Visual Explanations From Deep 3D Convolutional Neural Networks for Alzheimer's Disease Classification

Chengliang Yang, Anand Rangarajan, Ph.D., and Sanjay Ranka, Ph.D.
Dept. of Computer & Information Science & Engineering
University of Florida, Gainesville, FL 32611, USA
ximen14@ufl.edu, anand@cise.ufl.edu, ranka@cise.ufl.edu

Abstract

We develop three efficient approaches for generating visual explanations from 3D convolutional neural networks (3D-CNNs) for Alzheimer's disease classification. One approach conducts sensitivity analysis on hierarchical 3D image segmentation, and the other two visualize network activations on a spatial map. Visual checks and a quantitative localization benchmark indicate that all approaches identify important brain parts for Alzheimer's disease diagnosis. Comparative analysis show that the sensitivity analysis based approach has difficulty handling loosely distributed cerebral cortex, and approaches based on visualization of activations are constrained by the resolution of the convolutional layer. The complementarity of these methods improves the understanding of 3D-CNNs in Alzheimer's disease classification from different perspectives.

1 Introduction

For years, medical informatics researchers have pursued data-driven methods to automate disease diagnosis procedures for early detection of many deadly diseases. Treatment of Alzheimer's disease, which has become the sixth leading cause of death in the United States [1], is one of the conditions that could benefit from computer-aided diagnostic techniques. A particular challenge of Alzheimer's disease is that it is difficult to detect in early stages before mental decline begins. But medical imaging holds promise for earlier diagnosis of Alzheimer's disease [2]. Magnetic resonance imaging (MRI), computed tomography (CT), and positron emission tomography (PET) scans contain information about the effects of Alzheimer's disease on the brains structure and functioning. But analyzing such scans is very time consuming for doctors and researchers because each scan contains millions of voxels.

Deep learning systems are one potential solution for processing medical images automatically to make diagnosing Alzheimer's disease more efficient. 3D convolutional neural networks (3D-CNN), taking only MRI brain scans and disease labels as input and trained end-to-end, are reported to be on par with the performance of traditional diagnostic methods in Alzheimer's disease classification [3, 4]. However, the process that 3D-CNNs use to arrive at their conclusions lacks transparency and cannot straightforwardly provide reasoning and explanations as human experts do in diagnosis. It is therefore difficult for human practitioners to trust such systems in evidence-centered areas like medical research.

The goal of this study is to break into the black box of 3D-CNNs for Alzheimer's disease classification. Particularly, we develop techniques to produce visual explanations that can indicate a 3D-CNN's spatial attention on MRI brain scans when making predictions. Our approaches give diagnosticians a better understanding of the behaviors of 3D-CNNs and provide greater confidence about integrating them into automated Alzheimer's disease diagnostic systems. In summary, the contributions of this study are as follows:

- We propose a hierarchical MRI image segmentation based approach for sensitivity analysis of 3D-CNNs, which can discriminate the importances of homogeneous brain regions at different levels for Alzheimer's disease classification.
- We extend two state-of-the-art approaches for explaining CNNs in 2D natural image classification to 3D MRI images, which can track the spatial attention of 3D-CNNs when predicting Alzheimer's disease.
- We compare the developed approaches qualitatively by examining the visual explanations generated. We also conduct quantitative comparisons for their ability to localize important parts of the brain in diagnosing Alzheimer's disease.

The rest of the paper is organized as follows. Section 2 surveys related work for this study. Section 3 describes the methods development, data, and experimental setup. Section 4 presents the qualitative and quantitative comparisons for proposed methods. Section 5 presents study conclusions.

2 Related Work

Works that are closely connected to this study are divided into three parts: 3D-CNNs for Alzheimer's disease classification, brain MRI segmentation, and visualizing and understanding CNNs for natural image classification.

3D-CNNs for Alzheimer's Disease Classification There are two major methods for using 3D convolutional neural networks for Alzheimer's disease classification from brain MRI scans. One uses 3D-CNNs to automatically extract generic features from MRIs and build other classifiers on top of them [5, 6]. The other trains the 3D-CNNs in an end-to-end manner that only takes MRI scans and labels as input [3, 4]. Both approaches achieve comparable performance [3]. The user has more control over the first method and thus can understand it better. The latter needs little input from humans so that it is easier to use.

Brain MRI Segmentation As one of the fundamental problems in neuroimaging, brain segmentation is the building block for many Alzheimer's disease diagnosis methods. Semantic segmentation methods such as FreeSurfer [7] enable brain volume calculations from MRI scans of Alzheimer's disease subjects [8]. Unsupervised hierarchical segmentation methods detect homogeneous regions and separate them from coarse to finer levels, providing more flexibility for multilevel analysis than the one-level semantic segmentation [9, 10].

Visualizing and Understanding CNNs for Natural Image Classification To explain the superior image classification performance for 2D-CNNs, researchers incorporate the spatial structure of the convolutional layer to visualize the discriminative object from activation maps [11, 12]. Sensitivity analysis by measuring the change of output class probability due to perturbed input is another popular method because it is not subject to the architectural constraints of CNNs. LIME, or local interpretable model-agnostic explanations [13], is a regression-based sensitivity analysis approach that examines perturbed superpixels to make CNN results more interpretable. The perturbed superpixels could be further learned to be more semantically meaningful [14, 15]. All these methods create a 2D spatial heatmap as a visual explanation that indicates where the CNN has focused to make its predictions. These can be extended to 3D for Alzheimer's disease classification.

3 Method

In this section, we describe the methods that can produce visual explanations of predictions of Alzheimer's disease from brain MRI scans by deep 3D convolutional neural networks (3D-CNNs). First, we summarize the deep learning models we deploy for the Alzheimer's disease classification task. Then, we present the brain MRI data for the study and describe how we use the data in experiments. Finally, we introduce the three approaches that we develop for explaining the 3D-CNNs, which are sensitivity analysis by 3D ultrametric contour map (SA-3DUCM), 3D class activation mapping (3D-CAM), and 3D gradient-weighted class activation mapping (3D-Grad-CAM).

