Personalized Medicine: Comparison of Techniques for the Automatic Diagnosis of Alzheimer's Disease

Óscar Darias Plasencia

Universidad Internacional de la Rioja, Logroño (España)

Date July 23rd, 2019

Abstract

In recent years, there has been a growing interest in automating the diagnosis of different diseases with Artificial Intelligence, including Alzheimer. This trend is trying to reduce human error and make diagnosis much more accurate, through developments in machine learning and deep learning by finding hidden patterns in medical images. In this context, this paper presents a state-of-the-art of different models used to support Alzheimer's diagnosis, emphasizing in the most suitable ones according to literature. Additionally, these models were implemented in an experimental scenario by means of simple tools, in order to learn about what extent their development is feasible by professionals without deep knowledge of medicine. Results show that convolutional neural networks allow the construction of quite powerful models, but their deployment in real environments is hardly feasible today. Finally, recommendations of all the followed process are emitted to guide professionals into the use of these models.

Keywords

Alzheimer, Deep Learning, Diagnosis, Medical Image Analysis

I. Introduction

In the last two decades, theoretical and practical advances of great relevance in Artificial Intelligence (AI) have been materializing, affecting a wide variety of fields, such as medicine, tourism, education, entertainment, among others [1]–[4]. Specifically, in the field of medicine, which is the general focus of this research, AI has proven its potential in various aspects such as: assistance with the formulation of a diagnosis, lesion detection, or cells and organs segmentation [5]–[7].

This growing interest for Artificial Intelligence among medical environments is due to the possibilities it offers for the automatization of several processes and the reduction of human error. Obviously, these two advantages are the same as in many other application areas. However, there is an extra incentive for AI research in medicine: the prospect of saving lives. Transferring the enormous advances that have been made in the last decade, especially in deep learning, could lead to significant improvements in the diagnosis or treatment of diseases, as well as in many other applications.

Within the paradigm of the application of AI to medicine, this research is focused on Alzheimer's diagnosis. This neurodegenerative disease is the most common type of dementia. The increase in life expectancy in recent decades has caused deaths from this disease to increase by 66% [8]. It is triggered by the malfunction of neurons, and patients end up being uncapable of performing the most basic tasks, such as walking or swallowing [8]. The problem is aggravated by the fact that the specific conditions that trigger the disease are largely unknown, resulting in major difficulties when preventing and treating its victims [8], [9].

The rise of deep learning in the past two decades has prompted research into solutions for automatic Alzheimer's diagnosis based on neuroimaging. A wide variety of different techniques have been used, but a clear turn towards the use of convolutional neural networks has been observed in the last decade, not only with Alzheimer but also with the diagnosis of several other diseases, as well as segmentation, computer-aided detection, and other medical applications [10]–[13]. Therefore, the variety of proposed options is enormous, and each one has its own particularities. For example, some of them use magnetic resonance imaging (MRI), while others opt for positron emission tomography images (PET).

There are also a lot of different machine learning models that can be used. In this research, the main ones are studied and implemented in an experimental environment, to see which one of them is the best option. Results show that fine-tuning a pretrained bidimensional convolutional neural network is the simplest model with which it is easier to obtain good performance.

The implementation of one of these alternatives usually requires of deep medical knowledge. Preprocessing the MRI or PET images can be very complicated, as well as extracting useful features from them. However, data scientists rarely have that knowledge, and its complexity can make it very difficult to get into this field. For that reason, this research aims to clarify whether or not the implementation of these models is feasible for regular data scientists. It also seeks to make the appropriate recommendations for any data scientist that is trying to initiate in this field or implement one of these solutions. For all of that, simple tools available, like *SimpleElastix* [14] and FSL BET [15], where used for preprocessing the images. For model design, Keras was the selected option [16]. Results show that, although decent results could be achieved, full implementation of this models in medical environments is hardly recommended.

II. Objectives and Methodology

The main objective of this research is to generate recommendations on the most appropriate machine learning models to support decision making in an Alzheimer's diagnosis system. This implies a detailed bibliographic review and the implementation of state-of-the-art models in an experimental environment, using a neuroimaging database. More specifically, the following three objectives guided this research:

1. Identification of the models that are providing the best

results in the bibliography. This implies a detailed examination of the whole field.

2. Implementation of the models or algorithms selected in the previous objective. The ADNI neuroimaging database [17] was used for training and testing.
3. Analyze and compare the results obtained by the different models and contextualize them with the state of the art. Emit the right recommendations based on these results.

The methodology was very straightforward, dividing the work in two main phases. The first one was an exploratory phase, based on studying the current state of the art. The following activities were performed on this phase:

- Systematic literature review, for better understanding the current state of the art in neuroimaging analysis and Alzheimer's automatic diagnosis.
- Identification of the models that have provided the best results, as well as the most important preprocessing techniques for neuroimaging.
- Selection of the best two models, algorithms or techniques.
- Study of several available tools for preprocessing the neuroimaging data.

The second phase was the descriptive phase, based on the implementation and comparison of the selected algorithms, using ADNI neuroimaging data preprocessed with the simple tools previously examined. The following activities where performed on this phase:

- Magnetic Resonance Imaging (MRI) data was prepared for its use for training the selected models.
- Prototyping and implementation of the selected models, using the preprocessed MRI data.
- Evaluation of the obtained results, contextualizing them with the state of the art.
- Emit the appropriate recommendations for data scientists without deep knowledge in medicine.

III. State of the Art

Medical image analysis is a key aspect of the diagnosis and treatment of several diseases. These types of images are a powerful source from where to extract useful biomarkers [18]. They are an important part of the Electronic Health Records (EHR) and are usually examined by dedicated professionals (radiologists). There are also a wide variety of image formats, being magnetic resonance imaging (MRI) and positron emission tomography (PET) the most common ones in Alzheimer's diagnosis. Over time, there has been increasing interest in the automatic processing of this type of images using Artificial Intelligence. In fact, AI is expected to play a very important role in the analysis of EHR in general, not just images.

