Expert Systems With Applications 220 (2023) 119709

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An evolutionary explainable deep learning approach for Alzheimer's MRI classification

Shakila Shojaei, Mohammad Saniee Abadeh *, Zahra Momeni

Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran

ARTICLE INFO

Keywords:
Alzheimer's Disease Classification
Convolutional Neural Networks (CNN)
Explainable Deep Learning
Genetic Algorithm

ABSTRACT

Deep Neural Networks (DNN) are prominent Machine Learning (ML) algorithms widely used, especially in medical tasks. Among them, Convolutional Neural Networks (CNN) are well-known for image-based tasks and have shown excellent performance. In contrast to this remarkable performance, one of their most fundamental drawbacks is their inability to clarify the cause of their outputs. Moreover, each ML algorithm needs to present an explanation of its output to the users to increase its reliability. Occlusion Map is a method used for this purpose and aims to find regions of an image that have a significant impact on determining the network's output, which does this through an iterative process of occluding different regions of images. In this study, we used Magnetic Resonance Imaging (MRI) scans from Alzheimer's Disease Neuroimaging Initiative (ADNI) and trained a 3D-CNN model to diagnose Alzheimer's Disease (AD) patients from cognitively normal (CN) subjects. We tried to combine a genetic algorithm-based Occlusion Map method with a set of Backpropagation-based explainability methods, and ultimately, we found a brain mask for AD patients. Also, by comparing the extracted brain regions with the studies in this field, we found that the extracted regions are significantly effective in diagnosing AD from the perspective of Alzheimer's specialists. Our model achieved an accuracy of 87% in 5-fold cross-validation, which is an acceptable accuracy compared to similar studies. We considered a 3D-CNN model with 96% validation accuracy (on unmasked data that includes all 96 distinct brain regions of the Harvard-Oxford brain atlas), which we used in the genetic algorithm phase to produce a suitable brain mask. Finally, using lrp_z_plus_fast explainability method, we achieved 93% validation accuracy with only 29 brain regions.

1. Introduction

In recent years, using deep learning algorithms to solve various problems has become very widespread. Their successful performance in many fields has made them the main subject of numerous scientific research. However, despite their growing use, they face various challenges (Pouyanfar et al., 2018). One of the most notable drawbacks of Deep Neural Networks (DNN) is lacking to provide a clear interpretation of their decisions due to their black-box nature (Galli et al., 2022). Nonetheless, nowadays, models are becoming more complex to enhance accuracy and other performance measurement criteria. Hence, making DNN more interpretable, especially in sensitive tasks, is one of the most significant challenges in increasing its reliability (Buhrmester et al., 2021).

A well-known category of DNN called Convolutional Neural Networks (CNN) is widely used in image analysis. By way of illustration, it is used for classification, segmentation, localization, and detection problems in medical image analysis. In fact, medical images such as X-ray, Computerized Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET), instead of being analyzed by trained and experienced professionals, are examined by intelligent systems, and they do tasks such as abnormally detection, localization, and segmentation (Sarvamangala and Kulkarni, 2021). MRI is a method of imaging that can produce clear images of any of the internal organs of the body (Berger, 2002) and is one of the common modalities that is frequently used in neuroimaging (Bowman, 2014). In recent years, besides diagnosing and classification of diseases like Alzheimer's (Pei et al., 2022), Parkinson's (Barbero-Gómez et al., 2021), Schizophrenia

Peer review under responsibility of Submissions with the production note 'Please add the Reproducibility Badge for this item' the Badge and the following footnote to be added:The code (and data) in this article has been certified as Reproducible by the CodeOcean: https://codeocean.com. More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physicalsciencesandengineering/computerscience/journals.

* Corresponding author.
E-mail address: saniee@modares.ac.ir (M. Saniee Abadeh).

https://doi.org/10.1016/j.eswa.2023.119709
Received 2 July 2022; Received in revised form 26 December 2022; Accepted 13 February 2023
Available online 17 February 2023

Table 1
Examples of AD-related research that used explainability methods in their work.

Research paper	Explainability method	Data	Objectives	Results
(Bae et al., 2021)	Occlusion Map	sMRI	- Developing a 3D-CNN to predict conversion from MCI to Alzheimer - Identifying structural brain regions that contribute to DAT conversion	82.4 % Validation Accuracy
(Tang et al., 2019)	Occlusion Map/ Gradient-based Class Activation Map	WSI	- Proposing a CNN for amyloid plaque classification and localization - Using XAI to show that trained models learned relevant features	99.3 % AUROC and 74.3 % AUPRC on Hold-out data
(Shahamat and Saniee Abadeh, 2020)	Occlusion Map	sMRI	-Binary classification of Alzheimer's and Autism using 3D-CNN -Identifying knowledgeable brain regions for Alzheimer's and Autism using genetic algorithm	85 % Validation Accuracy (using 41 brain regions for Alzheimer's classification)
(Böhle et al., 2019)	Layer-wise Relevance Propagation/ Guided Backpropagation	sMRI	Explaining CNNs by Computing average heatmaps across Alzheimer's and healthy control patients	91.00 % Validation Accuracy 87.96 % Test Accuracy
(Sudar et al., 2022)	Layer-wise Relevance Propagation	sMRI	-Using VGG-16 for Alzheimer's Classification -Identifying the stages of Alzheimer's using XAI	78.12 % Validation Accuracy
(Bron et al., 2021)	Guided Backpropagation	sMRI	-Validating the generalizability of Alzheimer's disease classification in the prediction of conversion from MCI to Alzheimer's Using CNN (and Support Vector Machines) -Visualizing the regions that contributed to the classifications	AD/CN: 93 % Validation & 87.6 % Test AUC MCIn/MCInc: 74 % Validation & 70 % Test AUC
(Chakraborty et al., 2021)	Class Activation Map	sMRI	-Multi-class classification of Alzheimer's using 3D-CNN -Using XAI to show that the model pays attention to the important regions	95.88 % Validation Accuracy
(Feng et al., 2018)	Class Activation Map	sMRI	- Using longitudinal scans in classification to increase the amount of data -Using an average class activation map to evaluate learned features by the 2D-CNN model	93 % Test Accuracy
(Yang et al., 2018)	Class Activation Map/ Gradient-based Class Activation Map	sMRI	Introduced approaches to produce a visual explanations of 3D-CNNs in Alzheimer's classification	79.4 % 5-fold Accuracy
(Iizuka et al., 2019)	Gradient-based Class Activation Map	SPECT	-Evaluating the usefulness of deep learning in diagnosing AD (and DLB) -Using XAI to evaluate learned features by the CNN	92.39 % Validation Accuracy (CN vs AD)
(Nakagawa et al., 2020)	Deep Taylor	sMRI, Biomarkers	- Assessing if deep survival analysis can predict the conversion of MCI to Alzheimer's - Using deep Taylor to evaluate the effectiveness of each ROI in the prediction of Alzheimer's disease conversion	83.5 % concordance index
(Tinauer et al., 2021)	Deep Taylor	sMRI	- Proposing a regularization technique to train CNNs using deep taylor	86.19 % Validation Accuracy
Proposed	Occlusion Map and Backpropagation-based methods	sMRI	-Binary classification of Alzheimer's using 3D-CNN -Combining two categories of explainability methods to extract the most important brain regions in Alzheimer's (using genetic algorithm)	93 % Validaion Accuracy (using 29 brain regions) 96.6 % Validaion Accuracy (using 36 brain regions)

DAT, Dementia of Alzheimer's type; AUROC, Areas Under the Receiver Operating Characteristic; AUPRC, Areas Under the Precision Recall Curve, Whole Slide Image; SPECT, Single photon emission computed tomograph; DLB, Dementia with Lewy bodies.

(Hu et al., 2022), and Multiple Sclerosis (MS) (Yılmaz Acar et al., 2022; Narayana et al., 2020), using MRI images and CNN, making the results interpretable (Böhle et al., 2019; Eitel et al., 2019; Oh et al., 2019; Zhang et al., 2021; Organisciak et al., 2022; Lin et al., 2022) has also been considered.

In our previous work (Shahamat and Saniee Abadeh, 2020), we presented a genetic algorithm-based explainability framework to extract knowledgeable brain regions in Autism and Alzheimer's Disease (AD). However, one of the drawbacks is that, despite the acceptable accuracy, the random nature of evolutionary algorithms leads to different results in different experiments. In this paper, we developed our previous work to achieve more robust and reliable results, reduce time cost, and improve accuracy. We proposed a hybrid explainability method in which two families of explainability methods called Backpropagation-based methods and Occlusion Map are combined. We first trained a 3D-CNN model on Alzheimer's MRI data to classify AD subjects and cognitively normal (CN), then, using Backpropagation-based Explainability methods, extracted a primary mask to use in Occlusion Map. Occlusion Map needs an iterative process to find a proper mask, and we used genetic algorithm for this purpose. Finally, the primary mask extracted in the previous stage was fed into the genetic algorithm to reach a suitable brain mask. This mask indicates those brain areas that are considered essential from our network's point of view. Using Backpropagation-based methods to start from better initial brain masks leads the algorithm to better final masks. At the same time, it can lead to faster convergence and less time cost by limiting the search space.

The research structure is as follows: in section 2, descriptions of related works are provided, and in section 3, first, the data used in this research are introduced, then, a comprehensive explanation of our framework is presented. Section 4 is devoted to experimental results, and finally, the research is summarized and discussed in Section 5.

2. Related works

This section discusses relevant works closely related to the topic under study. These works include eXplainable Artificial Intelligence (XAI) methods for deep learning visualization and AD classification. Several well-known works done on these topics are reviewed in this section.

2.1. XAI methods for deep learning Visualization

Due to the non-transparent nature of deep learning and its widening use in sensitive issues, researchers have shown huge interest in explainability methods in recent years. That is why deep learning explainability, especially in networks such as CNN, which are mainly used in image analysis is becoming a hot topic. From one point of view, explainability methods are divided into two categories: Attribution and

Feature Visualization. Feature Visualization refers to what a network or part of it has learned in the training process, while Attribution methods focus on determining those parts of the input the network is more likely to make decisions based on (Yu and Shi, 2018). In the Feature Visualization category, (Erhan et al., 2009) presents the Activation Maximization method, which belongs to a family called gradient-based methods. The primary purpose of Activation Maximization is to understand what type of image at the input causes the maximum activation of a particular neuron or layer of the network. (Zeiler and Fergus, 2014) introduces DeconvNet, in which the CNN is trained first and then, with the idea behind deconvolution network, reconstructs an image based on the heatmaps obtained in a specific layer. This reconstructed image indicates which image this heatmap had the most activation for. Guided Backpropagation is another method presented by (Springenberg et al., 2014). This method makes a slight change in vanilla backpropagation and amplifies positive gradients to make sharper output images. In the Attribution category, a method called Class Activation Map (CAM) was introduced (Zhou et al., 2016). In this method, heatmaps obtained from the last convolution layer are multiplied by their corresponding weights. In other words, by summing their weights, a heatmap is calculated for a specific class of objects which identifies those parts of an image that CNN uses to detect that particular class. Gradient-based Class Activation Map (Grad-CAM), which is derived from CAM, also falls into this category (Selvaraju et al., 2017). In Saliency Map, by considering an input image, the gradient of the classes is calculated using backpropagation, and it determines which pixels of the image increase the probability of a particular class in a classification task. Saliency Map can also be computed by Guided Backpropagation, Integrated Gradient (Sundararajan et al., 2017), SmoothGrad (Smilkov et al., 2017), etc. (Simonyan et al., 2013). Occlusion Map blocks different parts of the input image through masks iteratively and attempts to find the parts that increase the probability of a class in classification (Zeiler and Fergus, 2014). Layer-wise Relevance Propagation (LRP) tries to identify the pixels that cause input to be placed in a particular class by making backward connections in the network so that neurons that have stronger connections with higher layers are more affected by these connections (Bach et al., 2015). Deep Taylor also applies Taylor decomposition on each layer of the network backward to calculate the relevances (Montavon et al., 2017).

