sensors

Article

A Low-Cost Three-Dimensional DenseNet Neural Network for Alzheimer's Disease Early Discovery †

Braulio Solano-Rojas * and Ricardo Villalón-Fonseca

CITIC-ECCI, Universidad de Costa Rica, San José 11501, Costa Rica; ricardo.villalon@ucr.ac.cr
* Correspondence: braulio.solano@ucr.ac.cr; Tel.: +506-83670655

† This paper is an extended version of our paper published in ICOST 2020, Hammamet, Tunisia, 24–26 June 2020.

Abstract: Alzheimer's disease is the most prevalent dementia among the elderly population. Early detection is critical because it can help with future planning for those potentially affected. This paper uses a three-dimensional DenseNet architecture to detect Alzheimer's disease in magnetic resonance imaging. Our work is restricted to the use of freely available tools. We constructed a deep neural network classifier with metrics of $\overline{0.86}$ mean accuracy, $\overline{0.86}$ mean sensitivity (micro-average), $\overline{0.86}$ mean specificity (micro-average), and $\overline{0.91}$ area under the receiver operating characteristic curve (micro-average) for the task of discriminating between five different disease stages or classes. The use of tools available for free ensures the reproducibility of the study and the applicability of the classification system in developing countries.

Keywords: Alzheimer's disease; magnetic resonance imaging; optical sensors; image processing; deep learning; computer-aided detection; computer-aided diagnosis

Citation: Solano-Rojas, B.; Villalón-Fonseca, R. A Low-Cost Three-Dimensional DenseNet Neural Network for Alzheimer's Disease Early Discovery. *Sensors* **2021**, *21*, 1302. https://doi.org/10.3390/s21041302

Academic Editor: Mohamed Jmaiel

Received: 21 December 2020
Accepted: 25 January 2021
Published: 11 February 2021

1. Introduction

Alzheimer's Disease (AD) is the most prevalent dementia among the elderly population [1]. AD is a neurodegenerative disease without a known cure. Therefore, early detection strategies have become an important research focus in the absence of an effective treatment for AD. The proper identification of AD patients allows for adequate planning for the future and the necessary modifications to living accommodations and lifestyle [1]. Some dietary changes or medications have proven to slow disease progression [2]. Additionally, knowing AD's development can help families psychologically prepare for the future and the necessary changes in the attention that needs to be given to their family members [1].

The advancement of early AD detection has resulted in initiatives such as the Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI is a repository of images and biomarkers of healthy and AD-affected individuals available through their website (http://adni.loni.usc.edu). ADNI originated in 2003 as a public and private endeavor, led by Dr. Michael W. Weiner. ADNI's primary objective is to question whether medical images, other biomarkers, and clinical or neuropsychological evaluation can be united to scale AD's progress. Using the corpus of ADNI clinical images, it is possible to develop AD early detection software tools, commonly referred to as Computer-Aided Detection (CAD) [3]. CAD assists the clinical workers by providing a support decision system to help them detect diseases.

The potential of human error is a primary reason to develop software for the early detection of AD. The process of image-based early detection by health care professionals is affected by factors such as distractions, stress, fatigue, and inherent cognitive biases regarding the disease's specific conditions. Lee et al. [4] affirmed that about 75% of all medical mistakes were due to radiologists' diagnostic errors. Similarly, Graber et al. [5] observed that cognitive determinants contribute to approximately 74% of diagnostic errors.

Lee et al. [4] and Graber et al. [5] considered that the most common causes of diagnostic mistakes are inadequate systems, cognitive bias, fatigue, stress, and a high workload. Consequently, the development of CAD driven by artificial intelligence using Deep Neural Networks (DNNs) is increasing. These tools show remarkable proficiency in recognizing various diseases using clinical images. For example, the work by Rajpurkar et al. [6], where DNNs classified pneumonia in chest X-rays, demonstrated their power by outperforming radiologists.

The ability of DNNs for clinical image classification enables the development of diverse DNN architectures. These can be combined into computer-aided detection systems to identify disease progression and detection such as for AD. Hence, DNNs are relevant and a staple in clinical image analysis and diagnostics. A type of DNN, the Convolutional Neural Network (CNN), inspired by the inner workings of living organisms' visual cortex, is a recognized architecture for image analysis and is currently used in the field of computer vision. The scope of use of CNNs involves self-driving cars, drones, robotics, sports and recreation, intelligent surveillance and monitoring, and health and medicine [7].

In [8], we created a three-dimensional CNN architecture, specifically DenseNet-121, trained using the ADNI MRI image dataset. We measured the accuracy of detecting Alzheimer's disease with the implemented DNN architecture. This manuscript extends our previous work by further explaining our data preprocessing with the inclusion of figures. Additionally, we reevaluate our classifier using 16 runs with randomized data partitions instead of just one run. Finally, we include and discuss attention maps to convey what the constructed neural network perceives. Attention maps provide an interpretation of neural networks, often criticized for being black boxes that do not explain their output. Attention maps also help to analyze the neural network developed.

As in [8], we kept our low-cost approach. Our goal is to provide a technological artifact that can be used across many health care services to benefit people, particularly for developing countries with difficulties accessing specialized computing platforms.

This paper first provides some background definitions in Section 2 to support our work. In Section 3, we describe previous work in more detail. Then, in Section 4, we provide the methodology employed to realize this work. We present in Section 5 the results of the design choice. We discuss those results in Section 6. We finalize our discussion with concluding remarks and future work in Section 7.

2. Background

We begin with a brief review of the medical glossary used to provide a context for our research. We present distinct clinical grades of disease that we desire to classify. Later, we report two types of medical imaging used in the discovery and diagnosis of AD.

2.1. Clinical Disease Stages

Cognitively normal, significant memory concern, and mild cognitive impairment are the distinct stages before AD's clinical diagnosis.

2.1.1. Cognitively Normal

In the ADNI study, the control subjects are Cognitively Normal (CN) patients. They exhibit no symptoms of depression, mild cognitive impairment, or dementia. They are aging in a healthy manner [9].

2.1.2. Significant Memory Concern

Significant Memory Concern (SMC) is reported by the patient. SMC is quantified using the Cognitive Change Index and a Clinical Dementia Rating of zero. Subjective memory concerns are correlated with a greater possibility of progression, thereby lessening risk stratification amid normal controls and addressing the gap within healthy elderly controls and mild cognitive impairment. Nevertheless, SMC patients score inside the typical range for cognition [9].

2.1.3. Mild Cognitive Impairment

Mild Cognitive Impairment (MCI) patients have a subjective memory concern, either self-reported or through a clinician or informant. Nonetheless, everyday living activities are essentially maintained; there are no meaningful impairment levels in different cognitive domains, and no symptoms of dementia exist. Levels of MCI (early or late) are defined using the Wechsler Memory Scale Logical Memory II [9].

2.1.4. Alzheimer's Disease

AD is the most common form of dementia. Dementia is a generic name for memory and other cognitive capabilities' decline. Dementia is severe enough to affect everyday life. AD is a gradual disease, where dementia signs progressively worsen beyond several years. Individuals suffer the loss of the capacity to converse and react to their surroundings. Current medicines cannot stop the disease from advancing. Medications can temporarily delay the worsening of dementia signs and enhance life quality for those with AD and their caregivers [1].

Because we intend to assess if those stages are identified from medical imaging, expressly magnetic resonance imaging, we proceed to describe two medical imaging methods.

2.2. Medical Imaging

Medical imaging is the procedure and technique of producing visual representations of the human body's inside for clinical examination and medical intervention. We present two types of medical imaging, although we are particularly interested in Magnetic Resonance Imaging (MRI) as input for DNNs. We also introduce Positron Emission Tomography (PET) because it is sometimes a modality that accompanies MRI. Next, we explain what MRI and PET are.

2.2.1. Magnetic Resonance Imaging

MRI is a non-invasive imaging technique that generates a detailed volumetric anatomical visualization without harmful radiation of human tissues. It is applied frequently for treatment monitoring and disease detection and diagnosis. It is based on advanced technology that excites and detects the difference in the direction of the rotational axis of protons located in the water that constitutes living tissues [10].