In the last 15 years, the use of these algorithms and AI systems has increased considerably in medical applications. There are three main factors associated to this:

- Increased quantity and quality of data. In this regard, the field is getting close to Big Data.
- Attempts to reduce human error, since radiologists are limited by several factors like speed or experience, and can make mistakes [11], [19].
- The rise of AI itself and, more specifically, Deep Learning. If this type of systems had not shown such outstanding performance in recent decades [20], [21], adapting them for medical applications would not have been seen as feasible.

Although there have been multiple publications based on simple machine learning algorithms, like support vector machines (SVM) [19], and statistical methods like independent component analysis (ICA) [22], medical image analysis in the past five to ten years has been taken over by deep learning and convolutional neural networks (CNN) [11], [12].

These techniques have also been used in different applications within medical image analysis. They can be condensed in the following categories [10]–[13]:

- Computer-aided Detection (CADe). Identification of certain elements in the images, such as organs or cells. It is also useful for highlighting interesting areas for a physicist, such as lesions.
- Segmentation, which should not be confused with CADe. Isolation of entire image regions from the rest.
- Computer-aided Diagnosis (CADx), on which this work is focused. Diagnosis based on certain information, which could be explained, in simpler terms, as a classification task. In this case, medical images are used, hence the importance of convolutional neural networks. In the case of Alzheimer's disease, classification is usually made between three classes: normal cohort (NC), mild cognitive impairment (MCI) or Alzheimer's disease (AD).

In [13], an additional interesting category is mentioned; it could be referenced as deep feature learning. It is based on the design and implementation of systems capable of extracting useful features from the data. It allows to obtain higher level features that are invisible to the human eye and happen to be reusable in different scenarios. In Alzheimer's CADx, it is often used as a previous step for extracting useful features from the images or pre-training deep networks, by means of different algorithms like autoencoders [23]–[26]. However, this process is falling into disuse, requires extra developing stages and is no longer improving the overall results [27]. For those reasons, deep feature extraction was not considered during the descriptive phase.

In the end, the neural network represents the basis for the vast majority of the models used in the past five to ten years. These are usually trained in a supervised manner, but their non-supervised applications are also very important. In any case, images must be preprocessed so models can take full advantage of them.

Image Preprocessing

There are a number of medical imaging preprocessing techniques that are common to a wide variety of publications related to research on the automatic diagnosis of Alzheimer's disease. Originally, authors made use of manually defined features, which required very complex preprocessing techniques. However, the use of convolutional networks and autoencoders for automatic feature extraction slightly simplifies this process. In the end, two key procedures should be highlighted: the recording of medical images and the skull stripping.

Image registration consists on adapting a certain image to another reference image, which is called atlas, seeking that the same regions of both represent the same anatomical structures [28], [29]. This way, it would be simpler for a CNN or an autoencoder to identify a certain region of the images as relevant, since the same information would be represented in all of them. There is a great variety of registration algorithms [29], applicable not only to neuroimaging [25], [30]–[32] but also other fields of medicine, such as breast cancer [33].

Skull stripping is, as the name implies, the removal of information from the skull that appears on MRI images. The objective is to obtain a final image as clean as possible, containing only the information relevant to the task at hand. Obviously, in the case of Alzheimer's disease, none of the most relevant biomarkers are found in the skull. Therefore, certain publications use different techniques to extract the skull and other non-brain regions [30]–[32], or directly use an image

data set with the skull already extracted [34]. Fig. 1 shows how the skull occupies an important part of the images, especially in lower axial cuts. It is worth mentioning that, in the case of those studies that make use of PET images, skull-stripping is not mentioned, since this type of images do not present as much irrelevant information.

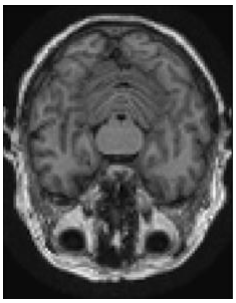

Fig. 1. Low axial cut of an MRI image without skull-stripping

Beyond these two procedures, there are other more generic techniques, such as the normalization of images. This can be broken down into two variants: intensity normalization and spatial normalization. The first of these is based on adapting the range of pixel values according to a certain criterion, such as reducing them to a certain range or subtracting the mean and dividing by the standard deviation, which is called whitening [35]. The second, on the other hand, consists of adapting the pixels (or voxels in 3D) so that they represent a certain space (or volume) [35]. For example, in [36] 3D images are adapted so each voxel represents $2mm^3$ of space. In this regard, image registration can be considered as a form of spatial normalization [35].

Alzheimer's CADx

In recent years, convolutional neural networks have started to take on an important role in Alzheimer's diagnosis [30], [34], [36]. That does not mean these models were not used in previous publications, but they used to be supported by other deep learning strategies, like the previously addressed Deep Feature Extraction. It is in the past three years that they have been used directly, building excellent performing models that, at the same time, are much simpler that previous attempts.

AlexNet [21] was a turning point in the history of Deep Learning. It proved how a CNN was able to get excellent results in image classification. In the following years, its ideas would be further developed, giving rise to other important architectures like VGGNet [37], Inception [38] and ResNet [39]. Although they were originally designed for working with ImageNet data [40], their excellent results have made them the default choices for multiple applications. Most Deep Learning frameworks offer these networks already trained with ImageNet, so developers can fine-tune them for their own purposes.

Medical image analysis has not been any different. Although it is true that both domains have wide differences that limit the performance that can be obtained through fine-tunning these networks, this technique has still proved to be the most promising one in practice. Not only with Alzheimer, but also with the diagnosis of other diseases, like diabetic retinopathy [41], breast cancer [42] and skin cancer [43].

- In [41], authors fine-tune an Inception V3 model.
- In [42], authors try with both an Inception V3 model and a ResNet50.
- In [43], an Inception V3 model is also used for classification in 757 different disease classes.