3.1 Architecture of Deep 3D Convolutional Neural Networks

The architecture of the deep 3D convolutional neural networks (3D-CNN) for Alzheimer's disease classification in this study are based on the network architectures proposed by Korolev et al.[4]. Particularly, two types of 3D-CNNs are built for classifying brain MRI scans from an Alzheimer's disease cohort (AD) and a normal cohort (NC). The design ideas for both types of 3D-CNNs are rooted in successful 2D natural image classification models, specifically, VGGNet, the Very Deep Convolutional Networks [16], and ResNet, the Deep Residual Networks [17].

3D Very Deep Convolutional Networks (3D-VGGNet) VGGNet stacks many layer blocks containing narrow convolutional layers followed by max pooling layers. The 3D very deep convolutional network (3D-VGGNet) [4] for Alzheimer's disease classification is a direct application of this idea to 3D brain MRI scans. It contains four blocks of 3D convolutional layers and 3D max pooling layers, followed by a fully connected layer, a batch normalization layer [18], a dropout layer [19], another fully connected layer, and the softmax output layer to produce the probabilities of disease in the Alzheimer's disease cohort (AD) and the normal cohort (NC). The full network architecture of 3D-VGGNet is visualized in Figure 1 (left). To optimize model parameters, the ADAM optimizer [20] is used with a learning rate of 0.000027, a batch size of 5, and 150 training epochs. The two-class cross-entropy calculated from the probabilities output by the softmax layer and the ground-truth labels are used as loss functions.

3D Deep Residual Networks (3D-ResNet) Residual network is the most important building block of the state-of-the-art of 2D natural image classification [17, 21]. 3D deep residual networks (3D-ResNet) [4] for Alzheimer's disease classification prove their effectiveness in the 3D domain. We deploy this important type of 3D-CNN in this study and try to explain its predictions. Specifically, a six-residual-block architecture is built. Each residual block consists of two 3D convolutional layers with $3 \times 3 \times 3$ filters that have a batch normalization layer and a rectified-linear-unit nonlinearity layer (ReLU) [22] between them. Skip connections (identity mapping of a residual block) add a residual block element-by-element to the following residual block, explicitly enabling the following block to learn a residual mapping rather than a full mapping. This eases the learning process for deeper architectures and results in better performance. The full architecture of 3D-ResNet is depicted in Figure 1 (middle). For optimization, Nesterov accelerated stochastic gradient descent [23] is used. Optimization parameters are set as 0.001 for learning rate, 3 for batch size, and 150 for training epochs. The same loss function as 3D-VGGNet, the two-class cross-entropy function, is used.