2.2.2. Positron Emission Tomography

PET scans employ radiopharmaceuticals to produce volumetric images. This type of scan generates small particles named positrons. A positron is a particle with approximately equal mass to an electron, but oppositely charged. Positrons respond to electrons in the body, and when these two particles unite, they annihilate each other. This annihilation briefly generates an amount of energy in the configuration of two photons that shoot off in opposite directions. These photons are measured by the detectors in the PET scanner, and this information is employed to produce images of internal organs [11].

3. Previous Work

Our literature review focuses on the current state of deep learning for AD detection and how much further this field can be improved through convolutional neural networks. We employed IEEE (https://ieeexplore.ieee.org/) as the reference for artificial neural networks because, based on the journal rankings (https://www.scimagojr.com/) for the subject of artificial intelligence, it is the highest-ranked in both the SCImago Journal Rank (SJR) and H-Index. We used Google Scholar and Duck Duck Go for other publications.

In the IEEE Digital Library, we employed the search string "deep AND learning AND alzheimer AND mri" to identify the amount of cases where the classification and detection of AD used convolutional neural networks. We ran the query from the year 2016 to the year 2019. The search resulted in 81 documents recovered from the IEEE Digital

Library, including early access articles, journals, and conferences. As of 2020, the query result provided an additional 39 results. The new results indicate that the research field of Alzheimer's disease detection with CNNs remains relevant.

Previously, in [8], we selected by title, and if the title was cryptic, we also used the abstract. We focused on the implementation of convolutional neural networks. The selection resulted in 32 articles. We focused on literature that covered convolutional deep learning classifiers' study to identify AD in MRI and other modalities. We restricted the review to supervised learning, thus omitting architectures such as convolutional autoencoders. We also screened the 39 new articles and found 10 additional articles of interest. For the collected studies that met our selection criteria, we extracted the following information: (1) publication year, (2) architecture of the neural network, (3) if the MRI images were processed, (4) the modalities (number of inputs), (5) the number of classes used in the classifier, (6) accuracy, (7) sensitivity, (8) specificity, and finally, (9) the Area Under the Receiver Operating Characteristic curve (AUROC).

As stated in [8], we identified a concern when analyzing the collected papers and their data. Approximately 50% of papers report accuracy, but did not report sensitivity, specificity, or AUROC. It appears that there is a competition to achieve higher accuracy, although this measure is misleading. A classifier may report a high accuracy and still possess a low capability of correct prediction.

We concluded that the various research efforts are too distinct for an adequate comparison and contrast based on the collected information. We noticed the avoidance of multiclass classification. The most commonly reported classifier systems were binary, which usually results in higher accuracy. By increasing the number of classes, it is not uncommon to see the classifiers' performance decrease. However, multiclass models are more informative than binary classifiers. In our search, the multiclass classifiers identified had a maximum of four classes.

When looking for specific architectures, we noticed that there were not many entries that used the DenseNet architecture. We did encounter papers with other varieties of residual neural networks: three-dimensional VoxResNet [12], ResNet [13–15], and three-dimensional ResNet [16]. Only three documents used DenseNets, of which two [17,18] were three-dimensional, but with depthless DenseNets, and one [19] used deep DenseNets, but two-dimensional. The collected items' quantitative analysis did not generate a great contribution due to these defects. However, in the studies' review, we found documents of remarkable quality like [20]. We also judged that some of the manuscripts collected were not easily repeatable, for example [13].

Finally, when looking at the 10 additional papers acquired from the IEEE Digital Library, four used three-dimensional deep learning [21–24]. We believe that, as mentioned by [21], three-dimensional architectures are being initiated. We consider that this kind of architecture is growing in use. We also continued to find research efforts that only report high accuracy, but not other metrics [24–28]. Thus, the paradigm of obtaining the highest possible accuracy in classification remains. Last, we found it interesting that [21] reported MRI as a better modality than PET.

In contrast to previous research, we explored the process of constructing a multiclass neural network classifier system using tools available for free. Additionally, we report multiple performance metrics such as sensitivity, specificity, AUROC, and the model's accuracy, in contrast to focusing on reporting only the classifier's accuracy. Lastly, we are committed to reproducibility, describing our methods and parameters used in the next section.

4. Materials and Methods

This section describes how we collected the ADNI data and how we preprocessed them. Next, we introduce our development and how we created, using the Google Colaboratory tool, an Alzheimer's prediction model to achieve the goal of estimating the accuracy of the detection of Alzheimer's disease employing a three-dimensional DenseNet-121.

4.1. Data Retrieval

We used the beta advanced search functionality of ADNI with the following criteria: In Projects, we checked ADNI. In Research Group, we checked MCI, EMCI, AD, SMC, and CN. In Modality, we checked MRI. We only selected MRI and excluded PET based on an economic constraint criterion. PET requires radiopharmaceuticals. Therefore, it is more common to find MRI being used in economically restricted circumstances.

Regarding other search options, for Image Description, we used the string MPRAGE; for Acquisition Plane, we selected SAGITTAL; and finally, in Weighting, we selected T1. We kept the remaining search fields with default parameters. With these parameters, we collected 5556 magnetic resonance images with the demographics in Table 1.

Table 1. Demographics of the images.

Subject Age		Subject Cohort		Subject Sex	
50–75 years	2811	Cognitive Normal	1520	Male	2956
75–100 years	2745	Significant Memory Concern	186	Female	2599
		Early Mild Cognitive Impairment	1222	Unknown	1
		Mild Cognitive Impairment	1274		
		Late Mild Cognitive Impairment	636		
		Alzheimer's Disease	718		

The MRI acquired from ADNI is in the Digital Imaging and Communication On Medicine (DICOM) format. The images are a zipped archive of 55.5 GB, and the uncompressed files hold 138 GB. We decrease that quantity with image preprocessing, and we describe how and why in the following section.

4.2. Data Transformation

MRI data comprise groups of slices. We present in Figure 1 an extract of the MRI slides of a subject. The extract is not complete. We only show 42 slides out of 170. Each slice is an image, and the combination of images forms the MRI. All images or slices are a pixel matrix. Every slice has an associated spatial thickness because they reproduce reality. Furthermore, all pixels in each slice have a spacing, that is the space they represent. Consequently, the data are volumetric or rectangular cuboids.

Considering the volumetric nature of the data, we applied the next transformations to it. First, we transformed all volumetric pixels (voxels) to a spacing of $1 \times 1 \times 1$ millimeters. This transformation may add or delete slices or slice voxels. Next, we turned each slice into 256×256 voxels as follows. Some slices are not square. When they are not, we filled them in with black voxels. After they were equilateral, if they were not 256×256, we changed them to that dimension employing interpolation. Likewise, we made the cuboids have 256 slices utilizing interpolation. Cubes of $256 \times 256 \times 256$ resulted from this. From these cubes, we created a cut from Slice 40 to Slice 214, from Row 50 to Row 199, and from Column 40 to Column 239, to keep only the area with brain tissue. This cut allowed discarding the black voxels' edges and preserving the useful internal information (the brain). Because we made all the MRIs the same size, we assumed that the cut conserved to the brain. We did not apply methods like segmentation (splitting the brain using pattern recognition). Of those cuboids, we employed only half of the slices and half of every slice's rows and columns by dropping one in between for all. The latter diminished the dimensionality of the problem and the volume of the images considerably. Finally, we normalized the voxels' values to an interval of -1.0 to 1.0.