From another perspective, Visualization methods can be divided into Perturbation and Backpropagation methods (Joshi et al., 2021). Backpropagation-based methods try to find the most effective features in the network's output using backpropagating the gradients in the network. CAM, Activation Maximization, DeconvNet, and LRP are well-known methods in this category. On the other hand, Perturbation is a family of methods that, by making modifications in the input and monitoring the output change, try to find effective features. Occlusion Map is a famous method in this category.

Almost all explainability methods are getting attention in different fields, especially in medical science, and much research has applied them to their networks. Table 1 shows a list of research that used explainability methods in AD-related research. By doing so, besides interpreting the model, they tried to increase the accuracy or prevent it from decreasing. That is because there is a compromise between the explainability and the accuracy of the model, and generally the accuracy decreases as the explainability increases. More importantly, one of the motivations for using explainability methods is that they can detect overfitting and biases in the model which are important in evaluating the model's performance, besides measurement criteria like accuracy. (Böhle et al., 2019) used LRP and Guided Backpropagation to find the importance of each voxel in the classification task. They compute LRP and Guided Backpropagation heatmap for each AD individual and then generate an average heatmap for each class to find out crucial atlas-based brain regions. (Shahamat and Saniee Abadeh, 2020) used genetic algorithm and Occlusion Map to find knowledgeable atlas-based brain regions. Nonetheless, due to the randomness of evolutionary algorithms, the main problem of this framework is achieving different results in different experiments that decrease the reliability of the results. In this paper, we attempt to combine genetic algorithm-based occlusion with Backpropagation-based methods. Indeed, after extracting an average heatmap for AD patients, we compared the performance of varying Backpropagation-based methods in extracting important brain regions. Moreover, the final extracted brain regions are compared to Alzheimer's professionals' perspective which shows that our model is well-trained and trustable.

Table 2
A list of Alzheimer's classification research we reviewed.

Research papers	Data	Network	Task
(Yagis et al., 2020)	sMRI	3D-CNN	Binary Classification
(Duc et al., 2020)	fMRI	3D-CNN	Binary Classification
(Jo et al., 2020)	PET	3D-CNN	Binary Classification
(Shahamat and Saniee Abadeh, 2020)	sMRI	3D-CNN	Binary Classification
(Huang et al., 2019)	Multimodal	3D-CNN	Binary Classification
	sMRI	3D-CNN	Binary Classification
(Amini et al., 2021)	sMRI	3D-CNN	Multi-class Classification
(Ding et al., 2019)	fMRI	3D-CNN	Multi-class Classification
(Folego et al., 2020)	CT	3D-CNN	Multi-class Classification
(Venugopalan et al., 2021)	Multimodal	3D-CNN	Multi-class Classification
(Gao et al., 2017)	PET	3D-CNN / 2D-CNN	Multi-class Classification
(Pereira et al., 2020)	sMRI	2D-CNN	Multi-class Classification
(Bae et al., 2020)	sMRI	2D-CNN	Binary Classification
(Liu et al., 2018)	PET	2D-CNN	Binary Classification
(Ying et al., 2021)	Multimodal	2D-CNN	Binary Classification

2.2. CNN-Based Alzheimer's Disease classification

Earlier diagnosis of Alzheimer's Disease before symptoms appear, besides timely treatment, can prevent severe and irreversible damage to the brain. Therefore, research on various methods of early diagnosis of Alzheimer's is one of the most popular research areas in Alzheimer's. Neuroimaging is one of the areas that has shown much attention to this issue ("Earlier Diagnosis", 2022). On the other hand, nowadays, deep learning methods, especially CNN, are used in various tasks related to radiology (Zhu et al., 2019). Both 2D-CNN and 3D-CNN are being used in Alzheimer's Disease classification. Because neuroimages are three-dimensional images 2D-CNNs are unable to keep the spatial relationships between slices and losing these features can lead to less accurate decisions. Contrastingly, 3D-CNNs keep these features but are more complex than 2D-CNN (Ebrahimighahnavieh et al., 2020). In that regard, numerous research used 3D-CNN for Binary Classification on structural MRI (Yagis et al., 2020), fMRI (Duc et al., 2020), PET (Jo et al., 2020), and Multimodal data (Huang et al., 2019). Likewise (Amini et al., 2021; Ding et al., 2019; Folego et al., 2020; Gao et al., 2017; Venugopalan et al., 2021) applied multi-class classification on Alzheimer's structural MRI, functional MRI, CT, PET, and Multimodal data, respectively. Among the above studies, (Gao et al., 2017) implemented a 2D-CNN as well. Moreover, (Bae et al., 2020; Liu et al., 2018; Pereira et al., 2020; Ying et al., 2021) used 2D-CNN in classification based on mentioned modalities. (Yagis et al., 2020) performed binary classification using 3D-CNN (which is similar to this paper) on 200 MRI scans

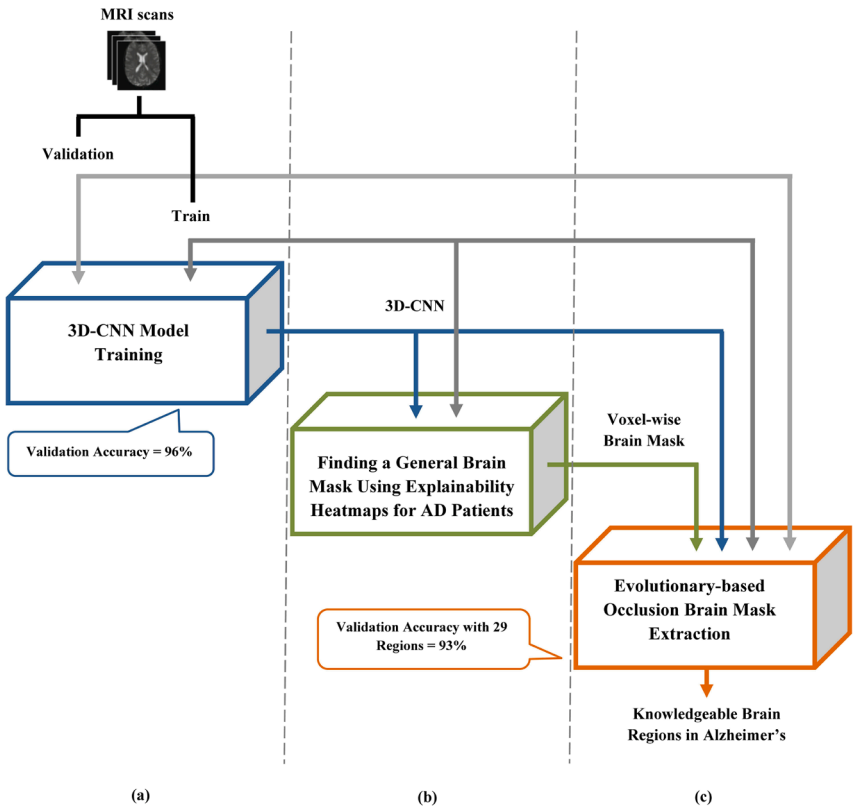

Fig. 1. Overall flowchart of the proposed framework. The proposed framework consists of three general sections. a) The neural network is trained on pre-processed Alzheimer's MRI scans. b) Using Backpropagation-based explainability methods, an early voxel-wise brain mask is generated for AD patients. c) Genetic algorithm is used to extract a suitable atlas-based brain mask for AD.

from ADNI (100 CE and 100 healthy control (HC) subjects) and 200 MRI scans from OASIS (100 CE and 100 HC subjects). They achieved 73.4 % accuracy on ADNI and 69.9 % on OASIS. Similarly (Shahamat and Saniee Abadeh, 2020) applied another 3D-CNN-based binary classification on 140 MRI scans (70 normal controls (NC) and 70 AD) from ADNI and obtained 5-fold classification accuracy of 85 %. Also, used a bigger dataset (475 CE and 494 HC) and reached an accuracy of 77 % in the classification of Alzheimer's. Table 2 shows a comparison between Alzheimer's classification research we reviewed.

3. Material and methods

This section introduces the data used in our research, then a comprehensive illustration of our method is given.

3.1. Dataset and Pre-processing

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is an association consisting of Canadian and American universities and medical centers that aims to develop standardized imaging techniques and biomarker procedures for normal, mild cognitive impairment (MCI) and AD subjects (Petersen et al., 2010). A total of 145 samples (74 AD and 71 CN) of structural MRI scans from ADNI repository (Jack et al., 2008) has been used in this paper. To avoid possible distribution shifts, we used scans from a specific range of ages. Secondly, we tried to divide the dataset so that the training and validation datasets are balanced in terms of gender and label.

All scans are pre-processed using FSL software (Jenkinson et al, 2012). First, using FMRIB's Linear Image Registration Tool (FLIRT) in FSL each scan is taken to the standard MNI152_T1_2_mm space. These MRI scans, which have the size of $91 \times 109 \times 91$, are then converted to the size of $80 \times 80 \times 80$ by cropping margins that do not contain important information.

3.2. Proposed method

Our goal is first to find the most important brain regions in AD from our network's perspective. Secondly, we want to examine which of the Backpropagation-based explainability methods generate heatmaps that are better able to find these areas. Fig. 1 shows the overall flowchart of the proposed framework. After data pre-processing and training the model, we use the iNNvestigate repository (Alber et al, 2019), which includes Backpropagation-based methods, to generate heatmaps. For each Backpropagation-based method, we generate heatmaps of the training data belonging to the AD patients (used in the network training phase). Then in a process that is similar to the weighted mean, we obtain an overall voxel-wise heatmap and call it an average heatmap, which identifies the most substantial areas. Using this average heatmap and single brain region atlases extracted from Harvard–Oxford cortical structural atlas, we generate an atlas-based mask for the brain areas. This mask is given as input to the genetic algorithm and generates a proper initial population (initial masks) that helps the algorithm start from a better location of search space. It means that many improper brain masks that the framework can consider are pruned from the beginning. Finally, with the convergence of the genetic algorithm, we will have an appropriate brain mask. To perceive the framework's performance, we apply the extracted brain mask to the validation data, give the masked validation data to the model, and compare its accuracy with the accuracy of the unmasked validation data (consisting of all brain regions).

