CURRICULUM LEARNING FOR EARLY ALZHEIMER'S DISEASE DIAGNOSIS

Catarina Mendes Faustino Gracias - catarina.gracias@hotmail.com

Instituto Superior Técnico, Lisbon, Portugal

October 2021

ABSTRACT

The early and asymptomatic stages of Alzheimer's Disease (AD), such as mild cognitive impairment (MCI), are hard to classify, even by experienced physicians. Deep learning approaches, such as convolutional neural networks (CNNs), have been shown to help, achieving similar or even better results. Although these methods have the advantage that features are automatically extracted from images rather than handcrafted, they do not allow for incorporating medical knowledge. In this thesis we propose to implement curriculum learning (CL) strategies into CNNs designed to diagnose healthy subjects, MCI and AD, as a way to incorporate medical knowledge to boost the networks performance for early AD diagnosis. CL is a training strategy of the networks that tries to mimic the way humans, in this case doctors, learn. Several CL strategies were implemented and compared to commonly used baseline deep learning models. The results showed that they clearly improve the F1-score (up to 3.3%) and overall accuracy (up to 4.5%), particularly that of MCI (up to 11.3%).

***Index Terms*—** Alzheimer's Disease; Curriculum Learning; Convolutional Neural Network; Mild Cognitive Impairment; Medical Knowledge

1. INTRODUCTION

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder and one of the leading causes of death in developed countries, since there is not a cure available yet [1]. In the early and asymptomatic stages of AD, patients are classified as having mild cognitive impairment (MCI), while healthy patients are denominated as normal controls (NC). The clinical research developed towards finding therapeutics and a cure for AD highly depends on the ability to diagnose patients accurately and at an early stage of the disease, when it is still possible to delay the onset of AD. AD diagnosis is performed by medical doctors, who have access to patient's information: medical images, genetic data and cognitive tests, such as Mini Mental State Examination (MMSE) and Clinica Dementia Ratio (CDR). However, MCI stages are not easily identified solely by following these traditional diagnostic approaches. Consequently, AD research benefits from the use of deep learning methods to make faster, earlier and more accurate diagnosis [2, 3]. Currently, Convolutional Neural Networks (CNNs), which allow features being automatically extracted rather than handcrafted, have already been successful in AD diagnosis through the classification of medical images [3]. Nevertheless, these recent approaches still have some drawbacks, such as not being optimized to incorporate medical knowledge and the vulnerability to overfitting problems, which are often related to the the small size of available medical datasets.

In this thesis, as a way to overcome these bottlenecks, we propose to develop and evaluate novel curriculum learning (CL) strategies to more accurately diagnose early AD. They will incorporate medical knowledge, such as the doctors training pattern, scores of the patient's cognitive tests and regions of interest (ROI) for AD diagnosis, into neural networks.

2. BACKGROUND AND RELATED WORK

2.1. Alzheimer's disease

AD slowly destroys memory and the person's ability to reason and function independently, being the advancing of age one of its greatest risk factors [3]. It can be characterized as a combination of cognitive, motor and behavioural deterioration, which eventually becomes overwhelming and devastating both to patients and their families [4].

The appropriate monitoring of AD's biomarkers results in earlier diagnosis and better patient care. The most common methods to measure the evolution of AD's biomarkers are 18F-Fluorodeoxyglucose - Positron Emission Tomography (FDG-PET), Magnetic Resonance Imaging (MRI) and cognitive tests. The first two allow to measure functional and structural changes in the brain, respectively, whereas the later assess the cognition level of the patients. Early diagnosis includes recognition of the pre-demented conditions, before clinical symptoms develop, allowing to identify those who would benefit from therapeutic intervention and therefore, delaying the onset of the disease. Moreover,

earlier diagnosis can be extremely helpful to signal patients to clinical trials, which is a crucial step for cure development [5].

FDG-PET: Imaging technique that provides information about physiological and biochemical processes of the body. In FDG-PET measurements, patients with AD have characteristic reductions in regional brain activity (temporoparietal hypometabolism), which are progressive and correlate with dementia severity [6].