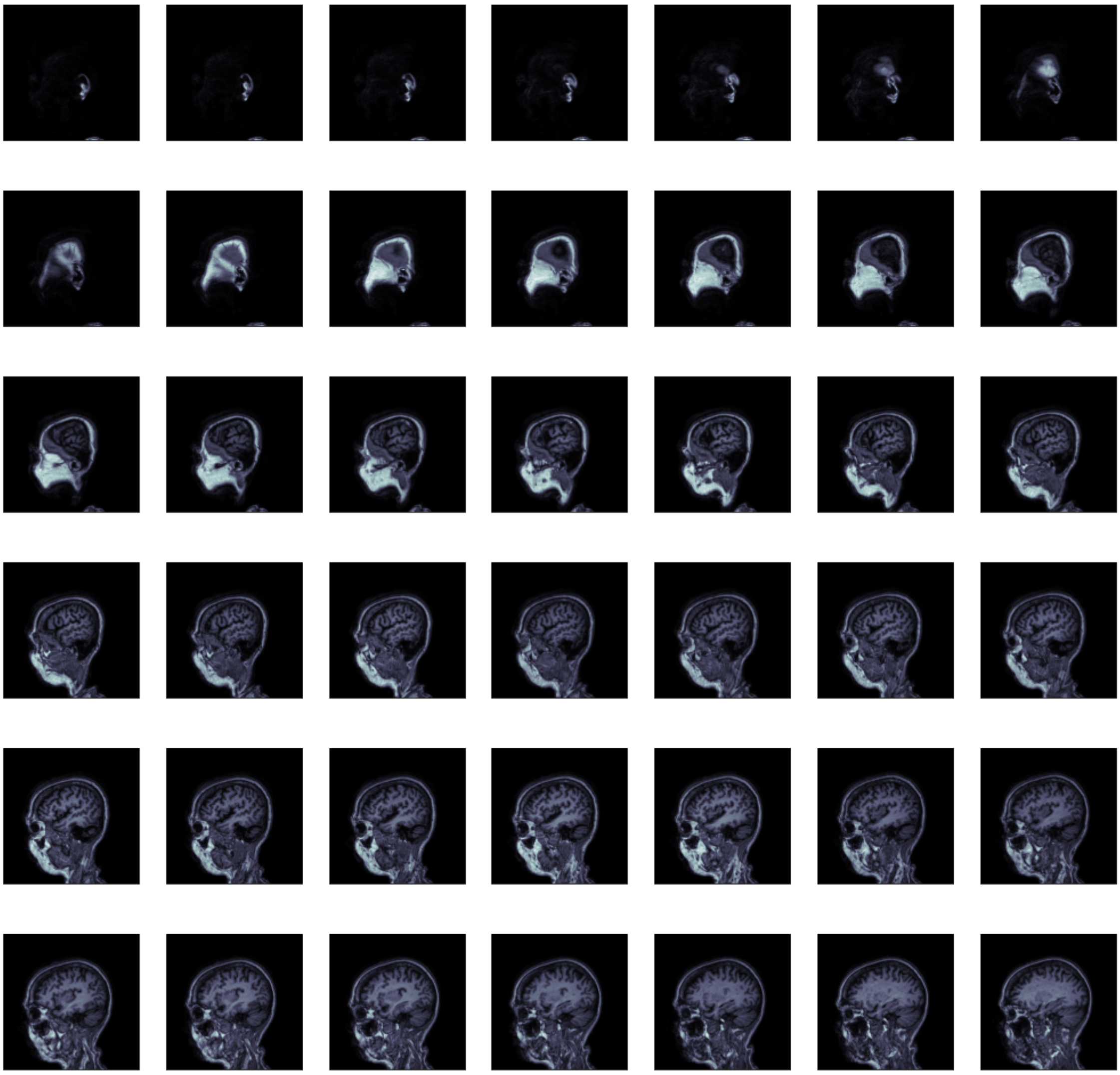


Figure 1. Volumetric data from a subject's MRI (extracted slice).

In Figure 2, we show an extract of a subject's processed slices to illustrate these transforms. It is the same subject as in Figure 1. The transform was still volumetric data. The extract is not complete because we only show 42 slices out of 87. As can be seen from the slices, we cut rectangles from the original slices, and these rectangles are compressed, but still show the brain patterns.

Data transformation can be done both previously or online. We implemented both. Nevertheless, to sustain a low-cost goal, we employed a script to apply the transformation previously to the job of neural network training. We used the already transformed MRI. Preprocessing beforehand may be achieved on a laptop or desktop computer. Although that would take hours, it is not a process that would demand more than a day on contemporary hardware.

After data transformation, the images were only 13.5 GB. We decreased the size of the images by approximately more than ten times. This compression helped minimize our development's neural network storage needs and training time. The next section explains that development.

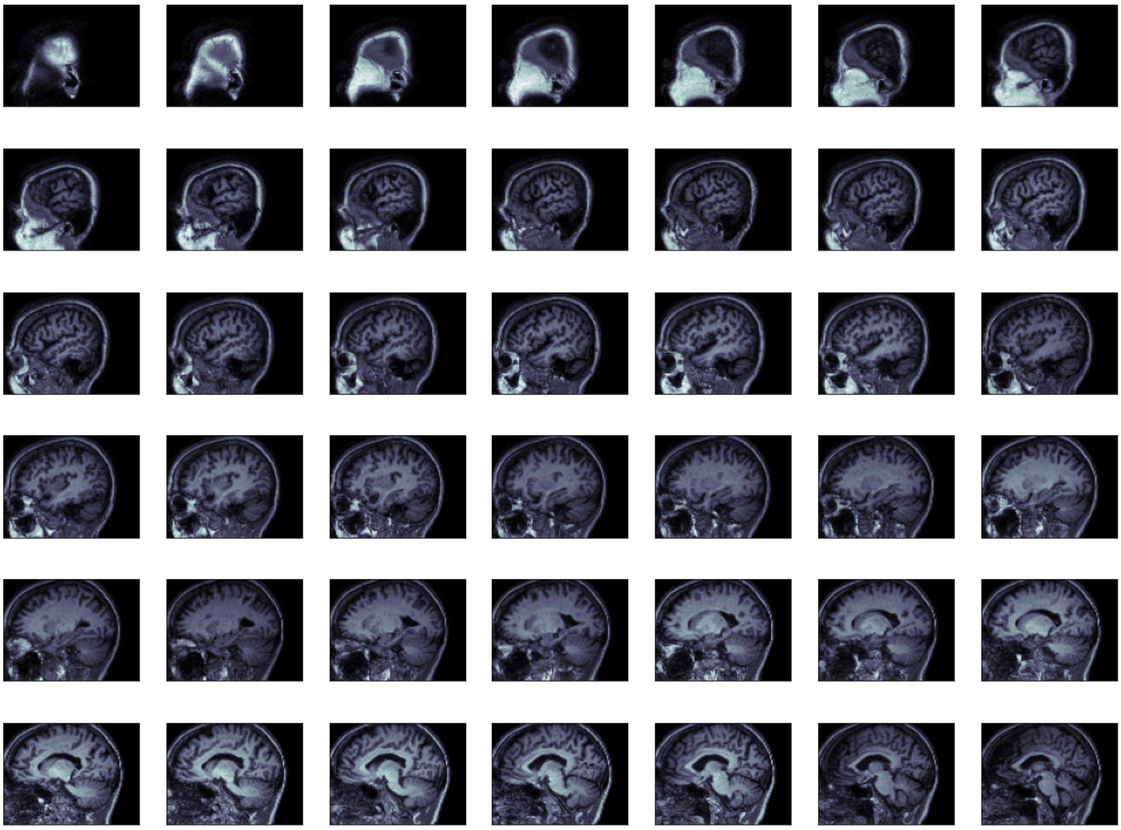


Figure 2. Processed volumetric data from a subject's MRI (extract of the processed slices).

4.3. Model Construction

We selected a convolutional DNN of the DenseNet Bottleneck-Compressed architecture because this kind of neural network structure has an exemplary performance with fewer training parameters [29], hence using fewer resources. We based our implementation on Hara et al. [30]. They based their implementation on the two-dimensional programming available in PyTorch. However, their implementation is not general purpose, but specific. It implements video and incorporates sample_size and sample_duration as variables related to video sample size and duration. We generalized their implementation, and now, it operates with all types of cuboids. Additionally, we added a channel parameter because the implementation only considered three channels (usually red, green, and blue colors), yet the MRI is monochromatic.

With this new programming, we set the training process of the neural network with the next parameters.