3.2.1. Convolutional Neural network architecture

Convolutional Neural Networks are significant feedforward networks that automatically extract features from data (Li et al., 2021). We

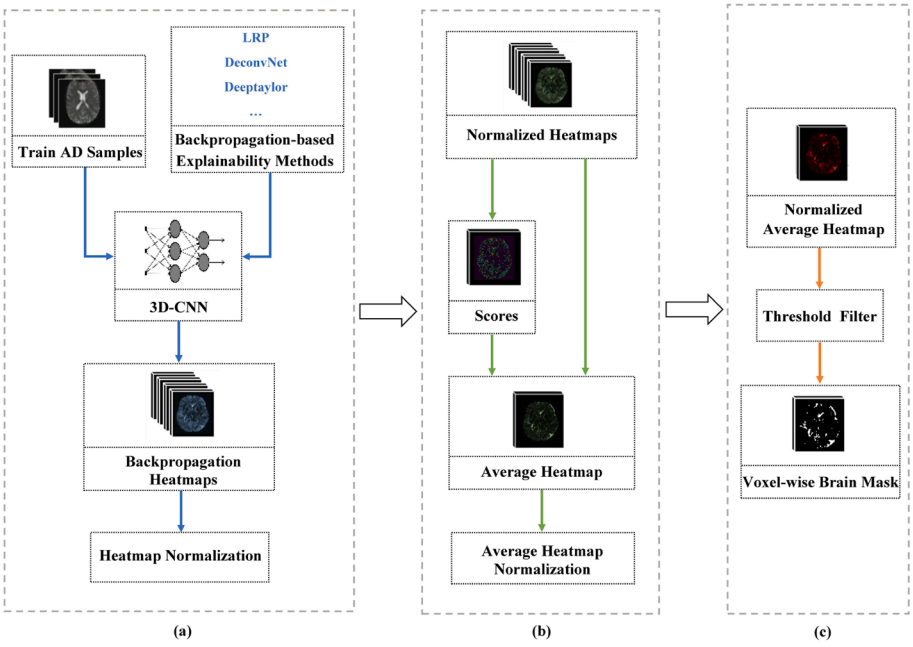

Fig. 2. Extracting Voxel-wise Brain Mask. a) Heatmaps of MRI scans of all AD patients are obtained (by each explainability method). b) With a procedure similar to the weighted mean, an average heatmap of AD patients' heatmaps is calculated. c) By ignoring the insignificant parts of the average heatmap, a voxel-wise brain mask is obtained for AD.

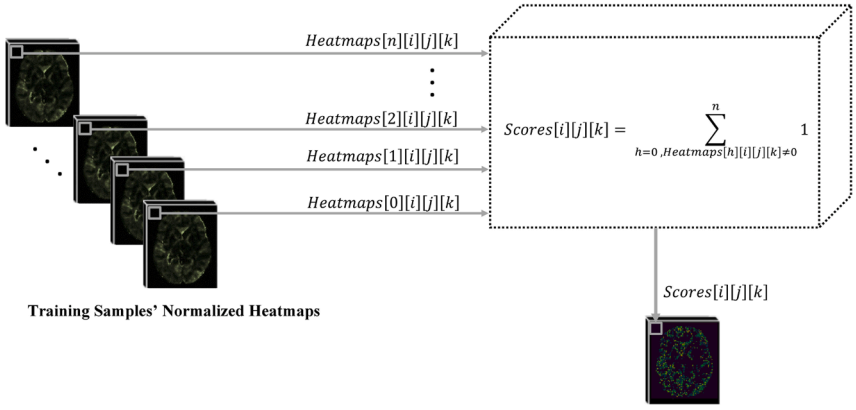

Fig. 3. Calculating the voxel scores. For each voxel, a score is computed by counting the number of AD samples' heatmaps in which the corresponding voxels have non-zero values.

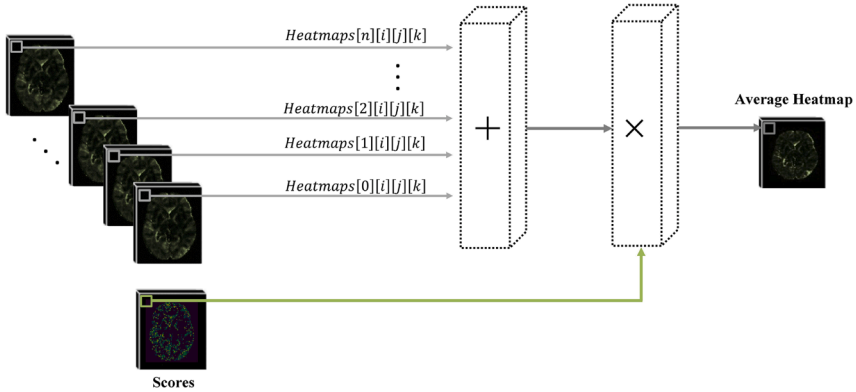

Fig. 4. Calculating Average Heatmap. For each voxel of the average heatmap, first, the sum of all corresponding voxels is calculated, then the result is multiplied by the corresponding voxel in Scores.

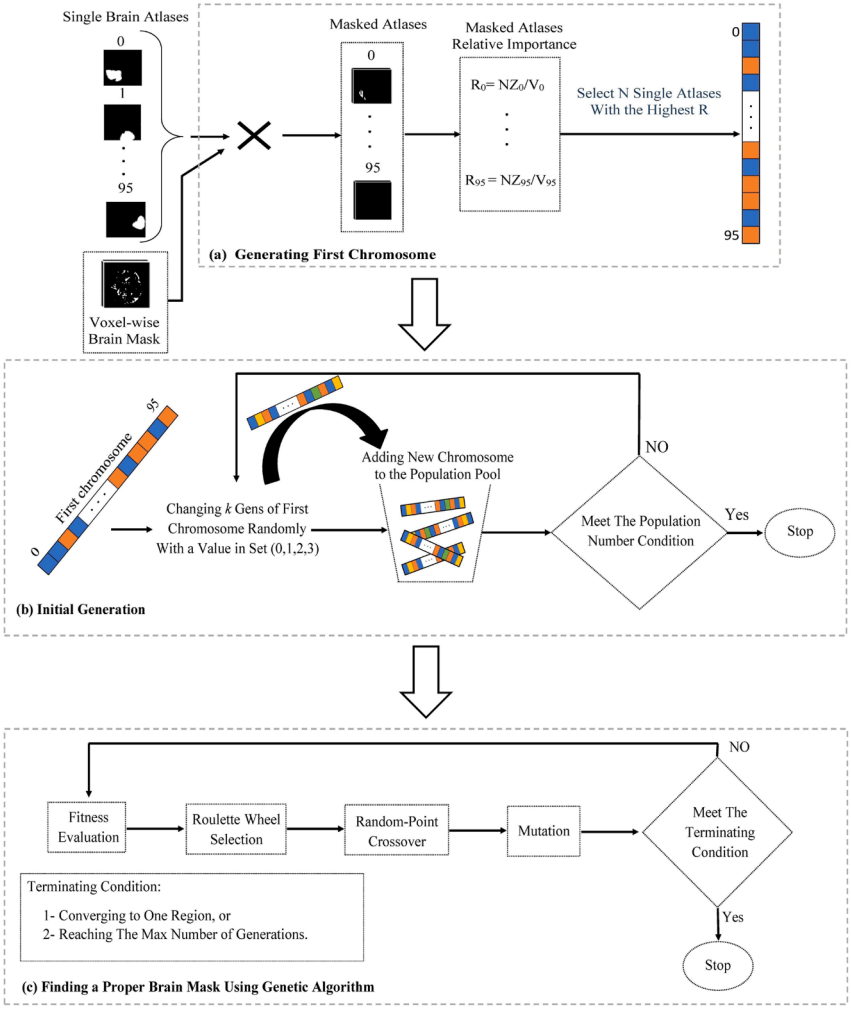

Fig. 5. Using genetic algorithm to extract a proper brain mask for AD patients. a) First, an initial brain mask is generated using explainability heatmaps and single brain atlases. b) Using this mask, a few other masks are generated for the first generation of the algorithm. c) Finally, in the genetic phase, we reach a proper brain mask for Alzheimer's Disease. N, number of desired regions to be selected; NZ_i, number of non-zero voxels in i_{th} masked atlas; V_i, number of voxels in i_{th} single atlas; R_i, relative importance of i_{th} atlas.

trained a 3D-CNN using MRI scans for a binary classification task that diagnoses AD patients from CN ones, and then we used it to extract a suitable brain mask. The network consists of three 3D convolution layers. A dropout layer is installed before the first convolution layer, and one is installed between the two dense layers to prevent overfitting. Also, the sigmoid activation function is used in dense layers. Finally, CNN classifies the data into two classes.

3.2.2. Voxel-wise brain mask extraction

After training CNN, we generate an average heatmap that shows the effectiveness of each voxel in the network output. Then using this average heatmap, a voxel-wise brain mask is computed. Fig. 2 illustrates the steps of extracting this mask. First, using each Backpropagation-based method a set of heatmaps is generated for all AD training samples. In the obtained heatmaps, for each voxel of the input images, the amount of effect that the voxel has on the network output is visible. In the next step, after normalizing the heatmaps, an array of voxel scores will be calculated based on all heatmaps (Fig. 3). Each voxel score counts the number of heatmaps in which the corresponding voxels has non-zero values. As Eq. (1) shows, Let *Heatmaps* be a 4D tensor that represents the normalized heatmaps of each explainability method we used, n be the number of AD samples in the training data, and h indicate an AD patient MRI scan, then $Heatmaps[h][i][j][k]$ is a single voxel of *h-th* AD training sample's heatmap and $Scores[i][j][k]$ is the number of heatmaps with non-zero values in the corresponding voxel.

$$Scores[i][j][k] = \sum_{\substack{h=0,\\ Heatmaps[h][i][j][k] \neq 0}}^{n} 1 \tag{1}$$

As Fig. 4 shows, to create the average heatmap, we use the Scores generated in the previous step (Fig. 3). We generally assume that the following two cases indicate the importance of a voxel. 1) The number of AD samples' heatmaps in which that voxel is considered essential (Scores that are equivalent to weights in a weighted mean). 2) The value of the corresponding voxel in each AD sample's normalized heatmap (in the normalized heatmaps, the closer a voxel is to one, the more important it is, and the closer it is to zero, the less important it is). We model these two factors as follows:

$$Average - Heatmap[i][j][k] = Scores[i][j][k] * \sum_{h=0}^{n} Heatmaps[h][i][j][k] \tag{2}$$

As the average heatmap is calculated, we have to normalize it again. To generate a voxel-wise mask, we need a threshold to ignore small values in the normalized average heatmap. This threshold depends on the explainability method, and we need tunning to find the best value for each method. Finally, we use this threshold and normalized average heatmap to generate the voxel-wise mask. To do so, first, an empty mask will be created with the same dimensions as the normalized average heatmap. Then for each voxel of the mask, we compare its corresponding voxel in the normalized average heatmap with the threshold. Voxels with a value higher than the threshold change to one and others remain

zero. In this way, we create a voxel-wise mask, and the farther the threshold is from zero, the more voxels will be pruned.

3.2.3. Genetic algorithm Occlusion

This section describes the main parts of the genetic algorithm used to extract an atlas-based brain mask. The genetic algorithm presented in (Shahamat and Saniee Abadeh, 2020) with some modifications has been used in this paper. Fig. 5 shows the general steps of applying the genetic algorithm.

3.2.3.1. Chromosome encoding. Each chromosome is made up of 96 genes, which are identified by the numbers 0 to 95. Each gene corresponds to a distinct region specified in the single brain atlases of Harvard–Oxford cortical structural atlas. Also, each gene can have one of the values in the set (0, 1, 2, 3). '0'-value genes refer to the regions that are not selected, and '1' means that the region is selected. '2' and '3' values mean that the corresponding regions are selected, and erosion and dilation operators have been applied to them, respectively. Indeed, each chromosome is a subset of distinct brain regions; However, by applying morphological operators (erosion and dilation), the margins of some regions can be slightly sharpened or thickened.