CDR: Global rating instrument used to characterize cognitive and functional performance. The CDR score is calculated on the basis of testing six different cognitive and behavioral domains: memory, orientation, judgment and problem solving, community affairs, home and hobbies performance, and personal care. The CDR is based on a discrete scale of 0–5, presented in Figure 1 (a) [7], which reflects the degree of Cognitive Impairment (CI).

MMSE: 30-point test used to measure thinking ability or "cognitive impairment". The MMSE scores are based in a continuous scale. The scores and the corresponding level of dementia are presented in Figure 1 (b) [8].

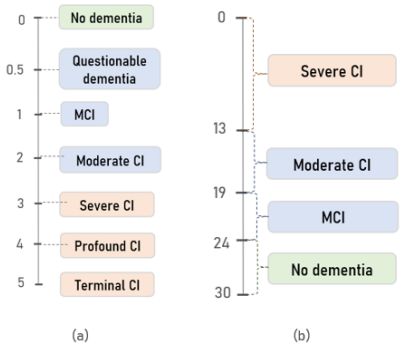

Fig. 1. (a) CDR test scores: (b) MMSE test scores.

2.2. Deep learning for AD diagnosis

2.2.1. Relevance of machine learning for AD diagnosis

With new technologies arising, the amount and diversity of patient data acquired over the years has exponentially increased, leading to complex and heterogeneous health datasets. Medical images, usually 3D images with high resolution, are the most widely used data for AD diagnosis but also the most complex to analyze, since they contain complex patterns [3]. However, to analyse thousands of images and learn their intrinsic discriminative patterns is extremely laborious, requiring a lot of practise and time, which most of clinicians do not have, even the most experienced ones. Consequently, neuroimaging was one of the first areas of neurology to benefit from the application of machine learning approaches to improve the interpretation of such medical images, and thereby, the diagnosis. Deep learning, which is a specific subfield of machine learning, have already proven its potential for different classification problems. The use of a particular type of deep neural network, known as a convolutional neural network (CNN), has led to significant performance improvements for image classification [9].

2.2.2. Convolutional neural networks

A CNN architecture consists of an input layer, that should receive image data, hidden layers and an output layer, which outputs the predicted label/class. The hidden layers are made up of several convolutional layers stacked with pooling layers, followed by fully-connected layers and a softmax or sigmoid layer in the end. The first layers work as automatic feature extractors, extracting discriminative features and the last layers allow task-specific classification using those same features [3].
CNNs were first introduced in 1989 by LeCun and colleagues [10]. Currently, for AD detection, the main competitor architectures are 3D CNNs and 2D CNNs (with or without recurrent neural networks (RNNs)) [3].

3D CNNs: Since neuroimaging techniques mostly provide 3D images, 3D CNNs became popular for AD detection. However, they are usually complex and associated with a large number of parameters, which combined with small sized datasets might result in overfitting [3]. Multiple AD studies use their own architectures, which can differ much on the number of convolutional layers used, their number of filters and activation function, while other focus on fine tuning well-known architectures. Basaia et al. [11] used twelve layers and Spasov et al. [12] used seven, both for predicting NC from AD and MCI subjects. Moreover, Bäckström et al. [13] achieved an effective 3D architecture by using five convolutional layers for feature extraction, followed by three fully-connected layers for AD/NC classification. Regarding well-known 3D architectures, Karasawa et al. [14] proposed an effective novel 3D CNN architecture, based on ResNet. Additionally, Cheng and Liu [15] used a 3D CNN structure inspired by LeNet with four convolutional layers for each image patch.

2D CNNs

2D CNNs were the first type of CNNs, which are specifically designed to recognize patterns in 2D images. Most of the studies that used 2D CNNs for 3D images either extract 2D information from the images by splitting volumetric data into image slices (without the use of RNNs) or they rely on the logic that a 3D image can be treated as a sequence of 2D images. In the former, studies using three convolutional layers are most common [16, 17]. Moreover,

Kazemi and Houghten [18] demonstrated that well known 2D structures, such as AlexNet and GoogLeNet performed well on fMRI images for classifying different stages of AD. In the latter, they use RNNs to extract the inter-slice features (similar structures in adjacent slices) while the 2D CNN captures the intra-slice features (similar structures in a single slice) [3]. They have been successfully applied to AD detection by Cheng and Liu [19] and Liu et al. [20].