Training Of the data retrieved from ADNI, we used 75% as the training dataset. The training dataset was taken randomly from the full data set.

Batch size We selected a batch size of 5 MRI images for training based on the experimental results of [20].

Testing The remaining 25% of the data was the testing dataset.

Channels The images were monochromatic; therefore, we set channels as 1 in the constructor of the neural network.

Classes At first, we set the number of classes to 6 in the neural network. Nevertheless, we chose to drop the SMC class because we believe it is a noisy subjective class. We finally set the number of classes to 5.

Dropout We used a dropout rate of 0.7 based on the investigation of [20] because it reduces overfitting.

Loss For the loss function, we employed cross-entropy. It is useful in classification problems with more than two classes.

Optimizer We used Stochastic Gradient Descent (SGD) as the optimizer. SGD is used in the event of unbalanced data like in our dataset.

Learning We employed a learning rate parameter of 0.1 in the SGD optimizer. At Epoch 80, we dropped the learning rate by 0.1 (we started at Epoch 0). That drop diminished the learning rate to a value of 0.01.

Momentum We used a standard momentum of 0.9 because an SGD optimizer with momentum typically attains flatter local minima.

Epochs We established the maximum number of epochs as 80 (in a run) because we used the Google Colaboratory platform. Due to platform's limitations, it was impossible to exceed 90 and reach the desired 110 epochs because it disconnected us. As a result, we saved the model at Epoch 80. Because of Google Colaboratory's constraints, explained next, we needed to wait 12 h before continuing the training. Then, we restarted training from the 80th epoch until Epoch 110.

According to other authors, the free-of-charge resources of Google Colaboratory "are far from enough to solve demanding real-world problems and are not scalable" [31]. However, we used Google Colaboratory to access the Graphics Processing Unit (GPU) computing. With the mentioned parameters, we pushed the Google Colaboratory platform's boundaries to deliver a state-of-the-art DNN. This choice has constraints and consequences. As described in [31], the free GPU backend can be used for a maximum of 12 h. However, we believe that we had less than the allocated 12 h on some occasions, with approximately 10 h of use. The subsequent disconnection from Google Colaboratory occurred, dropping the virtual machine with its corresponding GPU. When the user immediately reconnects, the platform provides a new machine. However, it grants only 3 h of GPU backend. After the 3 h, it is impossible to use a GPU backend for a defined amount of time (12 h). To overcome the 3 h restriction, the user has to wait 12 h after the first 12 h run, as we described in the parameters. Such constraints limit the process of training and testing to a maximum of 12 h, which depending on the task, might be unfeasible.

Google Colaboratory's constraints have other consequences. For example, it is common to validate or test neural networks throughout training, thus analyzing the neural network's accuracy and loss across all epochs. Nonetheless, to decrease the computation time, validation or testing of the DNN was only executed at the end of the training. We decided this since a validation cycle of 25% of the data needed around 2 or 3 min, requiring 1 h or more every 30 epochs. We accepted this trade-off. It is possible to save intermediate neural network states and evaluate those states after completing the training. However, this decision also signifies that methods like early stopping can not be applied. Finally, Google Colaboratory also has disk size limitations that we overcame with our data transformation.

Based on the specified configuration parameters and considering all the constraints associated with Google Colaboratory, we were able to achieve the results discussed in the following section.

5. Results

With the mentioned limitations, we progressively achieved our results by increasing the epochs. We first trained with six classes to 50 epochs. Figure 3 shows that, after training to 50 epochs, the SMC class is not properly discriminated. The column of the predicted SMC class is filled with zeros. Remarkably, the rest of the classes have an adequate level of correct classification. We decided to remove the SMC class from the complete dataset. The deletion of this class decreased the complete dataset from 5556 MRIs to 5370 MRIs.

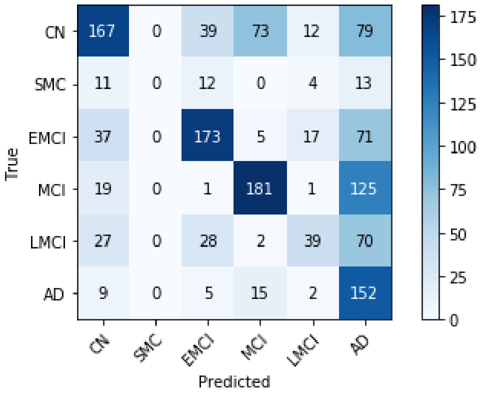
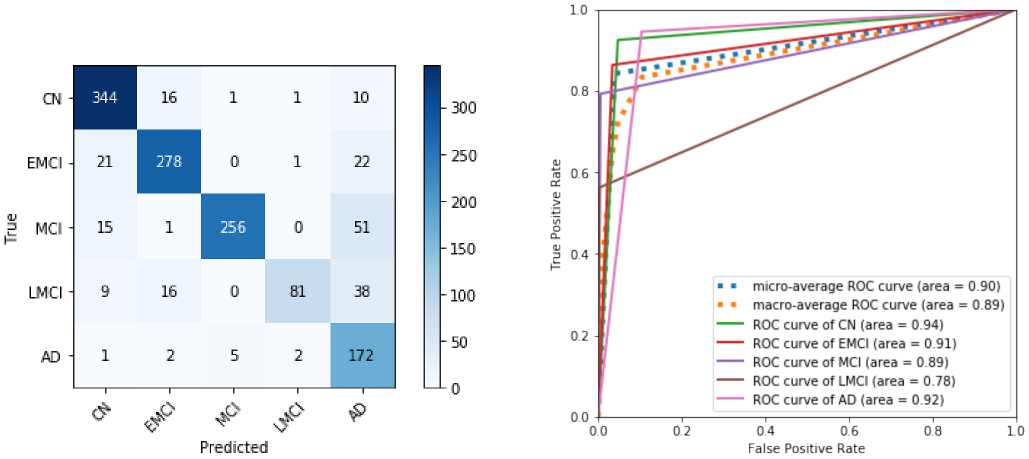


Figure 3. Confusion matrix with the SMC class at 50 epochs.

After removing the SMC class and training to 80 epochs, the neural network for the remaining classes showed good classification metrics. These results are shown in Figure 4a,b. The confusion matrix (Figure 4a) shows how most values are in the diagonal, with fewer incorrect predictions, especially compared to the confusion matrix that included the SMC class (Figure 3). The classifier's predictive power can be seen in Figure 4b, showing its potential for each class and all classes together.