3.2.3.2. Initial population. First, we multiply each single brain atlas in the voxel-wise mask. Then we count the number of non-zero voxels in the multiplication result (NZ_i) and divide it by the total number of voxels of that single atlas (V_i). In this way, considering the correspondence between the voxel-wise mask and each single brain atlas, a ratio is obtained that determines which of the single brain atlases had the most activated voxels in the voxel-wise mask (Eq. (3)). Let R be the ratio set, and N be the number of desired brain regions we are considering for the initial chromosome. From these 96 single atlases, we select N of them with the highest R and assign '1' to their corresponding genes in the primary chromosome.

$$R(i) = \frac{NZ_i}{V_i} \tag{3}$$

To generate the initial population, unlike our previous work (Shahamat and Saniee Abadeh, 2020), which does it randomly (using the roulette wheel selection method), we have done this more intelligently using Backpropagation-based heatmaps. For this purpose, we consider the primary chromosome and randomly replace its k genes with a value from the set (0, 1, 2, 3). We repeat this step until reaching the initial population.

3.2.3.3. Genetic operators. Fitness: The fitness function that measures the quality of a chromosome consists of two parts. As mentioned, each chromosome is a subset of the brain regions that together form a mask. From one aspect, the quality of this chromosome is related to the model's accuracy after applying this mask to the data. In contrast, paying attention to the number of selected regions is also essential. Let f_1 indicate the accuracy of the model, f_2 be the inverse of the number of selected regions, and α and β be the coefficients of f_1 and f_2, respectively to balance them, then:

$$Fitness = \alpha f_1 + \beta f_2 \tag{4}$$

To obtain f_1, the accuracy of the 3D-CNN is measured on the training data using Eq. (5).

$$f_1 = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

We also calculate f_2 as follows:

$$f_2 = \frac{1}{\sum_{i=0, chromosome[i]\neq 0}^{96} 1} \tag{6}$$

Selection: Although generating the initial population has become more intelligent using Backpropagation-based explainability methods, the selection operation to transfer the chromosomes to the next generation is done with roulette wheel selection. i.e., the probability of selecting each member of the population is commensurate with that member's fitness which is shown in Eq. (7).

$$Prob(i) = \frac{Fitness(i)}{\sum_{j=1}^{n} Fitness(j)} \tag{7}$$

Crossover: The single-point crossover is used. A random point in the parent chromosomes is considered the pivot, and the crossover operation is performed to generate new chromosomes.

Mutation: A gene is randomly selected from a chromosome and replaced with a new value from the set (0, 1, 2, 3).

Important parameters: *population size* is 20, *mutation* and *crossover* probabilities are 0.6 and 0.4, respectively. α is 0.02, which makes β be 0.98. also, *generation number* is at most 2000 with early stopping in case the algorithm reaches one region.

Table 3
Summary of our 3D-CNN layers and parameters.

Layer (type)	Output Shape	Param #
Input	(None, 80, 80, 80, 1)	0
Dropout	(None, 80, 80, 80, 1)	0
Conv3D	(None, 80, 80, 80, 8)	1008
MaxPooling3D	(None, 40, 40, 40, 8)	0
BatchNormalization	(None, 40, 40, 40, 8)	32
Conv3D	(None, 40, 40, 40, 16)	3472
MaxPooling3D	(None, 20, 20, 20, 16)	0
BatchNormalization	(None, 20, 20, 20, 16)	64
Conv3D	(None, 20, 20, 20, 32)	13,856
MaxPooling3D	(None, 10, 10, 10, 32)	0
BatchNormalization	(None, 10, 10, 10, 32)	128
Flatten	(None, 32000)	0
Dense	(None, 512)	16,384,512
Dropout	(None, 512)	0
Dense	(None, 1)	513
Total params: 16,403,585		
Trainable params: 16,403,473		
Non-trainable params: 112		

4. Experimental results

In this section, we comprehensively discuss our results. Section 4.1 is devoted to the Alzheimer's classification and its results. We compare the final result of the proposed method with previous research. We have also examined the important factors in training our 3D-CNN model. In section 4.2, we assess the various explainability methods used for the initial brain mask generation and discuss the results and crucial parameters of this part of the work as well. We also compare the proposed framework's results with our previous work in which the genetic algorithm starts randomly.

4.1. CNN model for Alzheimer's classification

The designed CNN in our previous work (Shahamat and Saniee Abadeh, 2020) with some changes has been trained on our new dataset. For example, we added Batch Normalization to our CNN, and in addition, the activation function of the final layer has changed too. Table 3 summarizes the network architecture and model parameters. We have the network input in the first layer, which takes MRI scans in 80 * 80 * 80 dimensions and gives them to a dropout layer with a keep probability of 95 %. The application of this layer is to prevent overfitting during training. Then we have three convolution layers, after each of which we first apply a ReLU activation. Next, a max-pooling with a window size of 2 * 2 * 2 halves each dimension of scans after each convolution layer. In the first convolution layer, we have eight filters with a kernel size of 5 * 5 * 5. In the second convolution layer, we have 16 filters with a kernel

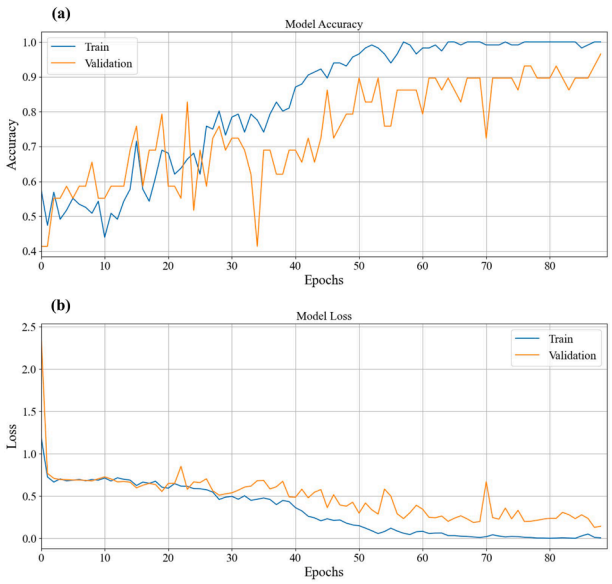

Fig. 6. Our 3D-CNN model accuracy (a) and loss (b) on train and validation data during training. This model is used as a part of fitness function in the genetic algorithm.

size of 3 * 3 * 3, and in the third one, we have 32 filters with a 3 * 3 * 3 kernel size. Then we come to the flattening layer, which prepares the data for the upcoming dense layer. The network has two dense layers with sigmoid activation functions. The first one has 32,000 inputs and 512 outputs, and the second one has 512 inputs and one output, which determines the classification result in one of the two classes based on the distance of the output value from zero and one. There is another dropout layer between these two dense layers with a keep probability of 95 %. The binary cross-entropy loss function and Adam optimizer with a learning rate of 0.00055 have been used to train the model. The batch size is four, and we trained the model for 100 epochs (2900 iterations); however, the training stopped after 90 epochs due to early stopping. It should also be mentioned that these hyperparameters were obtained by trial and error. We augmented the data using rotation to increase the number of MRI scans. We also used 5-fold cross-validation to evaluate the network and the model, which reached an accuracy of 87 %.

For the second part of the work, given in Section 4.2 (Extracting proper brain masks), we need to train a model that firstly we use to extract heatmaps using explainability methods and secondly use its train accuracy during the genetic process as a part of the fitness function. Out of 145 MRI scans, we used 116 scans as train data and the remaining 29 scans for validation in the training phase. Fig. 6 shows the accuracy and loss of both train and validation data during training. This model reached an accuracy of 96 %, precision of 94 %, recall of 100 %, and area under the curve (AUC) of 95.8 % on validation data.

Table 4 compares the results of the proposed work with some of the outstanding similar research on ADNI data. We presented the results of three other studies. For (Shahamat and Saniee Abadeh, 2020), we have shown the results with unmasked data. The proposed method's 5-fold accuracy is 87 %, which is higher than similar studies presented here.

4.2. Extracting proper brain masks

Twenty-five methods available in the iNNvestigate analyzer have been used to produce primary heatmaps. Each time, considering the AD training samples, one of these methods is applied to the base model (the model that is used in the genetic algorithm phase), and the heatmaps are obtained. After producing the average heatmap, we have to prune some low-effect voxels to reach the voxel-wise mask, which is done by setting a pruning threshold. For each of these explainability methods, this threshold must be obtained through tunning. In this way, voxels are pruned to the extent that in the next step, which is the conversion of this voxel-wise mask into an atlas-based mask, we attain the desired number of brain regions; therefore, these methods can be compared better. For this purpose, we have set this threshold so that each method reaches 60 to 70 regions. After producing an early atlas-based mask, we get to the point where we need to encode this mask into our genetic chromosome to generate the primary chromosome (a chromosome with 96 genes with gene values of 0 or 1, which indicates the selection or non-selection of the corresponding region). In this step, given that we want to select a specific number of brain regions, several regions may be pruned again. After many trials, we preferred to have 60 non-zero regions for the initial chromosome. Accordingly, if after ignoring some regions using the threshold filter, the number of regions is still more than 60, some of the less effective ones will be pruned again to reach 60 regions. This initial chromosome enters the production phase of the initial population. To generate the initial population, the chromosome is duplicated to the number of the population size, which has been considered 20. In each duplicate, we randomly replace the k genes with one value from the set (0, 1, 2, 3). In all our experiments, k is six; however, for a larger population size, a larger value of k should be used to further the diversity in the population.

After generating the initial population, the genetic algorithm starts, and if the algorithm reaches one region, the genetic phase ends. The maximum generation number of the genetic algorithm is 2000, which is usually only reached if the algorithm is located in a local optimum with a certain number of regions. The α and β values that control the fitness and convergence of the algorithm are 0.02 and 0.98, respectively. Larger α values will lead to later convergence.

Table 5 provides a complete report of the proposed framework applying various explainability methods. An initial chromosome with 60 regions has been created for all these experiments; however, since six (k) genes of each duplicate chromosome are changed to yield the initial population, the number of regions selected in each chromosome of the initial population is in the range of 54 to 66. Therefore, the number of selected regions in the fittest chromosome of the first generation of each experiment is in this range. The train and validation data used in the base model training are reused here. Before entering the genetic phase, the primary mask is produced using the AD training samples. In the genetic phase, according to the changes that occur in the mask in each

Table 4
Comparison of the proposed method with some outstanding previous research on similar data.

Reference	Method	Modality	#MRI Scans	Computational complexity	Network parameters	Accuracy
Proposed	3D-CNN (5-fold)	MRI	145	$O(n^5)$	≈ 16 Million	87 % (AD vs CN)
(Shahamat and Saniee Abadeh, 2020)	3D-CNN (5-fold)	MRI	140	$O(n^5)$	≈ 32 Million	85 % (AD vs CN)
(Yagis et al., 2020)	3D-CNN (5-fold)	MRI	200	$O(n^5)$	≈ 1.8 Million	73.4 % (AD vs HC)
(Pan et al., 2020)	123 distinct 2D-CNN (5-fold) + Ensemble Learning	MRI	299	$O(n^5)$	≈ 1.2 Million per 2D-CNN	84 % (AD vs HC)

AD, Alzheimer Disease; CN, Cognitively Normal; HC, Healthy Control.

Table 5
Comparison of using different explainability methods in the proposed framework to extract proper brain masks for AD patients. The number of regions in the first generation (RINFG) indicates the number of regions that our genetic algorithm started with.

