-term chronological changes from fragmented observations in disease progression. Clin Pharmacol Ther 105(2):436–447
10. Akti S, Kamar D, Anil Ozlu O, Soydemir I, Akcan M (2022) A comparative study of machine learning methods for predicting the evolution of brain connectivity from a baseline timepoint. J Neurosci Methods 368:109475
11. Petersen R, Aisen P, Beckett L, Donohue M (2010) Alzheimer's disease neuroimaging iniciative. Neurology 8:74
12. Bansal D, Chhikara R, Khanna K, Gupta P (2018) Comparative analysis of various machine learning algorithms for detecting dementia. Proc Comput Sci 132:1497–1502
13. Naidu C, Kumar D, Maheswari N, Sivagami M, Li G (2019) Prediction of alzheimer's disease using oasis dataset. Int J Recent Technol Eng 7:6s3
14. Battineni G, Chintalapudi N and Amenta F (2019) Machine learning in medicine: performance calculation of dementia

prediction by support vector machines (SVM). Informat Med Unlocked 16
15. Shanmuga E, Shahina A, Nayeemulla Khan A (2020) Dementia prediction on OASIS dataset using supervised and ensemble learning techniques. Int J Eng Adv Technol (IJEAT) 10(1):2249–8958
16. Yagis E, Luca C, Diciotti S, Marz C I, Workalemahu S, Seco de Herrara AG (2020) 3D convolutional neural networks for diagnosis of alzheimer's disease via structural MRI. In: IEEE 33rd international symposium on computer-based medical systems, pp 65–70
17. Saputra RA, Agustina C, Puspitasari D, Ramanda R and Pribadi D (2020) Detecting alzheimer's disease by the decision tree methods based on particle swarn optimization. J Phys Conf Ser, 1641
18. Saratxaga C, Moya I, Picón A, Acosta M, Moreno-Fernandez-de-Leceta A, Garrote E (2021) MRI deep learning-based solution for alzheimer's disease prediction. J Pers Med 11(902)
19. Khagi B, Lee B, Pyun Y, Kwon R (2019) CNN Models Performance Analysis on MRI images of OASIS dataset for distinction between Healthy and Alzheimer's patient. International Conference on Electronics, Information, and Communication (ICEIC) 1–4.
20. Gao S, Lima D (2021) A review of the application of deep learning in the detection of Alzheimer's. Int J Cognit Comput Eng 3:1–8
21. Jiang X, Zhang Y and Chang L (2020) Classification of alzheimer's disease via eight-layer convolutional neural network with batch normalization and dropout techniques. J Med Imaging Health Inform 10(5):1040–1048
22. Wang S-H, Zhou Q, Yang M and Zhang Y-D (2021) ADVIAN: Alzheimer's disease VGG-inspired attention network based on convolutional block attention module and multiple way data augmentation. Front Aging Neurosci, 13
23. OASIS-BRAINS (2021) Open access series of imaging studies. https://www.oasis-brains.org/ Accesed: 20 Sept 2021
24. Cortes C, Vapnik V (1995) Support vector machine. Mach Learn 20(3):273–297
25. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, New York
26. Deng N, Tian Y, Zhang C (2013) Support vector machines. CRC Press, Boca Ratón
27. Shalev-Shwartz S, Ben-David S (2014) Understanding machine learning: from theory to algorithms. Cambridge University Press, New York
28. Rosasco L (2017) Introductory machine learning notes. MIT Press, Cambridge
29. Mohri M, Rostamizadeh A, Talwalkar A (2018) Foundations of machine learning. MIT Press, Cambridge
30. Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press, Cambridge
31. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge
32. Alom M (2019) A state-of-the-art survey on deep learning theory and architecture. Electronics 8(3):292
33. Simonyan K and Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. Arxiv.
34. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. Arxiv.
35. Varian H (2005) Bootstrap Tutorial. Math J 9:768–775
36. James G, Witten D, Hastie T, Tibshirani R (2017) An introduction to statistical learning. Springer
37. Metz C (1978) Basic principles of ROC analysis. Seminar Nuclear Med 4:283
38. Chollet F (2018) Deep learning with python. Manning Publications, USA, 144
39. Martín Abadi AAPBEB (2015) TensorFlow: large-scale machine learning on heterogeneous systems. Tensorflow. [En línea]. Available: tensorflow.org..
40. Waldo Benítez GC (2022) GinaWaldo—GitHub. https://github.com/GinaWaldo/OASIS2-CNN. Accesed 21 Oct 2022
41. Balasundaram A, Srinivasan S, Prasad A, Malik J, Kumar A (2021) Hippocampus Segmentation-Based Alzheimer's Disease Diagnosis. Arab J Sci Eng 48:10249–10265
42. Morid M A, Borjali A and Del Fiol G (2021) A scoping review of transfer learning research on medical image analysis using ImageNet Computers. Biol Med, 128