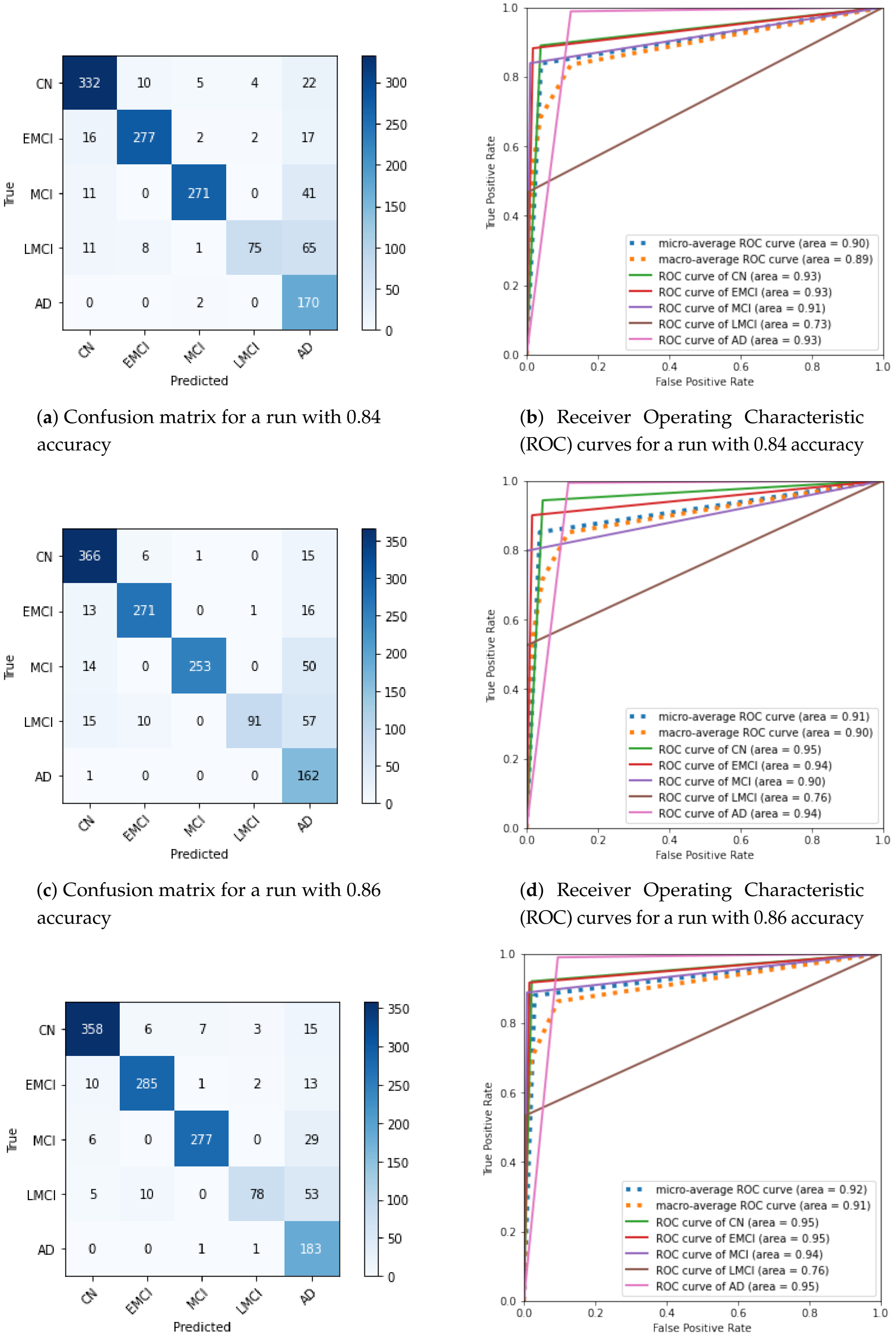


(**a**) Confusion matrix (**b**) Receiver operating characteristic curves

Figure 4. Evaluation plots of DenseNet-121 at 80 epochs.

The model was improved further, although it might be considered, based on Figure 4a, that our classifier already had a good performance. To further improve our model, we restarted the training from Epoch 80 to 110 epochs. The last predictive model had the metrics in Table 2 (page 11).

To produce evaluation metrics, we divided the entire dataset into four parts: 1342, 1342, 1343, and 1343 images. We then undertook a test run with each of these partitions. We did this four times, which resulted in 16 runs in total. We report the mean, minimum, and maximum values for specificity, sensitivity, f1-score, and support. For support, we also report the standard deviation because the classes were not balanced. Additionally, we show sample confusion matrices and plot the area under the Receiver Operating Characteristic (ROC) curves for the worse, median, and best accuracies. This is shown in Figure 5. The reported accuracies are: 0.84, 0.86, and 0.88 (respectively). We discuss the predictive performance and metrics in the next section.



(**a**) Confusion matrix for a run with 0.84 accuracy

(**b**) Receiver Operating Characteristic (ROC) curves for a run with 0.84 accuracy

(**c**) Confusion matrix for a run with 0.86 accuracy

(**d**) Receiver Operating Characteristic (ROC) curves for a run with 0.86 accuracy

(**e**) Confusion matrix for a run with 0.88 accuracy

(**f**) Receiver Operating Characteristic (ROC) curves for a run with 0.88 accuracy

Figure 5. Evaluation plots of DenseNet-121 at 110 epochs.

Table 2. Evaluation metrics of the obtained DNN at 110 epochs.

	Specificity (Precision)		**Sensitivity (Recall)**		**F1-Score**		**Support**		
Cognitive Normal	$\overline{0.92}$	$[0.89, 0.94]$	$\overline{0.93}$	$[0.89, 0.95]$	$\overline{0.92}$	$[0.89, 0.94]$	$\overline{380}$	$[350, 411]$	$\sigma^2 = 17.13$
Early Mild Cognitive Impairment	$\overline{0.94}$	$[0.92, 0.95]$	$\overline{0.90}$	$[0.87, 0.92]$	$\overline{0.92}$	$[0.90, 0.93]$	$\overline{305.5}$	$[290, 322]$	$\sigma^2 = 10.27$
Mild Cognitive Impairment	$\overline{0.98}$	$[0.96, 1.00]$	$\overline{0.85}$	$[0.80, 0.89]$	$\overline{0.91}$	$[0.89, 0.93]$	$\overline{318.5}$	$[301, 343]$	$\sigma^2 = 10.27$
Late Mild Cognitive Impairment	$\overline{0.94}$	$[0.90, 0.99]$	$\overline{0.51}$	$[0.42, 0.55]$	$\overline{0.66}$	$[0.58, 0.69]$	$\overline{159}$	$[145, 182]$	$\sigma^2 = 10.31$
Alzheimer's Disease	$\overline{0.59}$	$[0.52, 0.63]$	$\overline{0.99}$	$[0.98, 0.99]$	$\overline{0.73}$	$[0.68, 0.77]$	$\overline{179.5}$	$[163, 201]$	$\sigma^2 = 9.87$
Macro average	$\overline{0.87}$	$[0.85, 0.88]$	$\overline{0.83}$	$[0.81, 0.85]$	$\overline{0.83}$	$[0.80, 0.85]$	$\overline{1342.5}$	$[1342, 1343]$	$\sigma^2 = 0.5$
Weighted average	$\overline{0.89}$	$[0.88, 0.90]$	$\overline{0.86}$	$[0.84, 0.88]$	$\overline{0.86}$	$[0.84, 0.88]$	$\overline{1342.5}$	$[1342, 1343]$	$\sigma^2 = 0.5$
Accuracy					$\overline{0.86}$	$[0.84, 0.88]$			
Micro specificity (precision)					$\overline{0.86}$	$[0.84, 0.88]$			
Micro sensitivity (recall)					$\overline{0.86}$	$[0.84, 0.88]$			

6. Discussion

Our first finding was that significant memory concern was noisy for training. We inferred this from the results in Figure 3. This property can be because the class is biased and is likely formed of at least two classes: people who will not develop the disease and people who will. Besides, people who will develop the disease may possess distinct progression levels, being, in turn, a class formed of different classes. The size of the SMC class may be another reason why it is problematic. It is the smallest cohort by far. These may be the causes that make it difficult to classify. We decided to remove this class.

We obtained the confusion matrix in Figure 4a after dropping the SMC class and training to 80 epochs. In the matrix, we may see that the incorrect predictions are principally pessimistic. That is, there are more errors above the diagonal than under it. This arrangement of errors above the diagonal means that the classifier makes errors that predict the upper disease stages. This kind of error is clearly in favor of patients because a false positive is better than a false negative in diagnosing diseases. Figure 4b also shows that the area under each curve tends to 1.0; which demonstrates the classifier's diagnostic ability. Notwithstanding the faults of our classifier, it is a valuable classifier. Next, we still improve this model.

As we introduced in Section 5, we trained our final model up to 110 epochs. We extended what we previously undertook in [8]. We evaluated the training to 110 epochs with more rigor. We carried out 16 test runs, as described in Section 5.

We also created new confusion matrices and ROC curve plots (Figure 5). When analyzing the newly obtained confusion matrices, we noticed a possible pattern (Figure 5a,c,e). In the AD prediction column, we see an ascending number of false AD as there is more disease progression. There is possibly a feature that is recognized as AD as the disease progresses. Healthy people possess this feature. These false AD predictions require further investigation. However, without investigation, we believe that the solution would be the brain segmentation we propose in Section 6.1.