Row	Method	Pruning Threshold	RINFG	Best Results #Generation	#Region	Train Accuracy	Validation Accuracy
1	gradient	0.02	59	135	41	97.4 %	96.6 %
2	**gradient_baseline**	**0.02**	**61**	**161**	**38**	**97.4 %**	**93 %**
3	input_t_gradient	0.022	61	290	29	96.6 %	82.8 %
4	deconvnet	0.42	59	339	29	93 %	86.2 %
5	guided_backprop	0.03	57	213	31	94 %	89.7 %
6	**integrated_gradients**	**0.19**	**62**	**382**	**31**	**96.6 %**	**93 %**
7	smoothgrad	0.024	59	431	23	86.2 %	89.7 %
8	lrp_z	0.006	64	180	40	94.8 %	93 %
9	lrp_z_IB	0.008	63	1089	37	88.8 %	86.2 %
10	lrp_epsilon	0.0066	64	249	41	94.8 %	93 %
11	lrp_epsilon_IB	0.007	62	260	40	92.2 %	82.8 %
12	lrp_w_square	0.51	58	547	20	79.2 %	86.2 %
13	lrp_flat	0.66	60	293	29	90 %	82.8 %
14	**lrp_alpha_2_beta_1**	**0.0038**	**61**	**455**	**30**	**94 %**	**93 %**
15	lrp_alpha_2_beta_1_IB	0.002	59	446	30	95.7 %	89.7 %
16	**lrp_alpha_1_beta_0**	**0.0045**	**58**	**239**	**36**	**93 %**	**96.6 %**
17	lrp_alpha_1_beta_0_IB	0.0061	63	455	34	96.6 %	89.7 %
18	lrp_z_plus	0.0065	59	244	30	88.8 %	89.7 %
19	**lrp_z_plus_fast**	**0.0008**	**61**	**256**	**29**	**94 %**	**93 %**
20	lrp_sequential_preset_a	0.003	63	384	38	97.4 %	93 %
21	lrp_sequential_preset_b	0.002	58	331	35	98.3 %	89.7 %
22	lrp_sequential_preset_a_flat	0.019	60	74	47	94.8 %	89.7 %
23	lrp_sequential_preset_b_flat	0.015	62	420	28	96.6 %	89.7 %
24	lrp_sequential_preset_b_flat_until_idx	0.015	59	463	18	75.9 %	86.2 %
25	deep_taylor	0.012	60	329	30	94.8 %	89.7 %

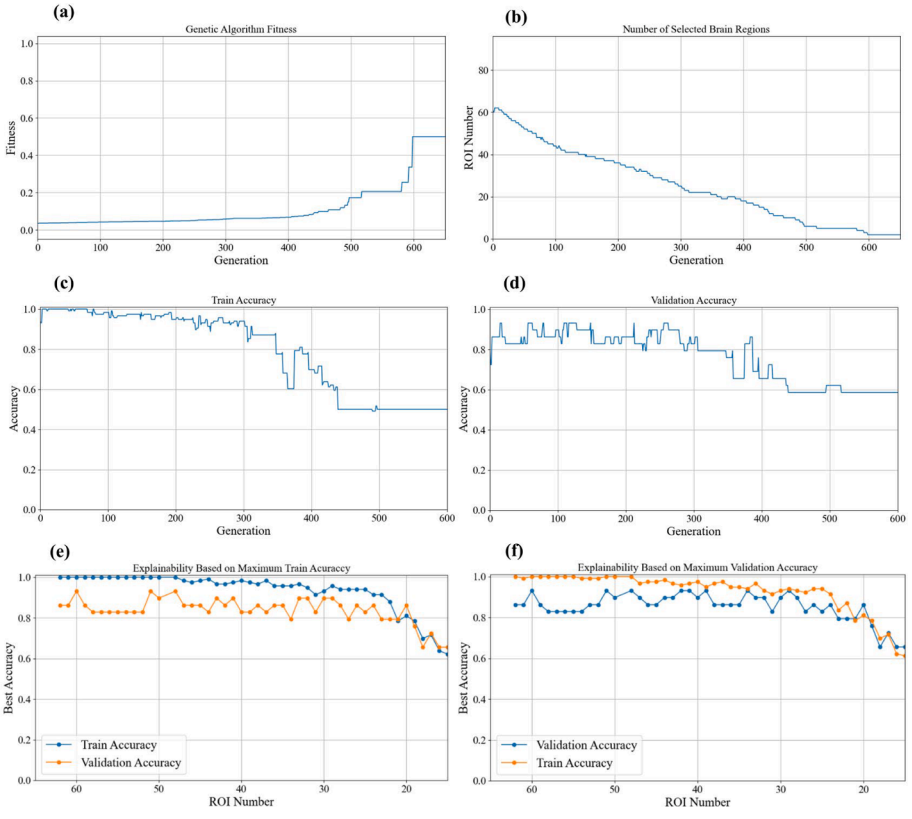

Fig. 7. Results of using lrp_z_plus_fast initial mask in genetic algorithm to generate a proper brain mask for Alzheimer's. a, b) show the genetic fitness value and the number of regions in the fittest mask of each generation, respectively. c, d) show the train and validation accuracy of the model on the masked data using the fittest brain mask of each generation. e, f) Specify the best train and validation accuracy for each number of regions reached by the algorithm, respectively. Fewer regions mean that the model is more explainable, and generally, with making a model more explainable, the accuracy decreases. Hence, these charts can provide a fair approximation of the model explainability so that we can choose a mask with fewer regions and higher model accuracy.

generation, all training samples are masked again, and the fitness is calculated (by counting the number of selected regions and the accuracy of the base model on the masked training samples). Then, the fittest mask (chromosome) is applied to the validation data to see its performance on the unseen data.

The last four columns of Table 5 contain information on the best brain masks extracted from each explainability method. Several experiments have been able to show outstanding performances. lrp_z_plus_fast method achieved 93 % validation accuracy with only 29 regions. Likewise, lrp_alpha_2_beta_1, integrated_gradients, gradient_baseline, and

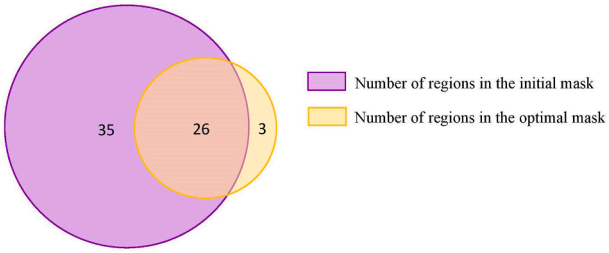

Fig. 8. Venn diagram of the brain regions in the initial mask and the optimal mask extracted by the genetic algorithm.

Table 6
Significant brain regions, extracted from the proposed framework.

Region ID	Brain Region	Reference
7	Left Frontal Orbital Cortex	(Qian et al., 2019; Tekin et al., 2001)
8	Left Frontal Pole	(Cajanus et al., 2019; Finger et al., 2017)
28	Left Parahippocampal Gyrus, anterior division	(Braak and Braak, 1990; Echávarri et al., 2011; Wang et al., 2016)
33	Left Postcentral Gyrus	(Peters et al., 2009; Yang et al., 2019)
47	Left Temporal Pole	(Arnold et al., 1994; Nag et al., 2018)
56	Right Frontal Pole	(Cajanus et al., 2019; Finger et al., 2017)
76	Right Parahippocampal Gyrus, anterior division	(Braak and Braak, 1990; Echávarri et al., 2011; Wang et al., 2016)
95	Right Temporal Pole	(Arnold et al., 1994; Nag et al., 2018)

lrp_sequential_preset_a reached 93 % accuracy with 30, 31, 38, and 38 regions, respectively. lrp_alpha_1_beta_0 method even has achieved 96 % accuracy with 36 regions. Since in the explainability topics, the aim is to reduce the factors influencing the network's output, it is essential to consider methods that with a small number of input regions do not have a decrease in accuracy or at least have a slight reduction. For example, the smoothgrad method reached 89.7 % accuracy with only 23 regions, which is excellent considering this number of brain regions. Furthermore, since evolutionary algorithms are known to have high time costs, another crucial factor is the time required to run the genetic algorithm (number of generations) to get the desired output. For example, gradient_baseline achieved 93 % accuracy for 38 regions only with 161 generations. Also, for lrp_z_plus_fast, which its genetic phase is shown in Fig. 7, we reached 93 % accuracy for 29 regions in the 256 generations. However, in our previous work, by starting the genetic randomly and without considering suitable initial solutions (masks) often, took more time to achieve a proper brain mask.

In addition to the ability of the proposed framework to extract the most important brain regions, it has also been able to partially solve another problem of our previous work. Since randomness is an inherent and inseparable feature of evolutionary algorithms (Kromer et al., 2013), its different experiments lead to different results, which makes it a little harder to trust the results. It means that different experiments, even with acceptable accuracy, may extract brain regions with a small overlap. In this framework, because we force the algorithm to reduce its randomness factors (using explainability heatmaps) if we repeat each of the experiments in Table 5 several times, the extracted regions have slight differences, which will increase the reliability of the framework.

Another subject that demonstrates the reliability of our framework is the number of areas shared between the optimal mask and the initial mask generated by each average heatmap. As Fig. 8 shows 26 regions of the optimal mask generated by our framework were also considered in the first generation of the genetic algorithm. It means that even without any optimization we extracted the most important regions by the initial

Table A1
Regions of Harvard-Oxford Atlas and their corresponding IDs.

ID	Region Name	ID	Region Name
0	Left Angular Gyrus	48	Right Angular Gyrus
1	Left Central Opercular Cortex	49	Right Central Opercular Cortex
2	Left Cingulate Gyrus, anterior division	50	Right Cingulate Gyrus, anterior division
3	Left Cingulate Gyrus, posterior division	51	Right Cingulate Gyrus, posterior division
4	Left Cuneal Cortex	52	Right Cuneal Cortex
5	Left Frontal Medial Cortex	53	Right Frontal Medial Cortex
6	Left Frontal Operculum Cortex	54	Right Frontal Operculum Cortex
7	Left Frontal Orbital Cortex	55	Right Frontal Orbital Cortex
8	Left Frontal Pole	56	Right Frontal Pole
9	Left Heschl's Gyrus (includes H1 and H2)	57	Right Heschl's Gyrus (includes H1 and H2)
10	Left Inferior Frontal Gyrus, pars opercularis	58	Right Inferior Frontal Gyrus, pars opercularis
11	Left Inferior Frontal Gyrus, pars triangularis	59	Right Inferior Frontal Gyrus, pars triangularis
12	Left Inferior Temporal Gyrus, anterior division	60	Right Inferior Temporal Gyrus, anterior division
13	Left Inferior Temporal Gyrus, posterior division	61	Right Inferior Temporal Gyrus, posterior division
14	Left Inferior Temporal Gyrus, temporooccipital part	62	Right Inferior Temporal Gyrus, temporooccipital part
15	Left Insular Cortex	63	Right Insular Cortex
16	Left Intracalcarine Cortex	64	Right Intracalcarine Cortex
17	Left Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex)	65	Right Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex)
18	Left Lateral Occipital Cortex, inferior division	66	Right Lateral Occipital Cortex, inferior division
19	Left Lateral Occipital Cortex, superior division	67	Right Lateral Occipital Cortex, superior division
20	Left Lingual Gyrus	68	Right Lingual Gyrus
21	Left Middle Frontal Gyrus	69	Right Middle Frontal Gyrus
22	Left Middle Temporal Gyrus, anterior division	70	Right Middle Temporal Gyrus, anterior division
23	Left Middle Temporal Gyrus, posterior division	71	Right Middle Temporal Gyrus, posterior division
24	Left Middle Temporal Gyrus, temporooccipital part	72	Right Middle Temporal Gyrus, temporooccipital part
25	Left Occipital Fusiform Gyrus	73	Right Occipital Fusiform Gyrus
26	Left Occipital Pole	74	Right Occipital Pole
27	Left Paracingulate Gyrus	75	Right Paracingulate Gyrus
28	Left Parahippocampal Gyrus, anterior division	76	Right Parahippocampal Gyrus, anterior division
29	Left Parahippocampal Gyrus, posterior division	77	Right Parahippocampal Gyrus, posterior division
30	Left Parietal Operculum Cortex	78	Right Parietal Operculum Cortex
31	Left Planum Polare	79	Right Planum Polare
32	Left Planum Temporale	80	Right Planum Temporale
33	Left Postcentral Gyrus	81	Right Postcentral Gyrus
34	Left Precentral Gyrus	82	Right Precentral Gyrus
35	Left Precuneous Cortex	83	Right Precuneous Cortex
36	Left Subcallosal Cortex	84	Right Subcallosal Cortex
37	Left Superior Frontal Gyrus	85	Right Superior Frontal Gyrus
38	Left Superior Parietal Lobule	86	Right Superior Parietal Lobule
39	Left Superior Temporal Gyrus, anterior division	87	Right Superior Temporal Gyrus, anterior division
40	Left Superior Temporal Gyrus, posterior division	88	Right Superior Temporal Gyrus, posterior division
41	Left Supracalcarine Cortex	89	Right Supracalcarine Cortex
42	Left Supramarginal Gyrus, anterior division	90	Right Supramarginal Gyrus, anterior division
43	Left Supramarginal Gyrus, posterior division	91	Right Supramarginal Gyrus, posterior division
44	Left Temporal Fusiform Cortex, anterior division	92	Right Temporal Fusiform Cortex, anterior division
45	Left Temporal Fusiform Cortex, posterior division	93	Right Temporal Fusiform Cortex, posterior division
46	Left Temporal Occipital Fusiform Cortex	94	Right Temporal Occipital Fusiform Cortex
47	Left Temporal Pole	95	Right Temporal Pole