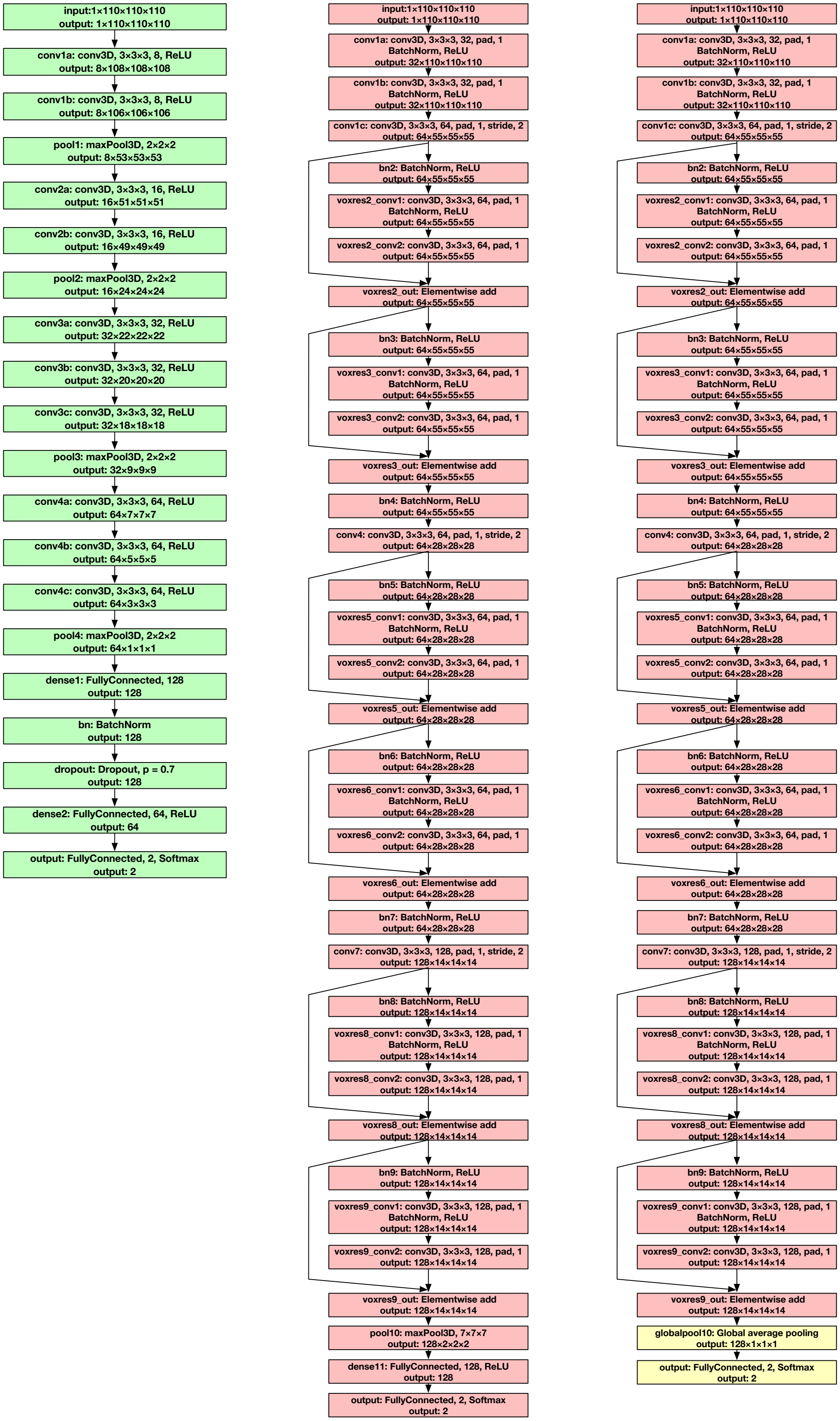


Figure 1: Left: The architecture of 3D-VGGNet; **Middle:** The architecture of 3D-ResNet; **Right:** The modified architecture of 3D-ResNet with global average pooling layer, 3D-ResNet-GAP, to produce 3D class activation mapping (3D-CAM). The only difference is that a global average pooling layer directly outputs to the softmax output layer (yellow boxes), replacing the original max pooling and fully connected layers.

3.2 Data and Experiment Setup

Brain MRI scans from the Alzheimer's Disease Neuroimaging Initiative [*] (ADNI) [24] are used for this study. Specifically, we used data from the "spatially normalized, masked, and N3-corrected T1 images" category to train the 3D-VGGNet and 3D-ResNet models to classify MRI scans from the Alzheimer's disease cohort (AD) and the normal cohort (NC). Each brain MRI scan is a 3D tensor of intensity values with size 110 × 110 × 110. As one subject could have more than one MRI scan in the database, to avoid potential information leak between the training and testing dataset, we only include the earliest MRI associated with each subject for this study. As a result, 47 MRI scans from the Alzheimer's disease cohort (AD) and 56 MRI scans from the normal cohort (NC) are selected for this study. We randomly set aside eight MRI scans (5 AD, 3 NC) for later visual explanation analysis. The rest of the dataset is used for training and testing the deep 3D convolutional neural networks (3D-CNNs).

For training and testing the 3D-VGGNet and 3D-ResNet models, we conduct five-fold cross-validation for five different splits of the dataset, totaling 25 training and testing rounds. As the batch size parameters are chosen as small numbers for both models (five for 3D-VGGNet and three for 3D-ResNet), we enforce that each batch in training contains samples from both the Alzheimer's disease cohort (AD) and normal cohort (NC) to stabilize the training process by avoiding biased loss.

3.3 Explaining the 3D-CNNs

In this section, we describe the methods that we develop for explaining the predictions of the 3D-CNNs in detail. We first revisit a baseline method using sensitivity analysis that can shed light on 3D-CNNs' attention [4]. Then we show how we used an unsupervised 3D hierarchical volumetric image segmentation approach, the 3D ultrametric contour map (3D-UCM) [10], to improve the baseline, which we call sensitivity analysis by 3D ultrametric contour map (SA-3DUCM). Next, we describe how the successful 2D visual explanation method, class activation mapping (CAM) [11] and its generalization, gradient-weighted class activation mapping (Grad-CAM) [12], are extended to 3D to explain predictions from 3D MRI scans. We call the two extended approaches 3D-CAM and 3D-Grad-CAM, respectively. As we mentioned, there are two major ways to explain the predictions of deep convolutional neural networks. One way applies perturbations to data and conducts sensitivity analysis. The baseline method and proposed SA-3DUCM approach belong to this category. The other way utilizes the architectural properties of CNNs to heuristically track the attention of neural networks. 3D-CAM and 3D-Grad-CAM fall into this category.

Baseline Approach A baseline approach is proposed alongside the work of 3D-VGGNet and 3D-ResNet [4] to shed light on 3D-CNN's attention when classifying MRI scans. To be specific, for every voxel in the MRI scan, its 7 × 7 × 7 neighborhood is occluded from the image, and then the 3D-CNN re-evaluates the probability of Alzheimer's disease from the partially occuluded image. The change of probability is used as the importance of that voxel. More formally, for the brain MRI volume V and each voxel of V at (x, y, z), we occlude the neighborhood $V_{x-3:x+3,y-3:y+3,z-3:z+3}$, resulting in a perturbed MRI volume occluded around (x, y, z), denoted by $OV_{(x,y,z)}$. We want to measure the change of probability of Alzheimer's disease of $OV_{(x,y,z)}$, predicted by the 3D-CNN, compared to the original volume V. This change is assigned to the voxel at (x, y, z). For a 3D heatmap, C, of the same size as V, to store these changes of probabilities as the importance score for all the voxels, the magnitude at (x, y, z) of C is calculated by

$$C_{x,y,z} = |P(OV_{(x,y,z)}) - P(V)| \tag{1}$$

where $P(\cdot)$ is one forward pass of the 3D-CNN to evaluate the probability of Alzheimer's disease from the MRI volumes, and $|\cdot|$ is the absolute value function.

This approach is a direct application of the one-at-a-time sensitivity analysis at the single voxel level to test how the uncertainty of the output probability of the 3D-CNN could be assigned to different voxels of the MRI scan. This is straightforward to implement; however, this approach suffers from three important problems. First, the 7 × 7 × 7 cubical neighborhoods are not necessarily semantically meaningful and could be across different brain segments, e.g., half in cerebral cortex and half in white matter. Thus, occlusion of such an area results in an unaccountable change of output probability. Second, this approach could only capture the impact of the 7 × 7 × 7 local areas. The importances of larger or smaller areas are not tested. Third, as we evaluate a new output probability for each voxel, this approach is extremely computationally intensive. An MRI scan of size 110 × 110 × 110 has over 1 million voxels, requiring the same number of forward passes through the 3D-CNN, which could take hours even in GPU-assisted systems.