In fact, Inception V3 [44], alongside ResNet, is the most common architecture in the bibliography. Table 1 shows the publications that make direct use of convolutional neural networks for Alzheimer's diagnosis and the previously mentioned diseases.

TABLE 1
PUBLICATIONS THAT USE CNN

Paper	Disease	Architecture
[30]	Alzheimer	Inception + LeNet5
[41]	Diabetic Retinopathy	Inception V3
[34]	Alzheimer	VoxCNN + VoxResNet
[43]	Skin cancer	Inception V3
[42]	Breast cancer	Inception V3 + ResNet50
[36]	Alzheimer	Inception V3

LeNet5 is a relatively old architecture derived from [20]. VoxCNN is a volumetric convolutional neural network derived from VGGNet. VoxResNet is a volumetric residual network [45].

Regarding the diagnosis of Alzheimer's disease, the first attempt to directly train convolutional neural networks did not use ImageNet weights or fine-tunning, but they did use two of one of these common architectures (Inception V1) and an older one from before AlexNet [30]. The authors argued that the limited data available would lead to overfitting when using more advanced and complex architectures. They combined both models and designed a decision-making algorithm, resulting in nearly 100% accuracy.

Subsequently, another publication sought to demonstrate that a CNN could be easily used to generate features and classify, both in a fully automatic way [34]. To do this, they simplified model creation to the maximum, using only 231 images. They built a 3D CNN inspired by VGGNet, which they called VoxCNN; and compared it with a VoxResNet model. Making a direct comparison with [26], the results were significantly worse, but they highlighted the simplicity of implementing their solution.

Finally, the most recent publication fine-tunes an InceptionV3 network with 18F-FDG PET images, using only the Keras framework [16] to implement the models and SciPy [46] to preprocess the images. By making a direct comparison with radiologists, they demonstrated that their model outperformed them statistically significantly, especially in anticipating the onset of Alzheimer more than six years in advance.

CHALLENGES

The different uses of Artificial Intelligence in clinical applications face a series of challenges, some of them similar to those found by these same techniques in other fields. The vast majority are data related, but there are also ethical and philosophical challenges or dilemmas.

The first and most obvious challenge is the limited availability of labeled data. Although it is true that this problem diminishes with the passage of time, it still is a major concern between researchers. Especially when comparing with other datasets like ImageNet. While there are millions of images in this data bank, OASIS Brains has MRI and PET data of only 1098 subjects [47]. While it is true that, normally, the number of classes in medical applications is also smaller (ImageNet works with 1000 classes), that is not always the case. For example, it has already been pointed out that, in [43], classification is made between 757 labels.

This problem usually leads to overfitting. A common solution was the extraction of several random patches from the images, both two-dimensional [23] and three-dimensional [24]. This technique bears some resemblance to how radiologists work, examining the images by regions [31]. Others extract the patches in a not totally random way, looking to relate multiple patches from each images to make use of contextual information [48].

Other solutions are data augmentation [36], much less common, or even the creation of synthetic images [49]. But probably the most

common technique in the last publications would be transfer learning. When fine-tunning an Inception network or a ResNet, it is expected to need less training data.

Related to the data availability problem, there is the unbalanced data problem. The negative class is usually over-represented when comparing to the positive class. This is expected, since obtaining the information of healthy patients is much easier. To make things worse, the negative class is often strongly correlated, while there is a lot of variation in the positive class. Research on this topic suggests that undersampling the over-represented class is never a good idea, while oversampling the under-represented one could help in some cases [50].

The structural variety of the images is another important topic. Aside from the number of different scan types that can be used (MRI, fMRI, PET, etc.), these can be used in different ways. Mainly, the decision is based on whether to use the three-dimensional data directly, since all these images are usually in 3D, or transform the images to 2D. 3D data should be the default choice, since it prevents the loss of information [24]–[26], [34]. However, there has been a tendency to transform the images into 2D, mainly because it is much faster and is easier to avoid overfitting [23], [30], [36]. In [24], a direct comparison is made between extracting 2D and 3D patches from the images, concluding that the differences are not that significant.

In addition to the technical challenges, there are several ethical and philosophical problems. HER are very sensible data, which not only limit the number of images that can be gathered, but also forces to make extremely careful use of them. Related to this, confidence in AI is another important issue, since there is still a great ignorance in the population about what AI is or how it really works. It is a very important and active research topic, not only related to medical applications [51], [52]. It is commonly referred to as the black box problem. In Alzheimer-related publications, there have been some attempts to ease this problem [36], but that is not the common trend.

IV. IMPLEMENTATION

The main conclusion that can be extracted from the bibliography study is that convolutional neural networks are the default option. Instead, the discussion focuses on how to preprocess the images and how to train the models. The data preprocessing techniques used in the descriptive phase are explained in the following paragraphs. With regards to the training process, there are two options:

- Perform fine-tunning of networks pre-trained with Imagenet. In this case, the images have to be converted to 2D, because these networks work with RGB images (two-dimensional with a third channel for color). At this point, it is also possible to make use of different architectures, such as Inception or ResNet.
- Use custom designed networks or with weights initialized with traditional techniques. The networks can work with information in 2D or 3D, depending on whether is preferred to reduce the cost of training or maintain the detail, respectively.

Data preprocessing, as well as development and validation of the models were done in Google Colaboratory [53], due to important limitations with the local hardware. This limited the experimentation with the models, since Colaboratory is an interactive tool, not intended for long training sessions in the background.

Dataset

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

From the data provided by ADNI, more than 3000 T1-weighted MRI images were obtained. The patients are between 55 and 92 years old and were assigned one of three possible labels: AD, MCI and NC. The complete distribution can be seen in Table 2.

Of all this data, 15% is always kept outside the training, to be used as a test set. At the same time, another 15% is separated from the remaining training set to be used as a validation set for evaluating the quality of the model as the training progresses (with each epoch).