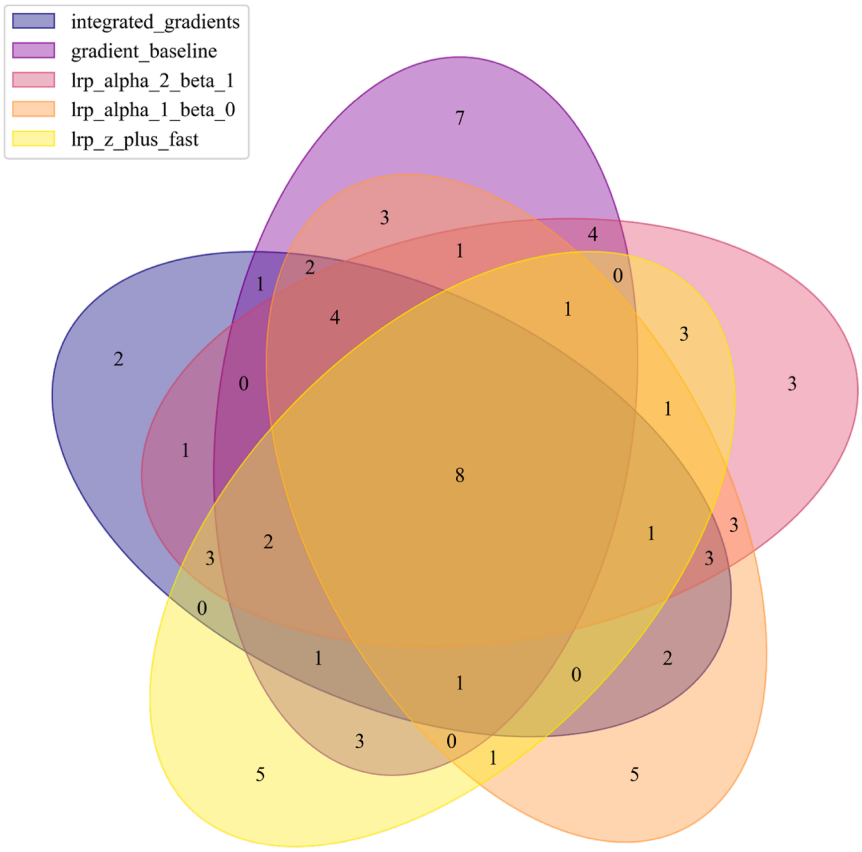

Fig. 9. Venn diagram of the number of brain regions extracted in five outstanding explainability methods consists of integrated_gradients, gradient_baseline, lrp_alpha_2_beta_1, lrp_alpha_1_beta_0, lrp_z_plus_fast. Intersected regions are 'Right Parahippocampal Gyrus, anterior division', 'Left Postcentral Gyrus', 'Right Temporal Pole', 'Left Parahippocampal Gyrus, anterior division', 'Left Frontal Pole', 'Left Temporal Pole', 'Right Frontal Pole', 'Left Frontal Orbital Cortex'.

mask generated from average heatmaps. In addition, as the optimization proceeded most of the less important regions were pruned and only three regions were added to the optimal mask.

5. Discussion and Conclusion

One of the critical issues in CNN explainability is finding the input regions which are more crucial in decision-making. More importantly, in professional topics such as medicine, where ordinary people cannot discern whether attention to a part of the image has been logical or not, the results need to be examined more closely. In this framework, using several outstanding experiments, a series of regions is extracted, which can be judged by an artificial intelligence expert only according to the accuracy of their output. To examine the results more accurately, it is necessary to compare the extracted regions with several studies in this field and see if our framework can identify the effective areas. Table 6 shows brain regions extracted based on several experiments performed. Firstly, we extracted the anterior Parahippocampal Gyrus, and it is shown that the earliest neuropathological changes in AD appear in the anterior part of the Parahippocampal Gyrus (Braak and Braak, 1990). Also, studies show that in the early stages of Alzheimer's healthy aging, aMCI, and mild AD patients are more distinguishable from Parahippocampal volume than Hippocampal volume (Echávarri et al., 2011). (Wang et al., 2016) too indicate that Parahippocampal Gyrus is one of the most vulnerable brain regions to Alzheimer's disease. Two other regions are the left and right Frontal Pole, which severely degenerate in Alzheimer's patients (Finger et al., 2017) and are related to aberrant motor behavior (Cajanus et al., 2019). Temporal Pole is another highly affected region (Arnold et al., 1994) and is among the most frequently involved areas with Alzheimer's (Nag et al., 2018). Moreover, it is confirmed that agitation and aberrant motor behavior in AD patients are associated with changes in the left Frontal Orbital Cortex (Tekin et al., 2001). Besides, one region which is almost regularly associated with delusions in Alzheimer's patients is Frontal Orbital Cortex (Qian et al., 2019). The left Postcentral Gyrus is one of the left hemisphere regions that is significantly reduced in Alzheimer's (Yang et al., 2019), and it is among the frontal regions that in verbal short-term memory tasks has less activation in AD patients compared to normal controls (Peters et al., 2009).Table A1.

The fact that despite the random nature of the genetic algorithm, the proposed framework agreed on significant regions even for different methods shows its appropriate performance. Moreover, as has been shown in Fig. 8 the average heatmap generated directly after applying explainability methods on the model (without any optimization using genetics) was consisting of all these 8 regions which shows that despite using a small dataset our model learned the important features. The results reported in Table 6 can be seen from another aspect in Fig. 8. This Venn diagram indicates the number of regions that were jointly selected in the experiments we used to extract brain regions. Also, Fig. 9 shows all these eight brain regions individually and together.Fig. 10.Fig. A1.

Conclusion: In this paper, we aimed to identify significant brain regions in diagnosing Alzheimer's Disease using a combination of explainability methods and genetic algorithm, and see if they can extract meaningful regions. To do so, we first trained a 3D-CNN with brain MRI scans. Then, we generated a primary Occlusion Map mask using Backpropagation-based explainability heatmaps and tried to improve it through genetic algorithm. This primary mask helps our genetic algorithm prune some inappropriate solutions (brain masks) and start from a better location of the search space. As a result, our algorithm reaches more accurate and reliable solutions with a lower time cost. In fact, not

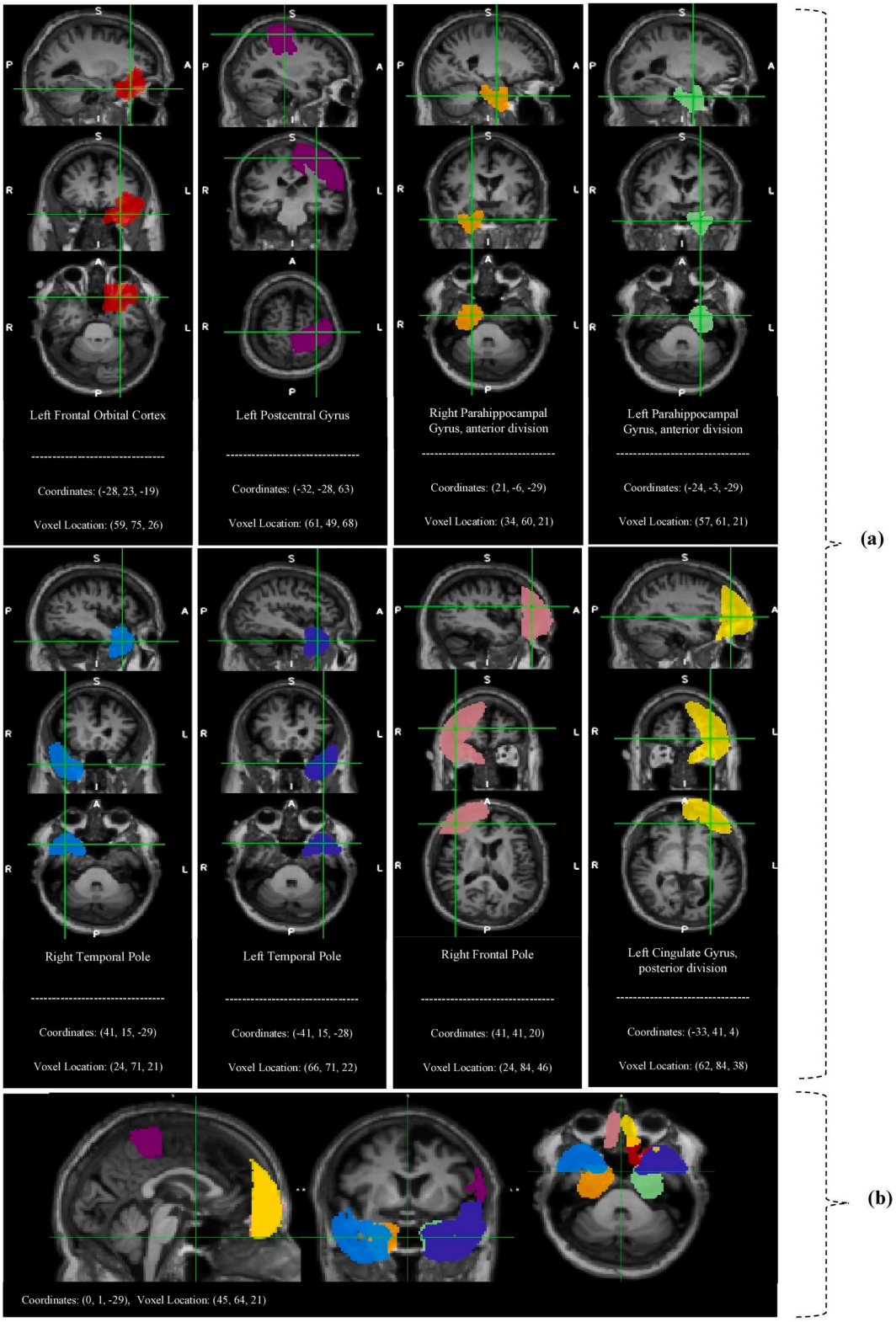

Fig. 10. Brain regions extracted from the proposed framework. a) shows each region individually. b) shows all extracted brain regions together.

only genetic algorithm is used to help XAI, but we have also used XAI methods to improve search in evolutionary algorithms. Our network reached a 5-fold accuracy of 87 % (on unmasked data). We used a 3D-CNN model with 96 % accuracy (on unmasked data), and by obtaining a proper mask of 29 regions using lrp_z_plus_fast, it reached 93 % accuracy. We extracted Parahippocampal Gyrus (anterior division), Left Postcentral Gyrus, Temporal Pole, Frontal Pole and, Left Frontal Orbital Cortex as the most significant brain regions related to Alzheimer's Disease, and we also showed that these extracted regions are meaningful.