As mentioned in Section 3, accuracy can be a misleading metric. The confusion matrix in Figure 5c has better accuracy than the confusion matrix in Figure 5a; however, it has more elements under the diagonal. It is preferable to have more elements above the diagonal in the confusion matrix in health care and clinical diagnostics. The opposite is optimistic, and in diagnostic terms, that can be a risk. A false negative can be counterproductive, as a healthy person is misdiagnosed, with deleterious consequences. It is common in health care clinical diagnostics to prefer a pessimistic approach to an optimistic one. We commonly opt for false positives over false negatives in that context.

As the ability to differentiate between most classes improved, we noticed a decrease in the late mild cognitive impairment class's performance metrics. We inferred that we advanced towards a local minimum solution that increased the other classes' predictive power, but departed from the LMCI class's accurate prediction. We believe that the need for class balance in the dataset is a possible cause of this observation. The LMCI is the class with the lowest number of instances (after the removal of SMC). The pitfall of class imbalance and the number of instances could potentially be addressed with data augmentation, as proposed in [32]. However, data augmentation would introduce a performance constraint and reduce the highest amount of epochs we can use during training. Data augmentation is a trade-off we decided not to make. Additional inspection of the LMCI prediction showed that many of the misclassified instances were labeled as AD. We can consider this a pessimistic behavior, and therefore, it was tolerated. Consequently, we accepted the final metrics and the compromise of not balancing the data through data augmentation because of the limitations imposed by Google Colaboratory.

We present additional classification metrics of this final model in Table 2. In that table, the lowest values are for AD specificity (precision) and LMCI sensitivity (recall). We could also include the sensitivity (recall) of MCI in the low numbers. The MCI 85% sensitivity (recall) will miss 25% of MCI patients; however, they will be classified pessimistically in an advanced disease stage. The AD class's low specificity is acceptable because it reaches almost 100% sensitivity (recall). Therefore, our approach is a useful diagnostic method for AD, although not suitable for screening. Screening methods must possess high specificity metrics, whereas diagnostic approaches should have high sensitivity scores. An explanation of why the specificity metric is low is the number of multiple instances misclassified as AD. Nevertheless, false positives are prophylactic and reduce future risks in a diagnostic task.

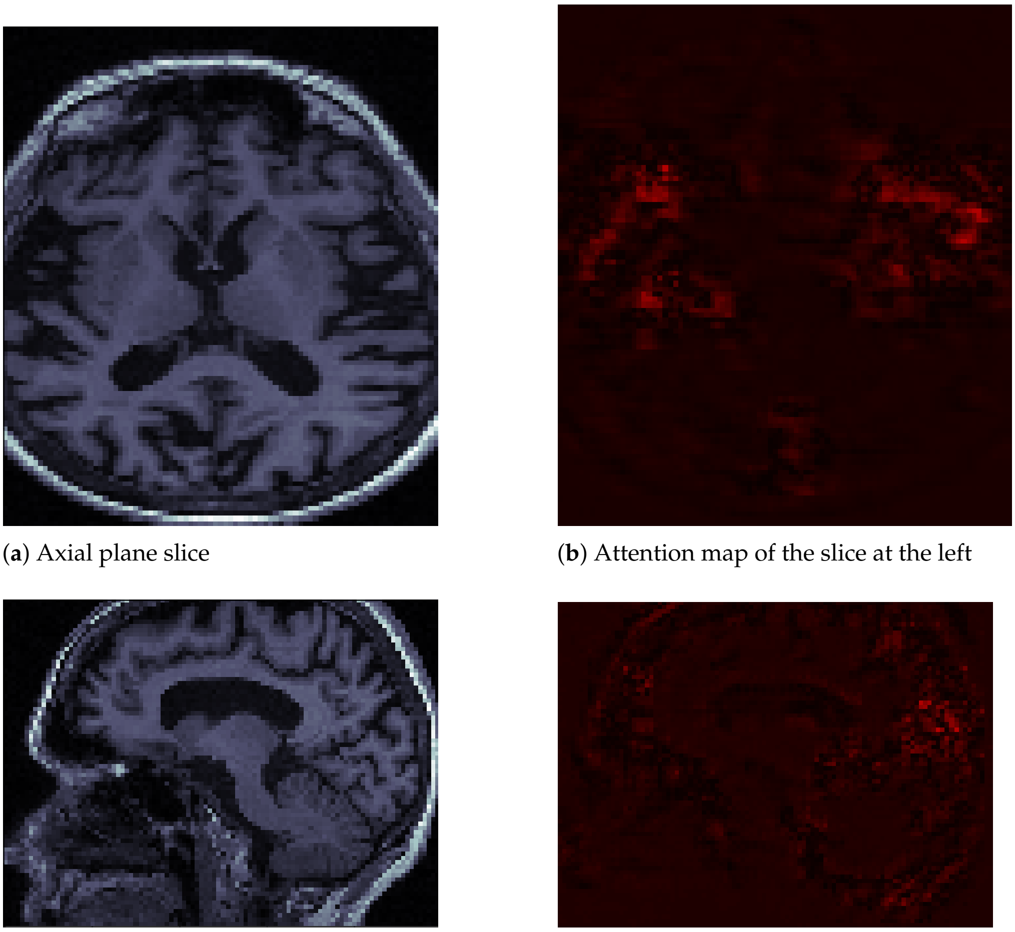
Our results also show that the AUROC and accuracy values do not contradict each other. That consistency supports that our obtained results are consequent with each other.

We constructed our models considering compromises due to the use of freely available tools. We developed methodologies that attempted to counter the limitations and achieve the best possible classification for diagnosis (high sensitivity). Our model's predictive ability has values of $\overline{0.86}$ mean accuracy, $\overline{0.86}$ mean specificity (micro), and $\overline{0.86}$ mean sensitivity (micro), which overall is considered acceptable. These metrics emphasize that our classifier can satisfactorily achieve the corresponding task of discriminating between the classes, even with our limitations. We decided to report final micro-average numbers instead of the macro-average. In a multiclass classification structure, the micro-average is preferred when there is a class imbalance. Still, the macro-average and the weighted average are higher. Finally, we only report the variance for support because, for the other columns, the variance is close to zero.

In the next section, we provide a study of what the neural network perceives.

6.1. Attention Maps

Attention maps are visual tools that explain deep convolutional neural networks [33]. We used the M3d-Cam tool [34] with the Guided Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm. We generated new images through M3d-CAM, and with these images, we could interpret what the neural network was emphasizing to make a decision. For example, in Figure 6a, we can see an image in the axial plane of a person with Alzheimer's disease. We extracted only one slice of the cuboid. To its right, in Figure 6b, we can see that the neural network focuses its attention on brain atrophy. We extracted the same slice number from the attention map cuboid. In Figure 6c,d, we see the same patient and the same features in the sagittal plane. We also only extracted one slice of that plane.