Sensitivity Analysis by 3D Ultrametric Contour Map (SA-3DUCM) We notice that the shortcomings of the baseline approach could be overcome by using a good segmentation of the brain volume instead of the 7 × 7 × 7

[*] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

local neighborhood around each voxel. Particularly, we occlude each segment in the segmentation, instead of the cubical neighborhoods, before re-evaluating the probabilities. To resolve each of the three problems of the baseline approach, the segmentation method should be semantically meaningful, hierarchical, and compact. Most specifically, to be semantically meaningful, the segmentation should separate different homogeneous parts of the brain volume well, e.g., separating cerebral cortex and white matter, so that changes of probability could be ascribed to specific segments. To be hierarchical, the segmentation method should provide a hierarchy of segmentations that capture both coarse level parts, such as the whole white matter, as well as finer level parts. In this way, we can test the importances for both small and large areas. To be compact, the segmentation method should avoid over-segmentation and generate a manageable number of segments for analysis. Thus, we can reduce the number of forward passes needed through the 3D-CNN from the number of voxels to the number of segments, which is usually three to four orders of magnitude less.

3D Ultrametric Contour Map (3DUCM) [10, 25] is an effective approach for unsupervised hierarchical 3D volumetric image segmentation, which is the 3D extension of the 2D state-of-the-art, Ultrametric Contour Map for natural image segmentation [26]. It provides compact hierarchical segmentation of high quality. For the brain MRI volume, V, it could generate a hierarchy of segmentation, $H = \{H_1, H_2, ..., H_N\}$, where each level $H_n = S_1^n \cup S_2^n \cup ... \cup S_{K_n}^n$ is a full segmentation of the volume V. We occlude each segment S_k^n , $k = 1, 2, ..., K_n$, $n = 1, 2, ..., N$, in V, denoting each resulting volume by OV_k^n , and re-evaluate the probability of Alzheimer's disease through one forward pass of the 3D-CNN. The change of probabilities compared to what is obtained from the original volume, $|P(OV_k^n) - P(V)|$, is assigned to every voxel in S_k^n. Since each voxel belongs to one segment at each level of the hierarchy, each voxel gets N quantities from the calculation, where N is the number of levels in the segmentation hierarchy. We compute the average quantity from the N quantities as the importance score for each voxel and store it in a heatmap C. So for a voxel of V at (x, y, z), assuming that it belongs to $S_{k_n}^n$, for each level of hierarchy H_n, we calculate the importance score for it as

$$C_{x,y,z} = \frac{1}{N} \sum_{n=1}^{N} |P(OV_{k_n}^n) - P(V)| \tag{2}$$

Since the 3DUCM hierarchical segmentation usually provides homogeneous segments of the brain MRI, we expect the importance heatmap C to distinguish important brain parts for Alzheimer's disease classification. In terms of computational burden, each level of the hierarchy contains at most hundreds of segments, and the hierarchy itself is no more than 20 levels. Thus, the number of forward passes needed to re-evaluate the probabilities is greatly reduced.

3D Class Activation Mapping (3D-CAM) One major problem with one-at-a-time sensitivity analysis based methods (baseline and SA-3DUCM) is that the correlations and interactions between segments of MRI volume are ignored. Although using the hierarchical segmentation method can cover most semantic segments from finer to coarser level, we cannot guarantee all combinations are tested. Therefore, we turn to methods based on the architectural properties of the 3D-CNN that directly visualize the activations of convolutional layers when predictions are made. Class activation mapping [11] designs a global average pooling layer on top of convolutional layers in natural images classification, which enables remarkable localization performance on important objects in the images in spite of the fact that the CNN is trained on image-level labels. This fits our problem well. Our Alzheimer's disease labels (Alzheimer's disease cohort (AD) and normal cohort (NC)) are used at MRI scan level during the training of the 3D-CNNs. Our goal is to obtain visual explanations that can highlight brain parts important for Alzheimer's disease classification. Thus, extending class activation mapping to 3D provides a way to do this.

The idea of class activation mapping is that the last convolution layer of the CNN contains the spatial information indicating discriminative regions to make classifications. To visualize these discriminative parts, class activation mapping creates a spatial heatmap out of the activations from the last convolutional layer. Specifically, class activation mapping adopts a global average pooling layer between the final convolutional layer and output layer, which enables projection of class weights of the output layer onto the activation maps in the convolutional layer. The 3D extension of class activation mapping based on 3D-ResNet is shown in Figure 1 (right). Instead of using a max pooling layer and a fully connected layer before output, the modified 3D-ResNet only uses a global average pooling layer (3D-ResNet-GAP). To be specific, for a given MRI volume V and a 3D-CNN, let $f_u(x, y, z)$ be the activation of unit u in the last convolutional layer at location (x, y, z). The global average pooling for unit u is $F_u = \frac{1}{Z} \sum_{x,y,z} f_u(x, y, z)$, where Z is the number of voxels in the corresponding convolutional layer. As the global average pooling layer is directly connected to the softmax output layer, by the definition of the softmax function, the probability of Alzheimer's disease, $P(V)$, given by

$$P(V) = \frac{\exp(\sum_u w_u^{AD} F_u)}{\exp(\sum_u w_u^{AD} F_u) + \exp(\sum_u w_u^{NC} F_u)} \tag{3}$$

where w_u^{AD} and w_u^{NC} are the class weights in the output layer for the Alzheimer's disease cohort (AD) and the normal cohort (NC), respectively. We ignore the bias term here because its impact is minimal on classification performance. Essentially, $\sum_u w_u^{AD} F_u$ and $\sum_u w_u^{NC} F_u$ are the class scores for AD and NC cohorts, respectively. By extending F_u in the class score, we have

$$\text{Score}(AD) = \sum_u w_u^{AD} F_u = \sum_u w_u^{AD} \frac{1}{Z} \sum_{x,y,z} f_u(x,y,z) = \frac{1}{Z} \sum_{x,y,z} \sum_u w_u^{AD} f_u(x,y,z) \tag{4}$$

The $\sum_u w_u^{AD} f_u(x,y,z)$ part of the quantity is defined for every spatial location (x,y,z) and their sum is proportional to the class score for Alzheimer's disease. As areas significantly negatively contributing to the class score are also important, we adopt the absolute value and define the class activation mapping for the AD cohort as

$$\text{3D-CAM}_{x,y,z}(AD) = |\sum_u w_u^{AD} f_u(x,y,z)| \tag{5}$$

which is essentially a heatmap of weighted sums of activations in every location (x,y,z) and can be easily calculated by one forward pass when the volume V is provided.