Image Preprocessing

MRI images were spatially normalized to an isotropic resolution of $2mm^3$, which reduces the resolution of most of them. Then, they were registered to a common atlas using affine registration with SimpleElastix [14]. The selected atlas was the MNI 305 [54]–[57]. As a result, the final resolution of the images was 78x110x86.

The registered images were skull-stripped using the FSL BET [15] tool implemented into Python by the library Nipype [58]. After a few tests, a fractional intensity threshold of 0.2 was used for the extraction. Fig. 2 shows three cuts of an example image result.

These preprocessing techniques may be quite complex, but they are carried out with an extremely simple approach. Libraries and open source tools are used and applied indistinctly to all images, seeking the best possible results, but without treating each image separately. This is a good way to test the extent to which the tools available today make it possible to deploy systems of this type, as individual image processing may not be affordable for data scientists with no in-depth experience in radiology.

In addition to the resulting 3D image dataset, a second dataset is created with the same images transformed into 2D. This is necessary for fine-tunning networks pre-trained with ImageNet. Keeping the simple approach, the transformation was inspired in the same procedure carried out in [36], but much simpler. The idea was to make multiple axial (horizontal) cuts and place them on the same plane to construct a two-dimensional image. The space between cuts was uniform, taking one of every two slices. In total, 16 different cuts were taken and placed on a "matrix" of 4x4 dimensions. The resulting two-dimensional image was replicated three times to adopt RGB dimensions. Thus, 16 cuts of size 110x86 result in an image of 440x344x3. Fig. 3 shows an image example result of this procedure.

It is worth noting that the images of both datasets were whitened for zero mean and unit standard deviation. They were also stored in several TFRecords files [59] for faster access during training.

TABLE 2.
CLASS DISTRIBUTION FOR MRI IMAGES

	Number of images
AD (Alzheimer's Disease)	636
MCI (Mild Cognitive Impairment)	1636
NC (Normal Cohort)	903
TOTAL	3175

Patients with MRI images are between 55 and 92 years old. The size of the data set is in line with what has been used for the main investigations referenced in section III. It is also important to note that the data set is considerably unbalanced.

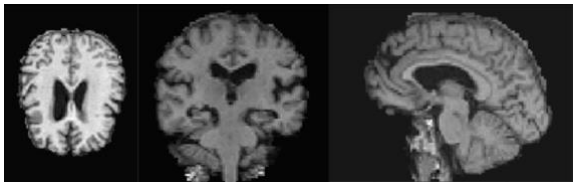

Fig. 2. Axial, coronal and sagittal cuts, respectively, of a registered and skull-stripped image

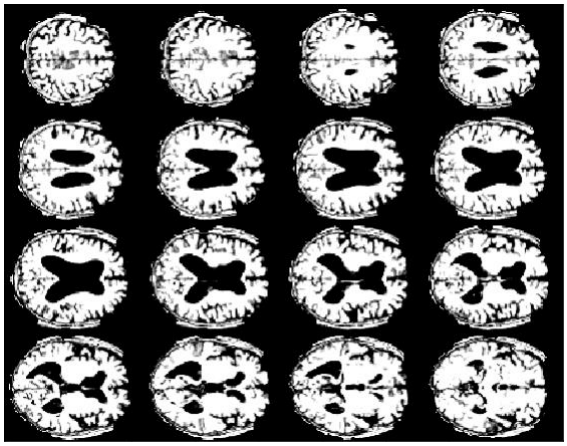

Fig. 3. Result image example after 2D conversion

Inception V3

Following the successful trends of recent years to adapt pre-trained Imagenet networks for medical applications [36], [41]–[43], the created 2D dataset was used for fine-tunning an Inception V3 network. The only modification was a final layer with 512 units and a dropout regularization of 80%, as well as a final softmax layer.

In a first training phase, only the two last layers were trained for 10 epochs with the Adam optimizer [60], a learning rate of 10^{-4} and a batch size of 8. A learning rate decay of 10^{-3} was also used. Next, the entirety of the network was fine-tuned, increasing the number of epochs to 70 and decreasing the learning rate decay to 10^{-7}, and keeping the rest of hyperparameters with the same values. Categorical cross-entropy was the lost function, and accuracy was also monitored.

For model evaluation, accuracy is not a very good metric because of class imbalance [50], so a method for drawing the ROC curve was implemented based on scikit-learn [61].

The model was also trained performing undersampling on the MCI class, which was over-represented when comparing with the other two.

ResNet3D

The second approach was based on a residual network adapted for volumetric data, implemented on Keras [62]. In this case, the 3D dataset was used for training, which makes the process much slower and increases the need for more images. In general, the more complex the network and the more information it has to process, the more observations it needs to learn from. Due to these limitations, tests were only made with the two smallest available networks, made of 18 and 34 layers.

The 18-layer 3D ResNet had its last layer removed and was added a fully connected layer with 512 units with 80% dropout and a final softmax layer. It is the same procedure performed for the Inception V3 model. It was trained for 50 epochs with the Adam optimizer, a learning rate of 10^{-5} and a batch size of 8. Categorical cross-entropy was used as the loss function, and accuracy was also monitored. Models fell easily into overfitting, so some experimentation was made with the regularization factor, which controls the L2 regularization. The best values ranged between 0.03 and 0.05.

The 34-layer 3D ResNet was trained with similar hyperparameters, but for 70 epochs. Training a 50-layer network was unbearable, since only three epochs took around 30 minutes.

V. RESULTS

Due to the hardware limitations and the use of Google Colaboratory, the models could not be properly validated. Several training attempts were made to study the potential of these networks, but cross validation and other stricter validation techniques were not possible. Despite that, interesting conclusions could be drawn from the results.

Fine-tuning the Inception V3 network consistently yields better results than training a 3D residual network, probably because of the difference in network complexity and the amount of available data. However, training on both models is highly irregular, and controlling overfitting becomes a very difficult task.