(**a**) Axial plane slice

(**b**) Attention map of the slice at the left

(**c**) Slice of an input image in the sagittal plane

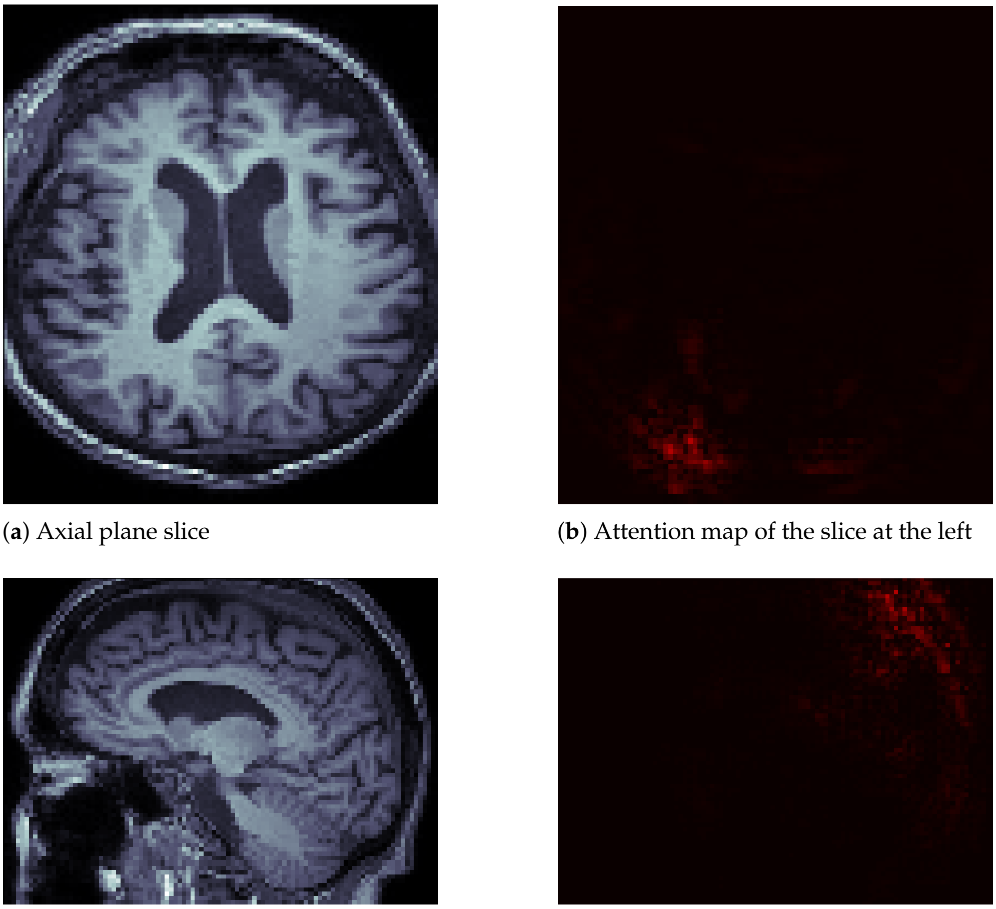
(**d**) Attention map of the slice at the left

Figure 6. Input image slices and their attention maps from an MRI of a patient with AD.

In the images in Figure 6, we can see how brain atrophy areas appear more prominent. We noticed this behavior in other images for which we produced attention maps. The behavior of the network was as expected in terms of identifying the regions of importance. However, we also noticed that sometimes the neural network gave importance to bony areas of the skull or even to areas outside the skull. In other words, the attention maps seemed to have a potential function as debugging tools. Besides, they allowed us to propose solutions. We believe that we must improve the preprocessing of images by removing the skull using a mask, as the skull is a potential source of noise. This background noise could explain some of the classifier's inaccurate results, as observed in the confusion matrices. The neural network may be learning uninformative features or patterns as a result of excess inputs. We could improve the results and reduce the inaccurate results by providing the CNN with only brain images, thus removing potential artifacts associated with the skull.

Additionally, we include the attention map for a non-affected individual in Figure 7 on page 14. The difference from the image in Figure 6 is that this image seems to generate less attention from the neural network. We observed the same behavior in other images of healthy people.

The M3d-CAM tool allowed us to better study the resulting neural network and generate new hypotheses to improve our current work. Furthermore, by incorporating the attention maps, we added a layer to our work that widens what we can achieve with our current and future classification systems. As an example, we now consider the use of brain segmentation. It is essential to include more visualization tools for neural networks that would allow debugging and understanding these architectures in depth.



(**a**) Axial plane slice

(**b**) Attention map of the slice at the left

(**c**) Slice of an input image in the sagittal plane

(**d**) Attention map of the slice at the left

Figure 7. Image slices and their attention maps from an MRI of a healthy person.

7. Conclusions

We explored a low-cost way to create a deep artificial convolutional neural network architecture for AD detection that proved its effective performance in this work. Our classification system has performance metrics of 0.86 mean accuracy, 0.86 mean sensitivity (micro-average), 0.86 mean specificity (micro-average), and 0.91 area under the receiver operating characteristic curve (micro-average).

The constructed classifier is a valuable and potentially viable option for developing countries' diagnostic systems if we add a user interface and interpretation to the model. The model presents high individual sensitivities (recall), a required staple for computer-assisted diagnostic systems. We can use such a tool in remote medicine scenarios and deployment for AD early detection. We may base a possible CAD on Chester's open implementation [35], a computerized disease prediction system for chest X-rays presented through the web. With the advent of recent tools such as TensorFlow.js and ONNX, we can transform models trained using PyTorch to work in a browser, and we can use WebGL for display [35]. This interface would permit prediction and also add explanation or interpretation through attention maps or heat maps. One hospital in Costa Rica has shown interest in such a tool.

By implementing attention maps, we discerned a potential focus for future improvement. We identified noisy features that could be the cause of inaccuracies. We could remove the noise by employing image processing. Using a mask, we would cut non-brain pixels. We do not attribute errors to the freely available tools used.

Additionally, we want to improve our model without using supervised learning. We will use reinforcement learning. We also will not use gradient methods for training, that is to say gradient descent and backpropagation. We will combine reinforcement learning with evolutionary computing.

We also want to perform a series of experiments. One possible experiment we could carry is the comparison of different architectures as, for instance, DenseNet-169, DenseNet-201, ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152, and ResNeXt-101. We could consider every architecture as a treatment for the data and perform an analysis of the variances for comparison. In that same line, we also want to compare against human professionals in the health sector. For instance, works such as [6] compared against humans.

Another possibility of experimentation consists of mixing different data sources. For example, we could use the entire ADNI tagged data set for training only, and we could use other data sources for validation. Among the possible data sources for validation are our local health centers or Open Access Series of Imaging Studies (OASIS) [36].

Finally, regarding the transparency and reproducibility of academic work, we contribute the source code of our DNN at [37]. The produced neural network is available for download at [38].

Author Contributions: Conceptualization, B.S.-R.; methodology, B.S.-R.; software, B.S.-R.; validation, B.S.-R. and R.V.-F.; formal analysis, B.S.-R.; investigation, B.S.-R.; resources, B.S.-R.; data curation, B.S.-R.; writing—original draft preparation, B.S.-R.; writing—review and editing, B.S.-R. and R.V.-F.; visualization, B.S.-R.; supervision, R.V.-F.; project administration, B.S.-R.; funding acquisition, B.S.-R. and R.V.-F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Centro de Investigaciones en Tecnologías de la Información y Comunicación (CITIC) and Escuela de Ciencias de la Computación e Informática (ECCI), Universidad de Costa Rica.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used is available at http://adni.loni.usc.edu.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
AUROC	Area Under the Receiver Operating Characteristic curve
CAD	Computer-Aided Diagnosis
CN	Cognitively Normal
CNN	Convolutional Neural Networks
DICOM	Digital Imaging and Communication On Medicine
EMCI	Early Mild Cognitive Impairment
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
LMCI	Late Mild Cognitive Impairment
MCI	Mild Cognitive Impairment
MPRAGE	Magnetization-Prepared Rapid Acquisition with Gradient Echo
PET	Positron Emission Tomography
ROC	Receiver Operating Characteristics
SGD	Stochastic Gradient Descent
SJR	SCImago Journal Rank
SMC	Significant Memory Concern