Though 3D-CAM is easy to obtain, and we expect it to highlight the important spatial areas for classification, there are two potential problems with this approach. First, as we modify the 3D-CNN architecture with the global average pooling layer, we need to re-train the model, possibly affecting the classification performance. Second, the resolution of the class activation mapping is of the same size as the last convolutional layer. We need to upsample it to the original MRI scan size to identify the discriminative regions, which means we would lose some details in the resulting heatmap. One solution could be to remove more layers and build the global average pooling layers on convolutional layers with higher resolution. But this could further decrease the classification performance.

3D Gradient-Weighted Class Activation Mapping (3D-Grad-CAM) To overcome class activation mapping's shortcoming of decreased classification performance, its generalization, gradient-weighted class activation mapping, is proposed in natural image classification [12]. This approach does not need to modify the 3D-CNN's architecture and thus will do no harm to classification performance. Since no re-training is required, it is more efficient to deploy in deep learning systems. The core idea is still to identify the important activations from feature maps in convolutional layers. Using the same notation as the previous part, we first calculated the gradient of the Score(AD) with respect to the activation of unit u at location (x,y,z), $f_u(x,y,z)$, in the last convolutional layer. Then, we use the global average pooling of the gradients, denoted by a_u^{AD}, as the importance weights for unit u for the Alzheimer's disease cohort (AD). That is,

$$a_u^{AD} = \frac{1}{Z} \sum_{x,y,z} \frac{\partial \text{Score}(AD)}{\partial f_u(x,y,z)} \tag{6}$$

where Z is the number of voxels in the corresponding convolutional layer. Then, we combined the unit weights with the activations, $f_u(x,y,z)$, to get the heatmap of 3D gradient-weighted class activation mapping.

$$\text{3D-Grad-CAM}_{x,y,z}(AD) = |\sum_u a_u^{AD} f_u(x,y,z)| \tag{7}$$

3D-Grad-CAM could be applied to a wider range of 3D-CNNs than 3D-CAM as long as the 3D-CNN has a fully convolutional layer. Also, it has been proven in 2D applications that CAM is a special case of Grad-CAM with the global average pooling layer [12]. It does not require re-training so it quickly generates the 3D-Grad-CAM heatmap with just one forward pass. However, 3D-Grad-CAM still suffers from the low resolution problem because the 3D-Grad-CAM is a coarse heatmap of the same size as the last convolutional layer. We could have calculated it with gradients and activations from lower convolutional layers, but there is no guarantee that the spatial activations wouldn't change in the upper layers.

In summary, in this section, we introduce four approaches to obtain visual explanation heatmaps for predictions from 3D-CNNs. The baseline approach and sensitivity analysis by 3D ultrametric contour map (SA-3DUCM) are completely model-agnostic and can handle any type of 3D-CNNs, but they might have problems with correlations and interactions between different segments of the brain volume. 3D class activation mapping (3D-CAM) and 3D gradient-weighted class activation mapping (3D-Grad-CAM) are weighted visualizations of the activation maps in the convolutional layer, which avoids dealing with the correlations and interactions problem. However, they are limited by the low resolution of the convolutional layers. Upsampled heatmaps might not be able to provide enough detail to accurately identify important regions. For computational efficiency, the baseline approach is the slowest because it does a forward pass for every voxel. 3D-CAM only needs one forward pass to generate the heatmap, but it requires

Method	AUC	ACC
3D-VGGNet	0.863±0.056	0.766±0.095
3D-ResNet	0.854±0.079	0.794±0.070
3D-ResNet-GAP	0.643±0.110	0.614±0.100
3D-ResNet-Shallow-GAP	0.751±0.083	0.585±0.122

Table 1: Classification performance of 3D-CNNs

very time-consuming re-training. SA-3DUCM needs a few hundred forwarded passes. 3D-Grad-CAM is the best because it does not require re-training and only needs one forward pass when generating the heatmap. In the next section, we will compare the models' performances in identifying of discriminative brain parts for Alzheimer's disease classification from MRI scans.

4 Results

In this section, we will present the classification performance of 3D-CNNs, visual comparisons of the heatmaps generated by the proposed visual explanation approaches, and a quantitative benchmark for the localization ability of the heatmaps in identifying important brain parts for Alzheimer's disease classification.

4.1 Alzheimer's Disease Classification Performance

We compare the classification performance of four different 3D-CNNs. These include 3D-VGGNet and 3D-ResNet as described. By implementing the 3D-CAM, we have a modified 3D-ResNet with global average pooling layer (GAP) as shown in Figure 1 (right), denoted as 3D-ResNet-GAP. The counterpart for 3D-VGGNet is not included because the classification performance drops too much, compared to 3D-VGGNet. Additionally, to obtain a higher resolution 3D-CAM, we remove the layers from `conv4` to `voxres9_out`, resulting in a shallow version of 3D-ResNet-GAP, which we call 3D-ResNet-Shallow-GAP. All four 3D-CNNs are trained for classifying the Alzheimer's cohort (AD) in comparison to the normal cohort (NC). Classification performance is measured by the area under the ROC curve (AUC) and classification accuracy (ACC). Cross-validation as described in Section 3.2 is conducted. Average AUC and ACC and their standard deviations are reported. The results are presented in Table 1. 3D-VGGNet and 3D-ResNet achieve good classification performances. However, there is a substantial drop in performance for 3DResNet-GAP and 3D-ResNet-Shallow-GAP, which means the global average pooling layer have a negative effect on classification performance.