Fig. 4 and Fig. 5 show the evolution of accuracy and loss for both models. Other training attempts yielded similar results, with stepped curves and irregular results. Only the ResNet3D kept a decent evolution of validation loss in comparison with the training loss, but overall results ended up being worse than with the Inception V3. Class imbalance and class similarity could be the main causes of this irregularities.

However, validation accuracy and validation loss were not the final evaluation metrics. Instead, ROC curves and AUC were computed for each class. Although, again, further validation would be helpful, strong models could be built by fine-tuning the Inception V3. Fig. 6 shows the best results that could be obtained in this case. Maintaining a very simple approach, the model is one of the strongest in three-class classification in the state of the art when looking at the AUC values. It is worth noting how the over-represented class is also the one with which the model has the most difficulties (MCI). The same happens in [36], where the over-representation is less notable, so this difficulties could be inherent in the very nature of the classes.

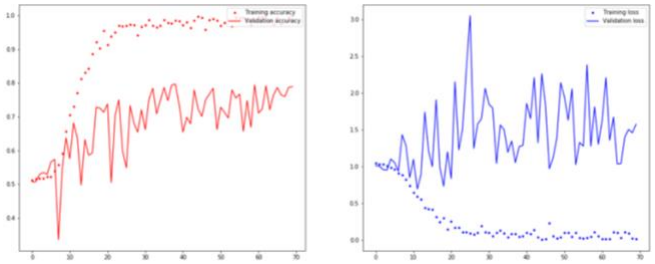

Fig. 4. Accuracy (left) and loss (right) during training for an Inception V3 network

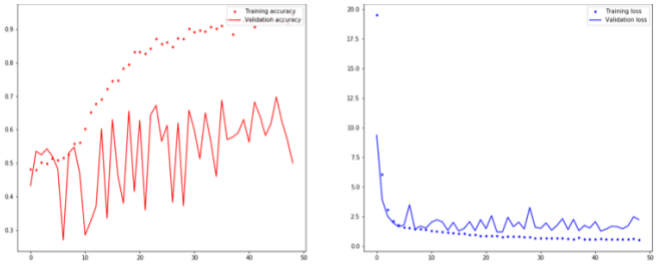

Fig. 5. Accuracy (left) and loss (right) during training for an 18-layer ResNet3D with a regularization factor of 0.05

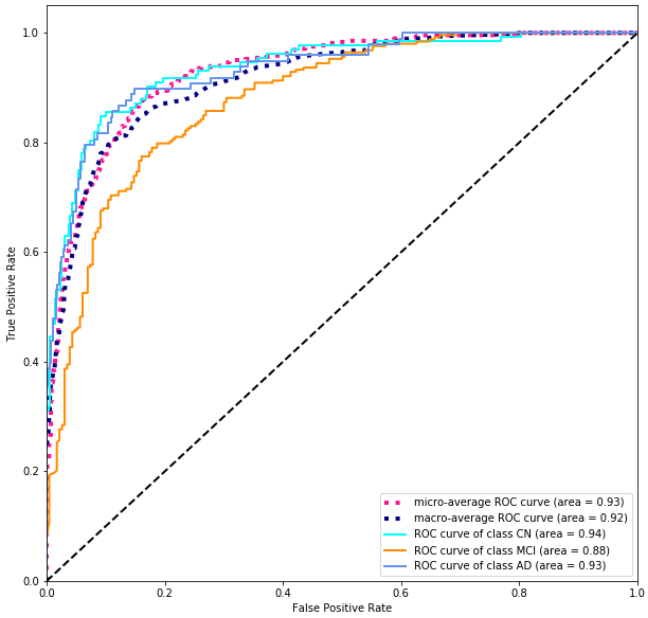

Fig. 6. Best results obtained by fine-tuning an Inception V3 network

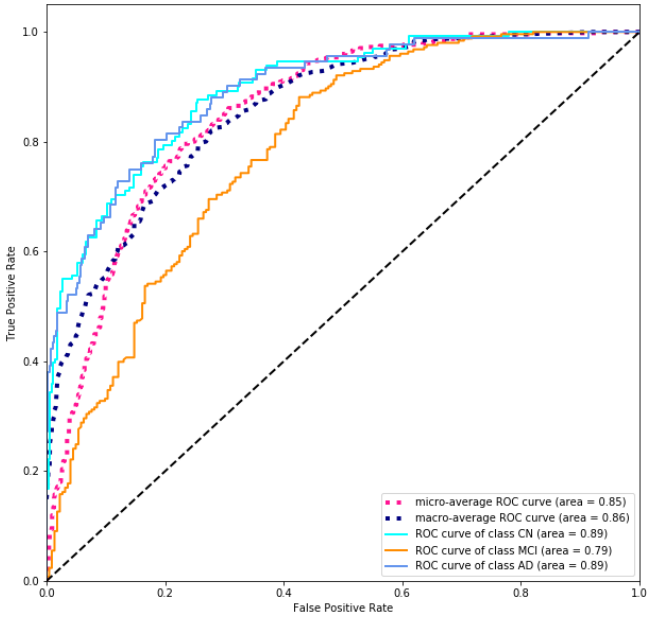

Fig. 7. Best results obtained with a ResNet3D. 18 layers and a regularization factor of 0.03

Fig. 7 shows the best results obtained by training a ResNet3D. Increasing the depth to 34 layers did not lead to better results, although experimentation with such depth was limited due to the slowness of training. Still, results are notably better by fine-tuning the Inception V3, so there does not seem to be any interesting potential in using a 3D architecture.

VI. ANALYSIS AND DISCUSSION

Contextualizing the results with the current state of the art, two publications should be considered. In [36], authors keep a similar approach based on 2D transformation and fine-tuning an Inception V3. Results obtained in that publication are around five points better, with an AUC of 0.98 with the AD class, against an AUC of 0.93 obtained in this research. However, there are key differences between both researches that should be noted.

First, the scan type. While T1-weighted MRI images were used in this work, 18FGD-PET images are used in [36]. This difference is important because MRI images are much cheaper and harmless for the patient [63]–[66]. For that reason, building a successful diagnosis system based on MRI images is much more desirable.