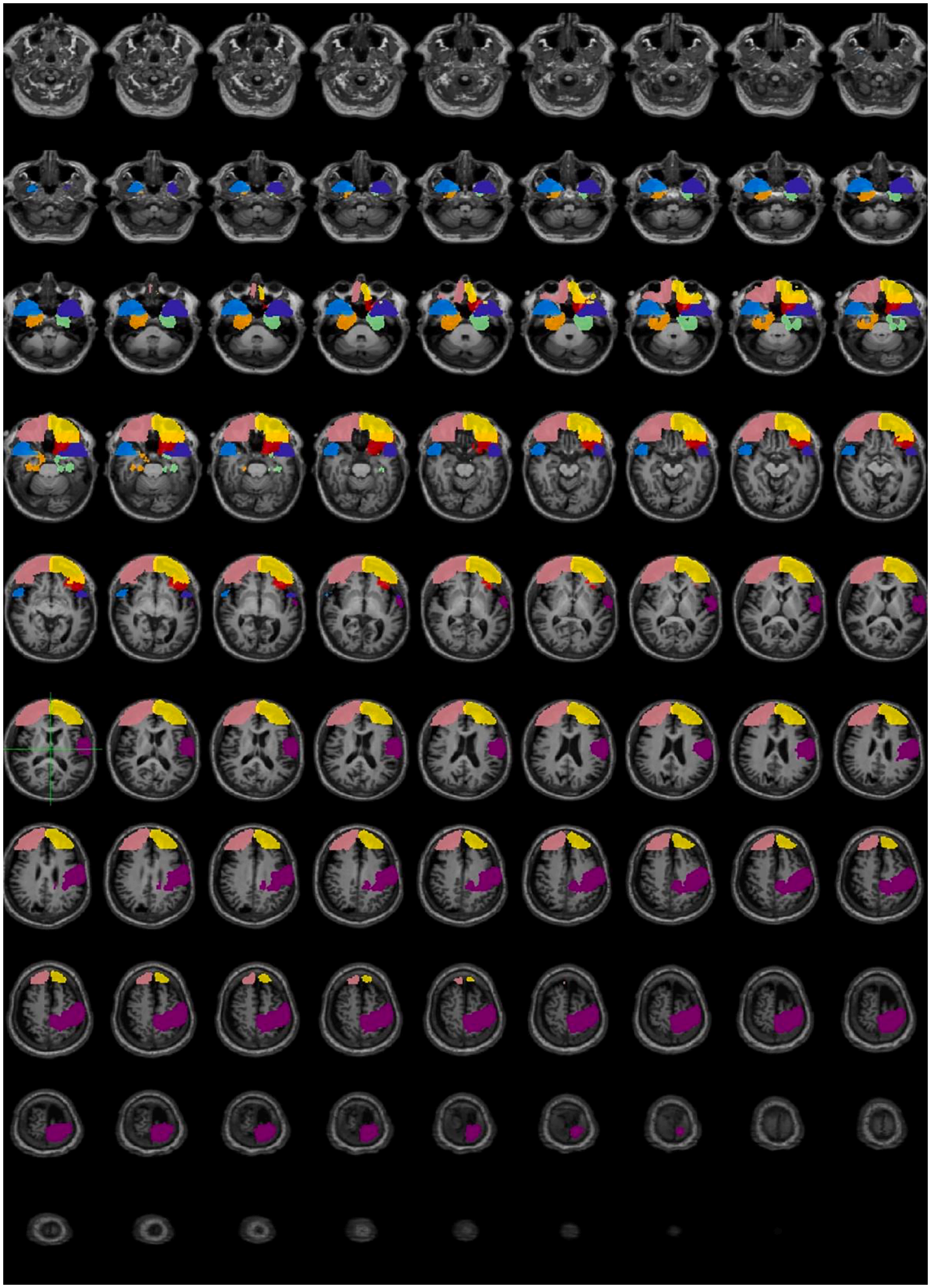

Fig. A1. Extracted brain regions in a sample MRI scan.

CRediT authorship contribution statement

Shakila Shojaei: Conceptualization, Methodology, Software, Validation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Mohammad Saniee Abadeh:** Conceptualization, Validation, Writing – review & editing, Supervision, Project administration. **Zahra Momeni:** Validation, Writing – review & editing, Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., … Kindermans, P. J. (2019). iNNvestigate neural networks! *Journal of Machine Learning Research, 20*.

Amini, M., Pedram, M., Moradi, A., & Ouchani, M. (2021). Diagnosis of Alzheimer's Disease Severity with fMRI Images Using Robust Multitask Feature Extraction Method and Convolutional Neural Network (CNN). *Comput. Math. Methods Med., 2021*, 5514839. https://doi.org/10.1155/2021/5514839

Arnold, S. E., Hyman, B. T., & Van Hoesen, G. W. (1994). Neuropathologic changes of the temporal pole in Alzheimer's disease and Pick's disease. *Arch. Neurol., 51*, 145–150. https://doi.org/10.1001/archneur.1994.00540140051014

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS One, 10*, e0130140.

Bae, J., Stocks, J., Heywood, A., Jung, Y., Jenkins, L., Hill, V., Katsaggelos, A., Popuri, K., Rosen, H., Beg, M.F., Wang, L., Alzheimer's Disease Neuroimaging Initiative, 2021. Transfer learning for predicting conversion from mild cognitive impairment to dementia of Alzheimer's type based on a three-dimensional convolutional neural network. Neurobiol. Aging 99, 53–64. https://doi.org/10.1016/j.neurobiolaging.2020.12.005.

Bae, J. B., Lee, S., Jung, W., Park, S., Kim, W., Oh, H., … Kim, K. W. (2020). Identification of Alzheimer's disease using a convolutional neural network model based on T1-weighted magnetic resonance imaging. *Sci. Rep., 10*, 22252. https://doi.org/10.1038/s41598-020-79243-9

Barbero-Gómez, J., Gutiérrez, P., Vargas, V., Vallejo-Casas, J., & Hervás-Martínez, C. (2021). An ordinal CNN approach for the assessment of neurological damage in Parkinson's disease patients. *Expert Systems with Applications, 182*, Article 115271. https://doi.org/10.1016/j.eswa.2021.115271

Berger, A. (2002). Magnetic resonance imaging. *BMJ, 324*, 35. https://doi.org/10.1136/bmj.324.7328.35

Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. (2019). Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification. *Front. Aging Neurosci., 11*, 194. https://doi.org/10.3389/fnagi.2019.00194

Bowman, F. D. (2014). Brain Imaging Analysis. *Annu Rev Stat Appl, 1*, 61–85. https://doi.org/10.1146/annurev-statistics-022513-115611

Braak, H., & Braak, E. (1990). Neurofibrillary changes confined to the entorhinal region and an abundance of cortical amyloid in cases of presenile and senile dementia. *Acta Neuropathol., 80*, 479–486. https://doi.org/10.1007/bf00294607

Bron, E. E., Klein, S., Papma, J. M., Jiskoot, L. C., Venkatraghavan, V., Linders, J., … van der Lugt, A. (2021). Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage: Clinical, 31*, Article 102712. https://doi.org/10.1016/j.nicl.2021.102712

Buhrmester, V., Münch, D., & Arens, M. (2021). Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey. *Machine Learning and Knowledge Extraction, 3*, 966–989. https://doi.org/10.3390/make3040048

Cajanus, A., Solje, E., Koikkalainen, J., Lötjönen, J., Suhonen, N.-M., Hallikainen, I., … Hall, A. (2019). The association between distinct frontal brain volumes and behavioral symptoms in mild cognitive impairment, Alzheimer's disease, and frontotemporal dementia. *Front. Neurol., 10*, 1059. https://doi.org/10.3389/fneur.2019.01059

Chakraborty, S., Sain, M., Park, J., & Aich, S. (2021). Early Detection of Alzheimer's Disease from 1.5 T MRI Scans Using 3D Convolutional Neural Network, in. In *Proceedings of International Conference on Smart Computing and Cyber Security. Springer Singapore* (pp. 15–28). https://doi.org/10.1007/978-981-15-7990-5_2

Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., … Franc, B. L. (2019). A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain. *Radiology, 290*, 456–464. https://doi.org/10.1148/radiol.2018180958

Duc, N. T., Ryu, S., Qureshi, M. N. I., Choi, M., Lee, K. H., & Lee, B. (2020). 3D-Deep Learning Based Automatic Diagnosis of Alzheimer's Disease with Joint MMSE Prediction Using Resting-State fMRI. *Neuroinformatics, 18*, 71–86. https://doi.org/10.1007/s12021-019-09419-w

Ebrahimighahnavieh, M. A., Luo, S., & Chiong, R. (2020). Deep learning to detect alzheimer's disease from neuroimaging: A systematic literature review. *Computer Methods and Programs in Biomedicine, 187*, Article 105242. https://doi.org/10.1016/j.cmpb.2019.105242

Echávarri, C., Aalten, P., Uylings, H. B. M., Jacobs, H. I. L., Visser, P. J., Gronenschild, E. H. B. M., … Burgmans, S. (2011). Atrophy in the parahippocampal gyrus as an early biomarker of Alzheimer's disease. *Brain Struct. Funct., 215*, 265–271. https://doi.org/10.1007/s00429-010-0283-8

Eitel, F., Soehler, E., Bellmann-Strobl, J., Brandt, A. U., Ruprecht, K., Giess, R. M., … Ritter, K. (2019). Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *Neuroimage Clin, 24*, Article 102003. https://doi.org/10.1016/j.nicl.2019.102003

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal, 1341*, 1.

Feng, X., Yang, J., Lipton, Z. C., Small, S. A., & Provenzano, F. A. (2018). Deep Learning on MRI Affirms the Prominence of the Hippocampal Formation in Alzheimer's Disease Classification. bioRxiv. *Alzheimer's Disease Neuroimaging Initiative*. https://doi.org/10.1101/456277

Finger, E., Zhang, J., Dickerson, B., Bureau, Y., Masellis, M., Alzheimer's Disease Neuroimaging Initiative, 2017. Disinhibition in Alzheimer's disease is associated with reduced right frontal pole cortical thickness. J. Alzheimers. Dis. 60, 1161–1170. https://doi.org/10.3233/JAD-170348.