Some of these methods shown slightly lower accuracies than the ones obtained with 3D CNNs. However, they train faster because 3D CNN are associated with a larger number of trainable parameters than 2D CNNs.

2.3. Curriculum learning

The process of data collection for medical domain studies is often associated with high costs and complexity. This is why these studies are usually characterized by limited samples, i.e. small-sized medical datasets [21]. As explained, the lack of data in deep learning models can lead to overfiting problems, which are usually solved by regularization or data augmentation techniques. These solutions, even though they effectively decrease model complexity, do not introduce any new information [22].

In recent years, introducing information beyond the one available in the dataset at hand has become a promising approach to address the problem of small-sized medical datasets. In the case of medical datasets the introduction of medical knowledge has been explored with promising results. One way of incorporating medical knowledge into deep learning models is through curriculum learning.

CL is a strategy of training machine learning models by mimicking the way humans learn. In this strategy, a curriculum is designed, which defines the order in which the data are presented to the model: the model is first trained with easier data (or tasks) and gradually more complex data (or tasks) are introduced, instead of being randomly presented [23]. Usually the curriculum is predefined (manual strategies). However, since defining a good curriculum manually is not an easy task, some strategies rely on learning the curriculum from the data, simultaneously with network training (automatic strategies).

CL have recently shown to improve the performance of CNNs for several medical image classification tasks [24, 25, 26]. Most approaches use a manual curriculum. For instance, Tang et al. [24] built a curriculum by categorizing the severity of patient injuries according to X-ray reports. By using it, they improved thoracic disease diagnosis from X-rays (AUC increased 3.19%). Haarburger et al. [25] used manually selected lesion-patch images for pre training the model and then fine tuned it with the whole MRI images, improving the AUC for breast cancer diagnosis by 27%. Automatic CL strategies have also been proposed. For example, Maicas et al. [26] proposed a meta learning approach for breast screening classification from DCE-MRI, which outperformed baseline approaches (AUC improved from 86% to 90%). Despite the recent success of CL strategies for medical image classification, they have still not been applied to networks for AD diagnosis.

3. METHODOLOGY

3.1. Data

The data used in the implemented strategies were collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. ADNI is a global research study that actively supports the investigationand development of treatments that slow or stop the progression of AD. FDG-PET images of 406 subjects, at baseline and at 6, 12 and 24 month follow-ups were used, labeled as Normal Control (NC), MCI or AD. For each image, the corresponding CDR and MMSE scores were also provided. The clinical profile of the groups studied is presented in Table 1.

Additionaly, ten provided ROIs were delineated and provided by an experienced physician, Professor Dr. Durval Campos Costa.

Table 1. Demographic and clinical profile of the groups studied $(mean \pm standard\,deviation)$.

Group	NC	MCI	AD
Number of subjects	104	207	95
Number of images	365	714	314
Age	76.9 ± 4.8	76 ± 7.3	76.5 ± 7.1
Sex (% M)	63.8	66.2	59.9
MMSE	29.1 ± 1.1	26.6 ± 3.2	21.6 ± 4.4
CDR	0.02 ± 0.2	0.5 ± 0.2	0.95 ± 0.5

3.1.1. Data pre-processing

All FDG-PET scans were normalized, averaged and co-registered by ADNI researchers, and were also further normalized in the range of [0,1] and cropped from 60x128x128 to 40x98x98, in order to remove most of the non-relevant area surrounding the brain.

The ten provided ROIs were rearranged into 8 different ones: symetrical ROIs with respect to the vertical axis of the coronal section of the brain were merged into one (since AD is not related to specific brain hemisphere) and also all ROIs were merged into one major ROI. Table 2 summarizes the information about the ROIs provided and the ones used in the project (first column).

Table 2. Available ROIs provided by Professor Dr. Durval Campos Costa, their name, percentage of brain area they occupy and the ROIs selected for this project.