Second, the simplicity-based approach. Although the 2D transformation procedure was very similar, authors from that publication performed much more complex steps. For example, "connected component analysis was used to derive the relevant imaging volume by selecting the cranial-most and caudal-most sections representing more than 100x100 mm^2 of brain parenchyma" [36]. A simpler process like the one performed in this research proves to be almost as valid. Also, image registration and skull-stripping were performed with easy-to-use tools.

With regard to the 3D networks, the results in [34] should be taken into account. That research also tries to maintain a very simple approach, by using directly 3D images and a volumetric residual network. When considering the best model obtained in this research, results are quite similar. For example, AUC for AD vs NC in [34] is around 0.87, while the AUC for both classes in Fig. 7 gets to 0.89. The key difference here is that this research performs three-class classification, while in [34] authors build multiple binary classifiers. The results obtained in this research suggest that multiclass classification is an equally feasible development path.

It is important to note that average values are considered in that research, comparing to the maximum values in this one. Further validation could not be performed due to hardware limitations.

Finally, it should be noted that these types of models still have a long way to go before they can be deployed in real clinical environments. Looking at the results directly, it would be difficult to say they are precise enough, and therefore their diagnosis shouldn't be taken into account. On the other hand, it is also true that other studies have carried out a direct comparison with radiologists and have issued surprising conclusions, stating that the models are significantly superior [36]. In such a case, the best models should always be directly compared with human professionals in order to extract a measure of real quality, and to estimate whether it is really feasible to deploy them in real environments.

VII. RECOMMENDATIONS

After a thorough analysis of the results and comparison with other important work of recent years, a series of recommendations can be made to a data scientist with no in-depth knowledge of medicine looking to build a CADx Alzheimer's diagnostic system. First of all, it should be noted that the use of convolutional neural networks should be the immediate option, since they capture practically all applications in medical image analysis in general, not only in Alzheimer's diagnosis.

Speaking of the data, it could be said that the ADNI database contains enough quality images to build good models. This is something that could be concluded from previous work [34]. However, processing this information is a bit more complicated, as the open source tools available are not perfect, especially when it comes to skull-stripping. It would be advisable to treat each group of images individually, seeking to adapt the fractional intensity threshold (in the case of FSL BET) as appropriate. The problem is that the only way to manage if the threshold value is correct is by directly observing the results, and this can be very expensive for very large sets of images. On the other hand, image registration tools work quite well. They are sufficient to treat images automatically, as long as they are registered to an atlas of the same modality (T1-weighted, in the case of this work) and all images are normalized to the same spatial resolution.

At this point, it should be noted that the higher the spatial resolution, the larger the resulting images will be, and each voxel will represent

smaller regions. This suggests that using an isotropic resolution of $1mm^3$ would be better for detail, but the larger size of the images would make the training process slower. Therefore, it is a matter of time and hardware capacity.

With regard to the structure of the images, it would be much more advisable to transform the images into 2D, even if it was with a very simple procedure such as the one used in this work. 3D networks are much more complicated to train, slower and get worse results. 2D networks, on the other hand, train faster, and there is a great variety of advanced models for fine-tuning. In addition, they are much more intuitive, in the sense that they require orienting the problem in a similar way as a radiologist would do, examining multiple cuts or slices individually.

Speaking of the scan types, both MRI and PET can be used. The latter do not require skull-stripping and have shown excellent results in other publications, but they are more expensive to obtain and more harmful to the patient. MRI, on the other hand, is inexpensive and harmless to patients, but requires more careful pre-processing. In the end, choice comes down to availability and stakeholder preferences.

It is also important to decide how training should be conducted. In general, using a recognized architecture, such as Inception or ResNet, should be the default option, as they have proven to be very good in a wide variety of applications. In the case of this work, better results have been achieved using small batch sizes and small learning rates. An adaptive optimizer, such as Adam, would also be very suitable. In addition, recent work has also followed these guidelines [34], [36].

Finally, it is important to note that achieving sufficiently good results is extremely complicated. The treatment of the MCI class tends to considerably worsen the performance of the models. For this reason, it would be advisable to reduce the classification to binary (AD vs NC) in critical applications where performance must be maximized at all costs.

Above all, the main recommendation would be not to deploy a system of this type. The diagnosis of a disease is an extremely sensitive issue, for which not only excellent results are necessary -which have not been achieved by the time of writing these lines- but also an in-depth examination of the reliability of the models. In the coming years, the increasing amount of data available should alleviate these problems but, for the time being, more and better research is required.

VIII. CONCLUSIONS

In this work, a series of contributions have been made. First, the state of the art of medical image analysis and Alzheimer's diagnosis with AI, until early 2019, has been analyzed and summarized. This is an excellent basis for researchers or other professionals looking to get started in this field.

Second, from the previous state of the art analysis, the most important Deep Learning models were identified. Later, they were implemented in an experimental scenario, looking to estimate the performance they are capable of providing in the task of Alzheimer's diagnosis. All while making use of simple tools available today to any developer or data scientist.

Last but not least, results of the previous experimentation were analyzed for emitting the appropriate recommendations to those professionals who seek to implement models of this type without having in-depth knowledge of medicine. It is concluded that it would be very difficult to produce a successful model of this type in a simple way, at least with the data available today.

The contributions of this work also open up new research lines that are worth being explored in the future. The first one is already addressed in the recommendations and suggests that a more thoughtful skull-stripping process could improve the results considerably. In particular, the fractional intensity threshold could be adjusted based on the normalization procedure that has been carried out on the images previously. In the ADNI database, images are marked with the possible transformations applied to them, with labels such as B1-corrected or N3-scaled. It would be within the reach of any data scientist to group the images by means of these labels and try to find the best fractional intensity threshold for each one.

Also, although deep feature extraction was discarded because it is gradually falling into disuse, it would be interesting to test it first hand in the case of Alzheimer's diagnosis. The problems brought about by the similarities between the images of different classes might be reduced when using this technique.