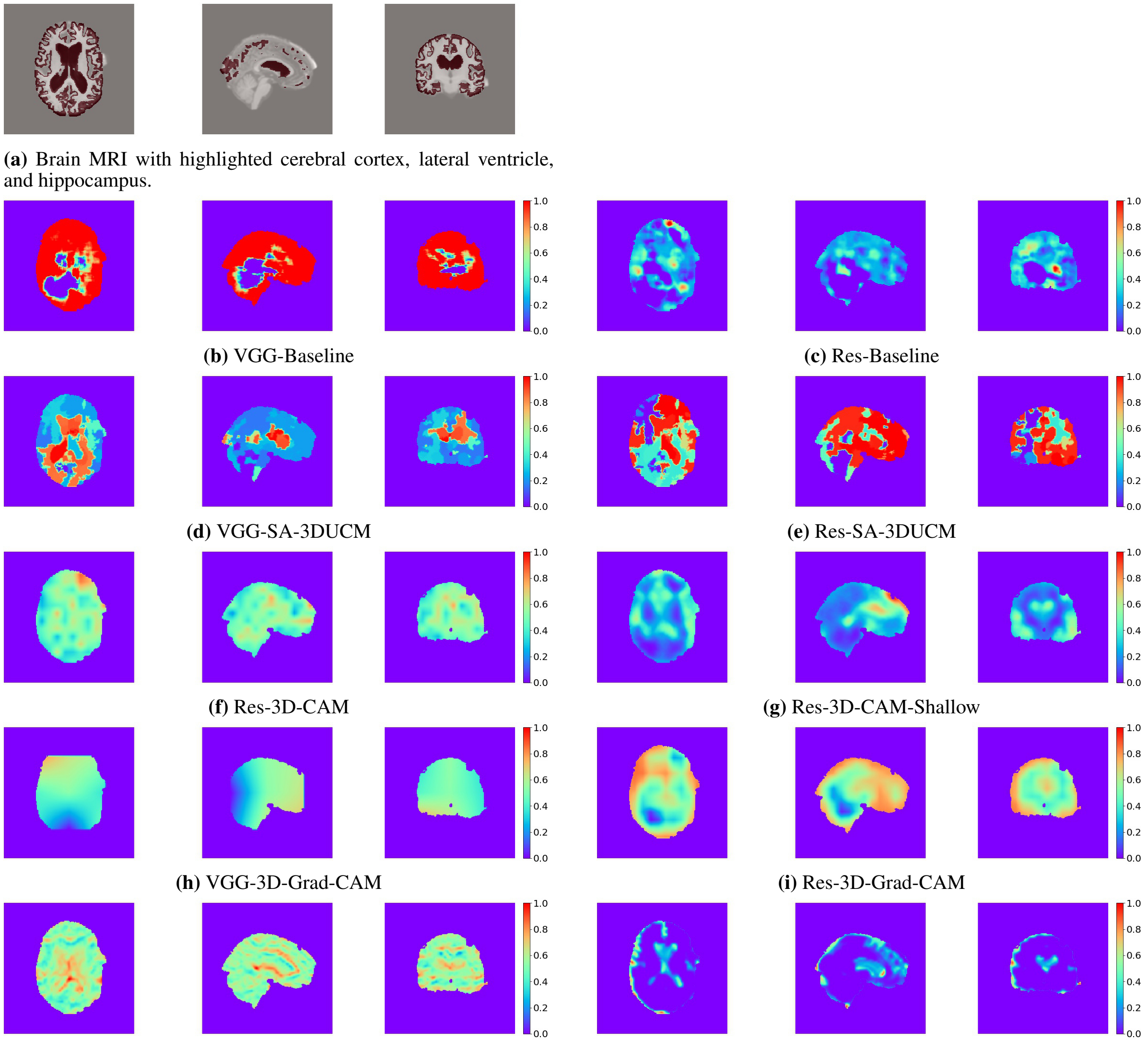
4.2 Qualitative Comparison for Visual Explanations

To visually check the quality of heatmaps generated by the introduced visual explanation methods, we take one MRI scan from the set-aside data for visual explanation analysis and present the heatmap from the horizontal, sagittal, and coronal sections. For comparison, we present the input brain MRI volume (Figure 2a) with highlighted areas of cerebral cortex, lateral ventricle, and hippocampus. These parts are believed to be important for Alzheimer's disease diagnosis by physicians [27, 28]. The ground-truth cerebral cortex, lateral ventricle, and hippocampus regions are segmented by the FreeSurfer software [7].

Baseline The resulting heatmaps are labeled as VGG-Baseline and Res-Baseline and are presented in Figure 2b and Figure 2c, respectively. We can see from the figures that in both situations, the baseline method does not find the important areas. The heatmaps are irregularly shaped because heterogeneous regions are used for sensitivity analysis. Overall, the baseline method fails to identify discriminative regions.

SA-3DUCM After incorporating hierarchical segmentations into sensitivity analysis, we find that the results greatly improves, compared to baseline. Figure 2d presents the heatmap made by applying SA-3DUCM to 3D-VGGNet (VGG-SA-3DUCM), and the heatmap in Figure 2e is made by applying SA-3DUCM to 3D-ResNet (Res-SA-3DUCM). In both situations, the approach differentiates the importances of different homogeneous regions. There are clear boundaries separating the regions. The lateral ventricle area stands out as the most discriminative part. However, the cerebral cortex areas are not well identified. This is because cerebral cortex is widely and loosely distributed in the brain so the cerebral cortex is usually not segmented as one area in hierarchical segmentations. SA-3DUCM tested the importance of different segments one by one. Thus, it is not able to capture the correlations between all segments that belong to the cerebral cortex.