References

1. National Institute on Aging. Alzheimer's Disease Fact Sheet. Available online: https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet (accessed on 13 May 2018).
2. MedlinePlus. Alzheimer's Disease. Available online: https://medlineplus.gov/alzheimersdisease.html (accessed on 13 May 2018).
3. Castellino, R.A. Computer aided detection (CAD): An overview. *Cancer Imaging* **2005**, *5*, 17–19. [CrossRef] [PubMed]
4. Lee, C.S.; Nagy, P.G.; Weaver, S.J.; Newman-Toker, D.E. Cognitive and System Factors Contributing to Diagnostic Errors in Radiology. *Am. J. Roentgenol.* **2013**, *201*, 611–617. [CrossRef] [PubMed]
5. Graber, M.; Franklin, N. Diagnostic Error in Internal Medicine. *Arch. Intern. Med.* **2005**, *165*. [CrossRef] [PubMed]
6. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv* **2017**, arXiv:1711.05225.
7. Khan, S.; Rahmani, H.; Shah, S.; Bennamoun, M. *A Guide to Convolutional Neural Networks for Computer Vision*; Number 1 in Synthesis Lectures on Computer Vision; Morgan & Claypool: San Rafael, CA, USA, 2018. [CrossRef]
8. Solano-Rojas, B.; Villalón-Fonseca, R.; Marín-Raventós, G. *Alzheimer's Disease Early Detection Using a Low Cost Three-Dimensional DenseNet-121 Architecture. The Impact of Digital Technologies on Public Health in Developed and Developing Countries*; Jmaiel, M., Mokhtari, M., Abdulrazak, B., Aloulou, H., Kallel, S., Eds.; Springer: Cham, Switzerlands, 2020; pp. 3–15.
9. Alzheimer's Disease Neuroimaging Initiative. Study Design. 2017. Available online: http://adni.loni.usc.edu/study-design/ (accessed on 13 May 2018).
10. National Institute of Biomedical Imaging and Bioengineering. Magnetic Resonance Imaging (MRI). Available online: https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri (accessed on 13 May 2018).
11. National Institute of Biomedical Imaging and Bioengineering. Nuclear Medicine. Available online: https://www.nibib.nih.gov/science-education/science-topics/nuclear-medicine#pid-1001 (accessed on 13 May 2018).
12. Korolev, S.; Safiullin, A.; Belyaev, M.; Dodonova, Y. Residual and plain convolutional neural networks for 3D brain MRI classification. In Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, Australia, 18–21 April 2017; pp. 835–838. [CrossRef]
13. Farooq, A.; Anwar, S.; Awais, M.; Alnowami, M. Artificial intelligence based smart diagnosis of alzheimer's disease and mild cognitive impairment. In Proceedings of the 2017 International Smart Cities Conference (ISC2), Wuxi, China, 14–17 September 2017; pp. 1–4. [CrossRef]
14. Farooq, A.; Anwar, S.; Awais, M.; Rehman, S. A deep CNN based multi-class classification of Alzheimer's disease using MRI. In Proceedings of the 2017 IEEE International Conference on Imaging Systems and Techniques (IST), Beijing, China, 18–20 October 2017; pp. 1–6. [CrossRef]
15. Yagis, E.; De Herrera, A.G.S.; Citi, L. Generalization Performance of Deep Learning Models in Neurodegenerative Disease Classification. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; pp. 1692–1698. [CrossRef]
16. Jin, D.; Xu, J.; Zhao, K.; Hu, F.; Yang, Z.; Liu, B.; Jiang, T.; Liu, Y. Attention-based 3D Convolutional Network for Alzheimer's Disease Diagnosis and Biomarkers Exploration. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 1047–1051. [CrossRef]
17. Cui, R.; Liu, M. Hippocampus Analysis by Combination of 3-D DenseNet and Shapes for Alzheimer's Disease Diagnosis. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 2099–2107. [CrossRef] [PubMed]
18. He, G.; Ping, A.; Wang, X.; Zhu, Y. Alzheimer's Disease Diagnosis Model Based on Three-Dimensional Full Convolutional DenseNet. In Proceedings of the 2019 10th International Conference on Information Technology in Medicine and Education (ITME), Qingdao, China, 23–25 August 2019; pp. 13–17. [CrossRef]
19. Jabason, E.; Ahmad, M.O.; Swamy, M.N.S. Classification of Alzheimer's Disease from MRI Data Using an Ensemble of Hybrid Deep Convolutional Neural Networks. In Proceedings of the 2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS), Dallas, TX, USA, 4–7 August 2019; pp. 481–484. [CrossRef]
20. Bäckström, K.; Nazari, M.; Gu, I.Y.; Jakola, A.S. An efficient 3D deep convolutional network for Alzheimer's disease diagnosis using MR images. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 149–153. [CrossRef]
21. Khagi, B.; Kwon, G.R. 3D CNN Design for the Classification of Alzheimer's Disease Using Brain MRI and PET. *IEEE Access* **2020**, *8*, 217830–217847. [CrossRef]
22. Niu, J.; Tang, X. 3D Residual Dense Convolutional Network for Diagnosis of Alzheimer's Disease and Mild Cognitive Impairment. In Proceedings of the 2020 IEEE International Conference on Mechatronics and Automation (ICMA), Beijing, China, 13–16 October 2020; pp. 1581–1586. [CrossRef]
23. Sahumbaiev, I.; Popov, A.; Ramírez, J.; Górriz, J.M.; Ortiz, A. 3D-CNN HadNet classification of MRI for Alzheimer's Disease diagnosis. In Proceedings of the 2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC), Sydney, Australia, 10–17 November 2018; pp. 1–4. [CrossRef]
24. Yagis, E.; Citi, L.; Diciotti, S.; Marzi, C.; Workalemahu Atnafu, S.; De Herrera, A.G.S. 3D Convolutional Neural Networks for Diagnosis of Alzheimer's Disease via Structural MRI. In Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 28–30 July 2020; pp. 65–70. [CrossRef]

25. Choi, J.Y.; Lee, B. Combining of Multiple Deep Networks via Ensemble Generalization Loss, Based on MRI Images, for Alzheimer's Disease Classification. *IEEE Signal Process. Lett.* **2020**, *27*, pp. 206–210. [CrossRef]
26. Jabason, E.; Ahmad, M.O.; Swamy, M.N.S. Hybrid Feature Fusion Using RNN and Pre-trained CNN for Classification of Alzheimer's Disease (Poster). In Proceedings of the 2019 22th International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2–5 July 2019; pp. 1–4.
27. Salehi, A.W.; Baglat, P.; Sharma, B.B.; Gupta, G.; Upadhya, A. A CNN Model: Earlier Diagnosis and Classification of Alzheimer Disease using MRI. In Proceedings of the 2020 International Conference on Smart Electronics and Communication (ICOSEC), Tamil Nadu, India, 7–9 October 2020; pp. 156–161. [CrossRef]
28. Vinutha, N.; Pattar, S.; Kumar, C.; Agarwal, A.; Shenoy, P.D.; Venugopal, K. A Convolution Neural Network based Classifier for Diagnosis of Alzheimer's Disease. In Proceedings of the 2018 Fourteenth International Conference on Information Processing (ICINPRO), Bangalore, India, 21–23 December 2018; pp. 1–6. [CrossRef]
29. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2016**, arXiv:cs.CV/1608.06993.
30. Hara, K.; Kataoka, H.; Satoh, Y. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6546–6555.
31. Carneiro, T.; Medeiros Da NóBrega, R.V.; Nepomuceno, T.; Bian, G.; De Albuquerque, V.H.C.; Filho, P.P.R. Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access* **2018**, *6*, 61677–61685. [CrossRef]
32. Cheng, D.; Liu, M. CNNs based multi-modality classification for AD diagnosis. In Proceedings of the 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, China, 14–16 October 2017; pp. 1–5. [CrossRef]
33. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
34. Gotkowski, K.; Gonzalez, C.; Bucher, A.; Mukhopadhyay, A. M3d-CAM: A PyTorch library to generate 3D data attention maps for medical deep learning. *arXiv* **2020**, arXiv:2007.00453.
35. Cohen, J.P.; Bertin, P.; Frappier, V. Chester: A Web Delivered Locally Computed Chest X-Ray Disease Prediction System. *arXiv* **2019**, arXiv:cs.CV/1901.11210.
36. LaMontagne, P.J.; Benzinger, T.L.; Morris, J.C.; Keefe, S.; Hornbeck, R.; Xiong, C.; Grant, E.; Hassenstab, J.; Moulder, K.; Vlassenko, A.; Raichle, M.E.; Cruchaga, C.; Marcus, D. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. *medRxiv* **2019**.10.1101/2019.12.13.19014902. [CrossRef]
37. Solano-Rojas, B. Use of 3D ResNets for Alzheimer Disease Early Detection. 2020. Available online: https://github.com/bsolano/Alzheimer-ResNets (accessed on 17 January 2021).
38. Solano-Rojas, B. ICOST 2020 - Alzheimer Resnet. 2020. Available online: https://github.com/bsolano/Alzheimer-ResNets/releases/tag/v0.1-icost (accessed on 17 January 2021).