Other possible improvements would be using higher resolution images (3T scans from ADNI) or trying several data augmentation techniques.

REFERENCES

[1] S. E. Dilsizian y E. L. Siegel, «Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment», *Curr. Cardiol. Rep.*, vol. 16, n.º 1, p. 441, 2014.

[2] L. M. Valetts, S. B. Navarro, y R. F. Gesa, «Modelling Collaborative Competence Level Using Machine Learning Techniques.», en *e-Learning*, 2008, pp. 56-60.

[3] U. Gretzel, M. Sigala, Z. Xiang, y C. Koo, «Smart tourism: foundations and developments», *Electron. Mark.*, vol. 25, n.º 3, pp. 179-188, 2015.

[4] I. Millington y J. Funge, *Artificial intelligence for games*. CRC Press, 2009.

[5] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, y B. J. Erickson, «Deep learning for brain MRI segmentation: state of the art and future directions», *J. Digit. Imaging*, vol. 30, n.º 4, pp. 449-459, 2017.

[6] F. Ciompi *et al.*, «Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box», *Med. Image Anal.*, vol. 26, n.º 1, pp. 195-202, 2015.

[7] S.-C. Lo, S.-L. Lou, J.-S. Lin, M. T. Freedman, M. V Chien, y S. K. Mun, «Artificial convolution neural network techniques and applications for lung nodule detection», *IEEE Trans. Med. Imaging*, vol. 14, n.º 4, pp. 711-718, 1995.

[8] W. Thies y L. Bleiler, «2012 Alzheimer's disease facts and figures.», *Alzheimer's Dement. J. Alzheimer's Assoc.*, 2012.

[9] C. Reitz, C. Brayne, y R. Mayeux, «Epidemiology of Alzheimer disease», *Nat. Rev. Neurol.*, vol. 7, n.º 3, p. 137, 2011.

[10] H. Greenspan, B. van Ginneken, y R. M. Summers, «Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique», *IEEE Trans. Med. Imaging*, vol. 35, n.º 5, pp. 1153-1159, 2016.

[11] J. Ker, L. Wang, J. Rao, y T. Lim, «Deep Learning Applications in Medical Image Analysis», *IEEE Access*, vol. 6, pp. 9375-9389, 2018.

[12] G. Litjens *et al.*, «A survey on deep learning in medical image analysis», *Med. Image Anal.*, vol. 42, n.º 1995, pp. 60-88, 2017.

[13] D. Shen, G. Wu, y H.-I. Suk, «Deep learning in medical image analysis», *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221-248, 2017.

[14] K. Marstal, F. Berendsen, M. Staring, y S. Klein, «SimpleElastix: A user-friendly, multi-lingual library for medical image registration», en *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 134-142.

[15] S. M. Smith, «Fast robust automated brain extraction», *Hum. Brain Mapp.*, vol. 17, n.º 3, pp. 143-155, 2002.

[16] F. Chollet y others, «Keras». 2015.

[17] C. R. Jack *et al.*, «The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods», *J. Magn. Reson. Imaging An Off. J. Int. Soc. Magn. Reson. Med.*, vol. 27, n.º 4, pp. 685-691, 2008.

[18] J. P. B. O'Connor *et al.*, «Imaging biomarker roadmap for cancer studies.», *Nat. Rev. Clin. Oncol.*, vol. 14, n.º 3, pp. 169-186, 2017.

[19] S. Klöppel *et al.*, «Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method», *Brain*, vol. 131, n.º 11, pp. 2969-2974, 2008.

[20] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, y others, «Gradient-

based learning applied to document recognition», *Proc. IEEE*, vol. 86, n.º 11, pp. 2278-2324, 1998.

[21] A. Krizhevsky, I. Sutskever, y G. E. Hinton, «Imagenet classification with deep convolutional neural networks», en *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[22] W. Yang *et al.*, «Independent component analysis-based classification of Alzheimer's disease MRI data», *J. Alzheimer's Dis.*, vol. 24, n.º 4, pp. 775-783, 2011.

[23] A. Gupta, M. Ayhan, y A. Maida, «Natural image bases to represent neuroimaging data», en *International conference on machine learning*, 2013, pp. 987-994.

[24] A. Payan y G. Montana, «Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks», *arXiv Prepr. arXiv1502.02506*, 2015.

[25] E. Hosseini-Asl, R. Keynton, y A. El-Baz, «Alzheimer's disease diagnostics by adaptation of 3D convolutional network», en *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 126-130.

[26] E. Hosseini-Asl, G. Gimel'farb, y A. El-Baz, «Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network», *arXiv Prepr. arXiv1607.00556*, 2016.

[27] E. Hellström, «Feature learning with deep neural networks for keystroke biometrics: A study of supervised pre-training and autoencoders». 2018.

[28] R. P. Woods, J. C. Mazziotta, S. R. Cherry, y others, «MRI-PET registration with automated algorithm», *J. Comput. Assist. Tomogr.*, vol. 17, p. 536, 1993.

[29] A. Klein *et al.*, «Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration», *Neuroimage*, vol. 46, n.º 3, pp. 786-802, 2009.

[30] S. Sarraf, G. Tofighi, y others, «DeepAD: Alzheimer′ s disease classification via deep convolutional neural networks using MRI and fMRI», *BioRxiv*, p. 70441, 2016.

[31] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative, y others, «Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis», *Neuroimage*, vol. 101, pp. 569-582, 2014.

[32] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative, y others, «Latent feature representation with stacked auto-encoder for AD/MCI diagnosis», *Brain Struct. Funct.*, vol. 220, n.º 2, pp. 841-859, 2015.

[33] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, y D. J. Hawkes, «Nonrigid registration using free-form deformations: application to breast MR images», *IEEE Trans. Med. Imaging*, vol. 18, n.º 8, pp. 712-721, 1999.