Folego, G., Weiler, M., Casseb, R. F., Pires, R., & Rocha, A. (2020). Alzheimer's Disease Detection Through Whole-Brain 3D-CNN MRI. *Front Bioeng Biotechnol, 8*, Article 534592. https://doi.org/10.3389/fbioe.2020.534592

Galli, A., Piscitelli, M., Moscato, V., & Capozzoli, A. (2022). Bridging the gap between complexity and interpretability of a data analytics-based process for benchmarking energy performance of buildings. *Expert Systems with Applications, 206*, Article 117649. https://doi.org/10.1016/j.eswa.2022.117649

Gao, X. W., Hui, R., & Tian, Z. (2017). Classification of CT brain images based on deep learning networks. *Comput. Methods Programs Biomed., 138*, 49–56. https://doi.org/10.1016/j.cmpb.2016.10.007

Hu, M., Qian, X., Liu, S., Koh, A., Sim, K., Jiang, X., … Zhou, J. (2022). Structural and diffusion MRI based schizophrenia classification using 2D pretrained and 3D naive Convolutional Neural Networks. *Schizophrenia Research, 243*, 330–341. https://doi.org/10.1016/j.schres.2021.06.011

Huang, Y., Xu, J., Zhou, Y., Tong, T., Zhuang, X., Alzheimer's Disease Neuroimaging Initiative (ADNI), 2019. Diagnosis of Alzheimer's Disease via Multi-Modality 3D Convolutional Neural Network. Front. Neurosci. 13, 509. https://doi.org/10.3389/fnins.2019.00509.

Iizuka, T., Fukasawa, M., & Kameyama, M. (2019). Deep-learning-based imaging-classification identified cingulate island sign in dementia with Lewy bodies. *Sci. Rep., 9*, 8944. https://doi.org/10.1038/s41598-019-45415-5

[dataset]Jack Jr., C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., De-Carli, C. S., Krueger, G., Ward, H. A., Metzger, G. J., Scott, K. T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J. P., Fleisher, A. S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., & Weiner, M. W. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging, 27, pp. 685–691. doi:https://doi.org/10.1002/jmri.21049.

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). *FSL. NeuroImage, 62*(2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

Jo, T., Nho, K., Risacher, S.L., Saykin, A.J., Alzheimer's Neuroimaging Initiative, 2020. Deep learning detection of informative features in tau PET for Alzheimer's disease classification. BMC Bioinformatics 21, 496. https://doi.org/10.1186/s12859-020-03848-0.

Joshi, G., Walambe, R., & Kotecha, K. (2021). A Review on Explainability in Multimodal Deep Neural Nets. *IEEE Access, 1–1*. https://doi.org/10.1109/access.2021.3070212

Kromer, P., Snael, V., Zelinka, I., 2013. Randomness and chaos in genetic algorithms and differential evolution, in: 2013 5th International Conference on Intelligent Networking and Collaborative Systems. Presented at the 2013 International Conference on Intelligent Networking and Collaborative Systems (INCoS), IEEE. https://doi.org/10.1109/incos.2013.36.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Trans Neural Netw Learn Syst*, 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

Lin, Q., Niu, Y., Sui, J., Zhao, W., Zhuo, C., & Calhoun, V. (2022). SSPNet: An interpretable 3D-CNN for classification of schizophrenia using phase maps of resting-state complex-valued fMRI data. *Medical Image Analysis, 79*, Article 102430. https://doi.org/10.1016/j.media.2022.102430

Liu, M., Cheng, D., Yan, W., Alzheimer's Disease Neuroimaging Initiative, 2018. Classification of Alzheimer's Disease by Combination of Convolutional and Recurrent Neural Networks Using FDG-PET Images. Front. Neuroinform. 12, 35. https://doi.org/10.3389/fninf.2018.00035.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit., 65*, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008

Nag, S., Yu, L., Boyle, P. A., Leurgans, S. E., Bennett, D. A., & Schneider, J. A. (2018). TDP-43 pathology in anterior temporal pole cortex in aging and Alzheimer's disease. *Acta Neuropathol. Commun., 6*. https://doi.org/10.1186/s40478-018-0531-3

Nakagawa, T., Ishida, M., Naito, J., Nagai, A., Yamaguchi, S., Onoda, K., on behalf of the Alzheimer's Disease Neuroimaging Initiative. (2020). Prediction of conversion to Alzheimer's disease using deep survival analysis of MRI images. *Brain Communications*. https://doi.org/10.1093/braincomms/fcaa057

Narayana, P. A., Coronado, I., Sujit, S. J., Wolinsky, J. S., Lublin, F. D., & Gabr, R. E. (2020). Deep Learning for Predicting Enhancing Lesions in Multiple Sclerosis from Noncontrast MRI. *Radiology, 294*, 398–404. https://doi.org/10.1148/radiol.2019191061

Oh, K., Kim, W., Shen, G., Piao, Y., Kang, N.-I., Oh, I.-S., & Chung, Y. C. (2019). Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization. *Schizophr. Res., 212*, 186–195. https://doi.org/10.1016/j.schres.2019.07.034

Organisciak, D., Shum, H., Nwoye, E., & Woo, W. (2022). RobIn: A robust interpretable deep network for schizophrenia diagnosis. *Expert Systems with Applications, 201*, Article 117158. https://doi.org/10.1016/j.eswa.2022.117158

Pan, D., Zeng, A., Jia, L., Huang, Y., Frizzell, T., & Song, X. (2020). Early Detection of Alzheimer's Disease Using Magnetic Resonance Imaging: A Novel Approach Combining Convolutional Neural Networks and Ensemble Learning. *Front. Neurosci., 14*, 259. https://doi.org/10.3389/fnins.2020.00259

Pei, Z., Wan, Z., Zhang, Y., Wang, M., Leng, C., & Yang, Y. (2022). Multi-scale attention-based pseudo-3D convolution neural network for Alzheimer's disease diagnosis using structural MRI. *Pattern Recognition, 131*, Article 108825. https://doi.org/10.1016/j.patcog.2022.108825

Pereira, M., Fantini, I., Lotufo, R., Rittner, L., 2020. An extended-2D CNN for multiclass Alzheimer's Disease diagnosis through structural MRI, in: Medical Imaging 2020: Computer-Aided Diagnosis. Presented at the Medical Imaging 2020: Computer-Aided Diagnosis, SPIE, pp. 438–444. https://doi.org/10.1117/12.2550753.

Peters, F., Collette, F., Degueldre, C., Sterpenich, V., Majerus, S., & Salmon, E. (2009). The neural correlates of verbal short-term memory in Alzheimer's disease: An fMRI study. *Brain, 132*, 1833–1846. https://doi.org/10.1093/brain/awp075

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., … Weiner, M. W. (2010). Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology, 74*, 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., … Iyengar, S. S. (2018). A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Comput. Surv., 51*, 1–36. https://doi.org/10.1145/3234150

Qian, W., Schweizer, T. A., Churchill, N. W., Millikin, C., Ismail, Z., Smith, E. E., … Fischer, C. E. (2019). Gray matter changes associated with the development of delusions in Alzheimer disease. *Am. J. Geriatr. Psychiatry, 27*, 490–498. https://doi.org/10.1016/j.jagp.2018.09.016

Sarvamangala, D. R., & Kulkarni, R. V. (2021). Convolutional neural networks in medical image understanding: A survey. *Evol. Intell., 1–22*. https://doi.org/10.1007/s12065-020-00540-3

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV). Presented at the 2017 IEEE International Conference on Computer Vision (ICCV), IEEE. https://doi.org/10.1109/iccv.2017.74.

Shahamat, H., & Saniee Abadeh, M. (2020). Brain MRI analysis using a deep learning based evolutionary approach. *Neural Netw., 126*, 218–234. https://doi.org/10.1016/j.neunet.2020.03.017

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. https://doi.org/10.48550/arXiv.1312.6034.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M., 2017. SmoothGrad: removing noise by adding noise. https://doi.org/10.48550/arXiv.1706.03825.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2014. Striving for Simplicity: The All Convolutional Net. https://doi.org/10.48550/arXiv.1412.6806.

K. Sudar P. Nagaraj S. Nithisaa R. Aishwarya M. Aakash S. Lakshmi Alzheimer's Disease Analysis using Explainable Artificial Intelligence (XAI). 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) 2022 10.1109/icscds53736.2022.9760858.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic Attribution for Deep Networks, in: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research. PMLR, pp. 3319–3328.

Tang, Z., Chuang, K. V., DeCarli, C., Jin, L.-W., Beckett, L., Keiser, M. J., & Dugger, B. N. (2019). Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. *Nat. Commun., 10*, 2173. https://doi.org/10.1038/s41467-019-10212-1

Tekin, S., Mega, M. S., Masterman, D. M., Chow, T., Garakian, J., Vinters, H. V., & Cummings, J. L. (2001). Orbitofrontal and anterior cingulate cortex neurofibrillary tangle burden is associated with agitation in Alzheimer disease. *Ann. Neurol., 49*, 355–361. https://doi.org/10.1002/ana.72

Tinauer, C., Heber, S., Pirpamer, L., Damulina, A., Schmidt, R., Stollberger, R., … Langkammer, C. (2021). *Interpretable Brain Disease Classification and Relevance-Guided Deep Learning.*. https://doi.org/10.1101/2021.09.09.21263013

Venugopalan, J., Tong, L., Hassanzadeh, H. R., & Wang, M. D. (2021). Multimodal deep learning models for early detection of Alzheimer's disease stage. *Sci. Rep., 11*, 3254. https://doi.org/10.1038/s41598-020-74399-w

Wang, M., Roussos, P., McKenzie, A., Zhou, X., Kajiwara, Y., Brennand, K. J., … Zhang, B. (2016). Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. *Genome Med., 8*, 104. https://doi.org/10.1186/s13073-016-0355-3

Yagis, E., Citi, L., Diciotti, S., Marzi, C., Workalemahu Atnafu, S., G. Seco De Herrera, A., 2020. 3D convolutional neural networks for diagnosis of Alzheimer's disease via structural MRI, in: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). Presented at the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), IEEE. https://doi.org/10.1109/cbms49503.2020.00020.

Yang, C., Rangarajan, A., & Ranka, S. (2018). Visual Explanations From Deep 3D Convolutional Neural Networks for Alzheimer's Disease Classification. *AMIA Annu. Symp. Proc., 2018*, 1571–1580.

Yang, H., Xu, H., Li, Q., Jin, Y., Jiang, W., Wang, J., … Wang, T. (2019). Study of brain morphology change in Alzheimer's disease and amnestic mild cognitive impairment compared with normal controls. *Gen Psychiatr, 32*, e100005.

Yılmaz Acar, Z., Başçiftçi, F., & Ekmekci, A. (2022). A Convolutional Neural Network model for identifying Multiple Sclerosis on brain FLAIR MRI. *Sustainable Computing: Informatics and Systems, 35*, Article 100706. https://doi.org/10.1016/j.suscom.2022.100706

Ying, Q., Xing, X., Liu, L., Lin, A.-L., Jacobs, N., Liang, G., 2021. Multi-Modal Data Analysis for Alzheimer's Disease Diagnosis: An Ensemble Model Using Imagery and Genetic Features. bioRxiv. https://doi.org/10.1101/2021.05.07.443184.

Yu, R., & Shi, L. (2018). A user-based taxonomy for deep learning visualization. *Visual Informatics, 2*, 147–154. https://doi.org/10.1016/j.visinf.2018.09.001

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014* (pp. 818–833). Springer International Publishing. https://doi.org/10.1007/978-3-319-10590-1_53.

Zhang, Y., Hong, D., McClement, D., Oladosu, O., Pridham, G., & Slaney, G. (2021). Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods, 353*, Article 109098. https://doi.org/10.1016/j.jneumeth.2021.109098

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929.

Zhu, G., Jiang, B., Tong, L., Xie, Y., Zaharchuk, G., & Wintermark, M. (2019). Applications of Deep Learning to Neuro-Imaging Techniques. *Front. Neurol., 10*, 869. https://doi.org/10.3389/fneur.2019.00869