3D-CAM We only apply 3D class activation mapping (3D-CAM) to 3D-ResNet because 3D-VGGNet loses too much classification performance after using the global average pooling layer. The class activation mapping heatmap of 3D-ResNet-GAP is labeled as Res-3D-CAM and is presented in Figure 2f. The heatmap is blurry because it is upsampled from a 14 × 14 × 14 coarse heatmap. To get a higher resolution 3D class activation mapping heatmap, Figure 2g (Res-3D-CAM-Shallow) is obtained from 3D-ResNet-Shallow-GAP with more convolutional layers removed. It is upsampled from a 55 × 55 × 55 heatmap and thus provides more detail. It identifies the lateral ventricle and most parts of the cortex as important areas, which matches the human experts' approach.



(a) Brain MRI with highlighted cerebral cortex, lateral ventricle, and hippocampus.

(b) VGG-Baseline

(c) Res-Baseline

(d) VGG-SA-3DUCM

(e) Res-SA-3DUCM

(f) Res-3D-CAM

(g) Res-3D-CAM-Shallow

(h) VGG-3D-Grad-CAM

(i) Res-3D-Grad-CAM

(j) VGG-3D-Grad-CAM-Shallow

(k) Res-3D-Grad-CAM-Shallow

Figure 2: Horizontal, sagittal, and coronal view of the brain MRI and the visual explanation heatmaps.

3D-Grad-CAM The 3D gradient-weighted class activation mapping (3D-Grad-CAM) also has low resolution problems, especially when it is applied to 3D-VGGNet. Because the last convolutional layer of 3D-VGGNet is only of size 3 × 3 × 3, the resulting heatmap VGG-3D-Grad-CAM barely provides any information (Figure 2h). When we apply the same approach to a lower convolutional layer, conv2b, in 3D-VGGNet, the resulting heatmap, VGG-3D-Grad-CAM-Shallow (Figure 2j), is able to highlight part of the lateral ventricle. 3D-ResNet has the same situation. Res-3D-Grad-CAM (Figure 2i) and Res-3D-Grad-CAM-Shallow (Figure 2k) are generated by the 3D-Grad-CAM approach applied to voxres9_out (last convolutional layer) and bn4 (an intermediate convolutional layer) of 3D-ResNet. They are of size 14 × 14 × 14 and 55 × 55 × 55, respectively. Though both of them identify most of the lateral ventricle and the cerebral cortex as discriminative, Res-3D-Grad-CAM-Shallow is of higher resolution and more accurate. However, as we stated, upper convolutional layers could change the activation maps from the lower convolutional layers. Thus sometimes, we may not trust the heatmap from lower layers as a good representation of spatial attention of the 3D-CNN.

To summarize the qualitative comparisons, SA-3UCM has the same resolution as the original MRI volume and differentiates homogeneous regions well. However, it fails to identify the correlations from the fragmented cerebral cortex segments because of the one-at-a-time process in sensitivity analysis. 3D-Grad-CAM and 3D-CAM both produce more blurry heatmaps than SA-3DUCM because of upsampling. But they are able to highlight the cerebral cortex that

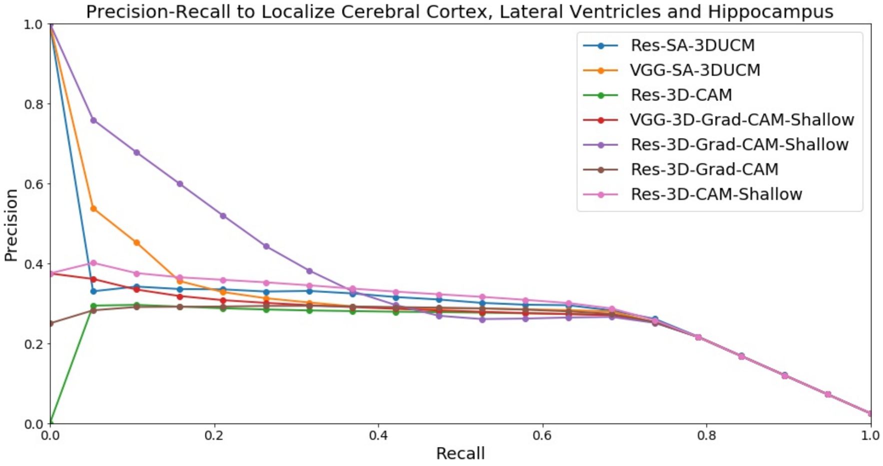


Figure 3: Precision-recall curve to localize cerebral cortex, lateral ventricle, and hippocampus regions using heatmaps.

is loosely distributed in the brain.

4.3 Quantitative Comparison for Localization

Visual comparisons of the heatmap give us a general idea how well different visual explanation methods work. But we wonder how well these heatmaps could localize important regions such as cerebral cortex, lateral ventricle, and hippocampus. To quantitatively compare localization ability, we plot the precision-recall curve for the heatmaps that we have visualized in the previous section to identify cerebral cortex, lateral ventricle, and hippocampus regions from the 8 MRI scans that are set aside for visual explanation analysis. VGG-Baseline, Res-Baseline, and VGG-3D-Grad-CAM are not included because they do not generate usable heatmaps in the visual comparisons. The results are presented in Figure 3.

From the results, we can see VGG-SA-3DUCM, Res-SA-3DUCM, and Res-3D-Grad-CAM-Shallow have high precision on the low recall end. This matches our visual comparisons as SA-3DUCM method puts the homogeneous lateral ventricle regions on top, and Res-3D-Grad-CAM-Shallow identifies cerebral cortex and lateral ventricle parts with high accuracy. However, the precision drops for all methods on the high recall end, implying no method is close to perfectly identifying all important regions. The reasons would be different. SA-3DUCM could not discriminate the cerebral cortex because of fragmented segments. 3D-CAM and 3D-Grad-CAM are limited by low resolution of the heatmaps.