[34] S. Korolev, A. Safiullin, M. Belyaev, y Y. Dodonova, «Residual and plain convolutional neural networks for 3D brain MRI classification», *2017 IEEE 14th Int. Symp. Biomed. Imaging (ISBI 2017)*, pp. 835-838, 2017.

[35] M. Rajchl, S. I. Ktena, y N. Pawlowski, «An Introduction to Biomedical Image Analysis with TensorFlow and DLTK», *Medium.com*, 2018. [En línea]. Disponible en: https://medium.com/tensorflow/an-introduction-to-biomedical-image-analysis-with-tensorflow-and-dltk-2c25304e7c13. [Accedido: 09-jun-2019].

[36] Y. Ding *et al.*, «A Deep learning model to predict a diagnosis of alzheimer disease by using 18F-FDG PET of the brain», *Radiology*, vol. 290, n.º 2, pp. 456-464, 2018.

[37] K. Simonyan y A. Zisserman, «Very deep convolutional networks for large-scale image recognition», *arXiv Prepr. arXiv1409.1556*, 2014.

[38] C. Szegedy *et al.*, «Going deeper with convolutions», en *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[39] K. He, X. Zhang, S. Ren, y J. Sun, «Deep Residual Learning for Image Recognition», 2015.

[40] «Imagenet», 2016. [En línea]. Disponible en: http://www.image-net.org/.

[41] V. Gulshan *et al.*, «Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs», *JAMA - J. Am. Med. Assoc.*, vol. 316, n.º 22, pp. 2402-2410, 2016.

[42] S. Vesal, N. Ravikumar, A. A. Davari, S. Ellmann, y A. Maier, «Classification of Breast Cancer Histology Images Using Transfer Learning», *PLoS One*, vol. 12, n.º 6, pp. 812-819, 2017.

[43] A. Esteva *et al.*, «Dermatologist-level classification of skin cancer with deep neural networks», *Nature*, vol. 542, n.º 7639, p. 115, 2017.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, y Z. Wojna, «Rethinking the inception architecture for computer vision», en *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818-2826.

[45] H. Chen, Q. Dou, L. Yu, y P.-A. Heng, «VoxResNet: Deep Voxelwise Residual Networks for Volumetric Brain Segmentation», *arXiv Prepr. arXiv1608.05895*, 2016.

[46] E. Jones, T. Oliphant, P. Peterson, y others, «SciPy: Open source scientific tools for Python». 2001.

[47] «OASIS Brains Datasets». [En línea]. Disponible en: https://www.oasis-brains.org/#data.

[48] T. Tong *et al.*, «Multiple instance learning for classification of dementia in brain MRI», *Med. Image Anal.*, vol. 18, n.º 5, pp. 808-818, 2014.

[49] E. Castro, A. Ulloa, S. M. Plis, J. A. Turner, y V. D. Calhoun, «Generation of synthetic structural magnetic resonance images for deep learning pre-training», en *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 2015, pp. 1057-1060.

[50] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, y G. D. Tourassi, «Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance», *Neural networks*, vol. 21, n.º 2-3, pp. 427-436, 2008.

[51] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, y K.-R. Müller, «Explaining nonlinear classification decisions with deep taylor decomposition», *Pattern Recognit.*, vol. 65, pp. 211-222, 2017.

[52] M. D. Zeiler y R. Fergus, «Visualizing and understanding convolutional networks», en *European conference on computer vision*, 2014, pp. 818-833.

[53] «Colaboratory: Frequently asked questions», 2018. .

[54] A. C. Evans, «An MRI-based stereotactic atlas from 250 young normal subjects», *Soc. neurosci. abstr*, *1992*, 1992.

[55] D. L. Collins, P. Neelin, T. M. Peters, y A. C. Evans, «Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space.», *J. Comput. Assist. Tomogr.*, vol. 18, n.º 2, pp. 192-205, 1994.

[56] A. C. Evans *et al.*, «Anatomical mapping of functional activation in stereotactic coordinate space», *Neuroimage*, vol. 1, n.º 1, pp. 43-53, 1992.

[57] A. C. Evans, D. L. Collins, S. R. Mills, E. D. Brown, R. L. Kelly, y T. M. Peters, «3D statistical neuroanatomical models from 305 MRI volumes», en *1993 IEEE conference record nuclear science symposium and medical imaging conference*, 1993, pp. 1813-1817.

[58] K. Gorgolewski *et al.*, «Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python», *Front. Neuroinform.*, vol. 5, p. 13, 2011.

[59] «Using TFRecords and tf.Example». [En línea]. Disponible en: https://www.tensorflow.org/tutorials/load_data/tf_records.

[60] D. P. Kingma y J. Ba, «Adam: A method for stochastic optimization», *arXiv Prepr. arXiv1412.6980*, 2014.

[61] F. Pedregosa *et al.*, «Scikit-learn: Machine Learning in {P}ython», *J. Mach. Learn. Res.*, vol. 12, pp. 2825-2830, 2011.

[62] JihongJu, «keras-resnet3d», *Github.com*, 2017. [En línea]. Disponible en: https://github.com/JihongJu/keras-resnet3d.

[63] M. D. Feldman, «Positron Emission Tomography (PET) for the Evaluation of Alzheimer's Disease/Dementia», *Calif. Technol. Assess. Forum*, 2004.

[64] J. M. Hoffman *et al.*, «FDG PET imaging in patients with pathologically verified dementia.», *J. Nucl. Med.*, vol. 41, n.º 11, pp. 1920-8, 2000.

[65] A. Nordberg, J. O. Rinne, A. Kadir, y B. Lngström, «The use of PET in Alzheimer disease», *Nat. Rev. Neurol.*, vol. 6, n.º 2, pp. 78-87, 2010.

[66] B. J. Pichler, A. Kolb, T. Nagele, y H.-P. Schlemmer, «PET/MRI: Paving the Way for the Next Generation of Clinical Multimodality Imaging Applications», *J. Nucl. Med.*, vol. 51, n.º 3, pp. 333-336, 2010.