ROIs		Name	Area (%)
1+2	1	Left lateral temporal	4.51
	2	Right lateral temporal	
3+4	3	Left mesial temporal	0.94
	4	Right measial temporal	
5	5	Inferior frontal gyrus/Orbitofrontal	0.84
6	6	Inferior anterior cingulate	0.71
7+8	7	Left dorsolateral parietal	2.66
	8	Right dorsolateral parietal	
9	9	Superior anterior cingulate	1.33
10	10	Posterior cingulate and precuneus	1.28
All	All Rois	———————-	12.29

3.1.2. Data division

To train and test the models a 5-fold cross-validation was performed. The subjects, and not the images, were separated into five folds, to guarantee that brain scans from the same subject were not present in different sets, avoiding data leakage. Five models were trained and each model used one of those folds for testing (around 20% of the dataset) and the remaining four for training (around 80% of the dataset). For each train, the subjects in the training set were further divided into subjects for training the model (80% out of the subjects of the original training test) and subjects for the validation of the model (20% out of the subjects of the original training test). In the end, to convert the subjects sets into image sets, all images from the same subject were added to the the correspondent set, originating the final training set (with 64% of the images), the final validation set (with 16% of the images) and the final test set (with 20% of the images).

3.2. Architecture and experimental design

The CL strategies were applied to a 3D-CNN. Its architecture consists of three convolutional blocks where the 3D convolutional layers are composed of 8, 16 and 32 filters, respectively, with ReLU activation function. Each convolutional layer is followed by a 3D max-pooling layer and a batch normalization layer. The output of the last convolution block is then flattened and fed into a fully connected classifier network, with 64 units and a softmax layer in the end, allowing the classification into 3 classes: NC, MCI and AD.

The experiments were performed on a single NVIDIA GeForce GTX 1070 GPU with 8GB of memory, in a machine with an Intel Core i7-6800K @ 3.40GHz CPU. For training the models, the ADAM optimizer was used and the categorical cross-entropy was chosen as the loss function. All models, expect for focal loss, were trained with a weighted training strategy, where the weight of the class was inversely proportional to the class frequencies in the train set. Moreover, a batch size of 16 was used, for a total number of 100 epochs, using an early stop criterion monitoring the validation loss with a patience of 50 epochs.

3.3. Curriculum learning strategies

To improve early AD diagnosis from medical images, curriculum learning strategies were applied to CNNs. Different strategies were implemented, nine manual, three automatic, eight use medical knowledge to build the curriculum (such as MMSE and ROI) and four do not. The manual strategies are further subdivided into complexity focused strategies, ROI focused strategies, mixed strategies and replicate automatic strategies, while the automatic ones are subdivided into self-paced learning and self-paced curriculum learning. All these strategies differ either on how the curriculum is built or on the information they use to build it.

3.3.1. Manual

Task strategy: In this approach, the samples are fed into the network ordered by task complexity. It follows the transfer learning proposal of [27], yet it is adapted to a CL strategy consisting in two rounds of training: in the first one the model is trained with only AD and NC samples (samples from only two classes and easier to distinguish between them), and in the second round the MCI samples are added (samples from three classes and harder to distinguish between them).

MMSE strategy: This strategy consists on feeding the network with samples and tasks ordered by difficulty. Therefore, the network first trains with a simpler task and the easier samples of that task and afterwards is challenged with more difficult ones. This strategy corresponds to 3 rounds of training, where in the end of the first two rounds, the last fully connected layer of the model (which contains the information about the predicted label) was replaced for randomly initicalez one. The MMSE score was used to define if a sample was easy or hard: an image was considered an easy sample if its label (NC, MCI and AD) and its corresponding MMSE score were in agreement. For example, according to the MMSE scale, a score between 24 and 30 is associated to no dementia (Figure 2), and all images labeled as NC with MMSE score in that range are considered easy samples. As depicted in Figure 2, in the first round only the easy samples of AD and NC (according to the MMSE) were included, to guarantee that the discriminative features of the

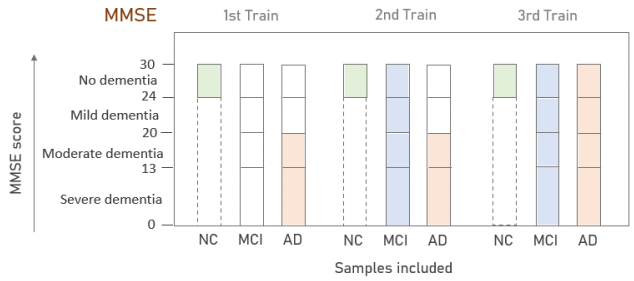

Fig. 2. Manually defined curriculum based on MMSE. The NC, MCI and AD samples included in each train are represented in green, blue and orange, respectively, and their MMSE scores are presented in the vertical axis.