Overall, both qualitative and quantitative comparisons indicate that all visual explanation methods have some limitations. The correct method may be chosen based on the specific goals. When the goal is to get the importance for a homogeneous region, SA-3DUCM is more suitable. If tracking the attention of the 3D-CNN is the goal, 3D-Grad-CAM is the preferred choice. Generally 3D-Grad-CAM is better than 3D-CAM because it does not modify the 3D-CNN architecture, requires less computation, and better localizes important regions.

5 Conclusion and Discussion

In this study, we develop three approaches for producing visual explanations from 3D-CNNs for Alzheimer's disease classification. All approaches can highlight important brain parts for diagnosis. However, they have limitations in different aspects. The one-at-a-time sensitivity analysis procedure of SA-3DUCM is not able to handle correlated or interacting images segments, causing underestimation of attention in the loosely distributed area such as cerebral cortex in our case. 3D-CAM and 3D-Grad-CAM build heatmaps from convolutional layer activations that have lower resolution than the original MRI scan, resulting in loss of details and decreased localization accuracy. Therefore, we suggest users choose the right approach based on their use cases for MRI analysis.

Though all approaches are developed for Alzheimer's disease classification, they are generic enough for other type of 3D image analysis. SA-3DUCM is completely model agnostic and can adapt to any classifiers taking 3D volumetric images as input. 3D-CAM and 3D-Grad-CAM can work on any deep learning model that has a 3D convolutional layer. They could be applied to other types of 3D medical images or even video analysis.

One common limitation of these approaches is that the visual explanation is still one step away from fully understanding the 3D-CNN. Human experts measure cerebral cortex thickness as a biomarker for diagnosis [29]. In the generated visual explanations, there is no such explicit summarized representation on top of the visual attention from the cerebral cortex. This leads to our future work of explicit biomarker representation learning from medical imaging to fully interpret the 3D-CNNs.

ACKNOWLEDGMENTS

This work is partially supported by NSF 1743050 to A.R. and S.R..

References

1. Jiaquan Xu, Sherry L Murphy, Kenneth D Kochanek, and Elizabeth Arias. Mortality in the united states, 2015. 2016.
2. Guy M McKhann, David S Knopman, Howard Chertkow, Bradley T Hyman, Clifford R Jack, Claudia H Kawas, William E Klunk, Walter J Koroshetz, Jennifer J Manly, Richard Mayeux, et al. The diagnosis of dementia due to alzheimers disease: Recommendations from the national institute on aging-alzheimers association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 7(3):263–269, 2011.
3. Alexander Khvostikov, Karim Aderghal, Jenny Benois-Pineau, Andrey Krylov, and Gwenaelle Catheline. 3d cnn-based classification using smri and md-dti images for alzheimer disease studies. *arXiv preprint arXiv:1801.05968*, 2018.
4. Sergey Korolev, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova. Residual and plain convolutional neural networks for 3d brain mri classification. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017.
5. Heung-Il Suk, Seong-Whan Lee, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. *NeuroImage*, 101:569–582, 2014.
6. Ehsan Hosseini-Asl, Georgy Gimel'farb, and Ayman El-Baz. Alzheimer's disease diagnostics by a deeply supervised adaptable 3d convolutional network. *arXiv preprint arXiv:1607.00556*, 2016.
7. Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.
8. Emma R Mulder, Remko A de Jong, Dirk L Knol, Ronald A van Schijndel, Keith S Cover, Pieter J Visser, Frederik Barkhof, Hugo Vrenken, Alzheimer's Disease Neuroimaging Initiative, et al. Hippocampal volume change measurement: quantitative assessment of the reproducibility of expert manual outlining and the automated methods freesurfer and first. *Neuroimage*, 92:169–181, 2014.
9. Jason J Corso, Eitan Sharon, Shishir Dube, Suzie El-Saden, Usha Sinha, and Alan Yuille. Efficient multilevel brain tumor segmentation with integrated bayesian model classification. *IEEE transactions on medical imaging*, 27(5):629–640, 2008.
10. Chengliang Yang, Manu Sethi, Anand Rangarajan, and Sanjay Ranka. Supervoxel-based segmentation of 3d volumetric images. In *Asian Conference on Computer Vision*, pages 37–53. Springer, 2016.
11. Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2921–2929. IEEE, 2016.
12. Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *See https://arxiv. org/abs/1610.02391 v3*, 7(8), 2016.
13. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.
14. Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *arXiv preprint arXiv:1704.03296*, 2017.
15. Chengliang Yang, Anand Rangarajan, and Sanjay Ranka. Global model interpretation via recursive partitioning. *arXiv preprint arXiv:1802.04253*, 2018.
16. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
17. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
18. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456, 2015.
19. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
20. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
21. Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5987–5995. IEEE, 2017.
22. Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
23. Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady AN USSR*, volume 269, pages 543–547, 1983.
24. Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877, 2005.
25. Xiaohui Huang, Chengliang Yang, Sanjay Ranka, and Anand Rangarajan. Supervoxel-based segmentation of 3d imagery with optical flow integration for spatiotemporal processing. *IPSJ Transactions on Computer Vision and Applications*, 10(1):9, 2018.
26. Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2011.
27. Kirsi Juottonen, Mikko P Laakso, Kaarina Partanen, and Hilkka Soininen. Comparative mr analysis of the entorhinal cortex and hippocampus in diagnosing alzheimer disease. *American Journal of Neuroradiology*, 20(1):139–144, 1999.
28. Qiwen Mu, Jingxia Xie, Zongyao Wen, Yaqin Weng, and Zhang Shuyun. A quantitative mr study of the hippocampal formation, the amygdala, and the temporal horn of the lateral ventricle in healthy subjects 40 to 90 years of age. *American Journal of Neuroradiology*, 20(2):207–211, 1999.
29. Bruce Fischl and Anders M Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences*, 97(20):11050–11055, 2000.