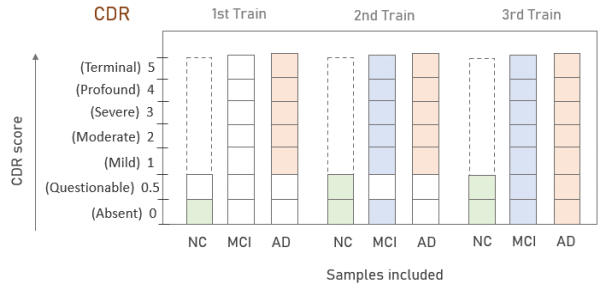

Fig. 3. Manually defined curriculum based on MMSE. The NC, MCI and AD samples included in each train are represented in green, blue and orange, respectively, and their CDR scores are presented in the vertical axis.

AD and NC concepts are well learnt, without noisy information. Then, in the second round, the MCI samples are added to the training data. In the last round, the AD hard samples are added to the training data, which now comprises all training samples.

CDR strategy: CDR test scores were used to manually build the curriculum schematized in Figure 3, also based on increasing complexity of the samples and increasing complexity of the tasks.

ROI focused strategy: This manual strategy focus on progressively adding to the training set more complex regions of the images. The model was first trained with the dataset images multiplied by a ROI mask (1 inside the ROI, 0 outside), then it was retrained (fine-tuned) using the complete images. Two ROIs were used, one containing all ROIs for AD (All ROIs) and another corresponding to the gyrus, cingulate and precuneus, which match the most discriminative regions for AD (ROI 5+9+10) [28].

Mix 1 strategy (based on CDR and ROI): It follows the strategy described in Figure 3, only that the first two rounds of training use the samples multiplied by the ROI mask, while the last train and test are performed using the complete images (without multiplying them by the ROI mask).

Mix 2 strategy (based on MMSE and ROI): Similarly to the strategy described in figure 2, only the first two rounds of training use the images multiplied by the ROI mask and the last train and test are performed using the complete images.

Mix 3 strategy (based on CDR, MMSE and ROI) In this strategy the model goes though two training stages: first it is fed with samples that are considered easy according to both MMSE and CDR scores, multiplied by the ROI mask. In the second train, the model trains with all (both easy and hard) complete brain scans.

Replicate strategy: In this strategy we manually built a curriculum equivalent to the one automatically generated in automatic CL strategies. It corresponds to three rounds of training, where in the first round, the model is fed only with AD labeled samples, adding the NC labeled samples in the second round and finishing in the third round by feeding the complete dataset into the model (AD, NC and MCI samples).

3.3.2. Automatic

Self-paced Learning: Self-paced learning (SPL) is an automatic CL strategy where data are sorted while training, based on sample training loss [29]. A threshold, λ, is defined and the samples with loss below (above) λ are considered easy (hard). During training the threshold is updated, according to a growing factor (= 1.5), from including only the lower loss samples, to including all samples in the final epochs. This strategy does not take prior medical knowledge into account.

***Self-paced Curriculum Learning* (SPCL)**: SPCL results from the merge of manual CL with SPL, taking into account both prior knowledge and the learning progress of the model during training. In this strategy, the predetermined curriculum, where prior knowledge is encoded, is given as input and updated at each epoch.

In this paper, a SPCL algorithm was implemented (Algorithm 1), inspired in the implementation performed in [30], yet adapted to the current classification problem.

In Algorithm 1, $training\_samples$ contains the samples the model should train with, at each epoch. Moreover, γ consists on the curriculum, which is updated during training through element wise multiplication ($\odot$) with the $losses$ vector. N represents the total number of training samples and E represents the total number of epochs.

The predefined curriculum, γ, and the growing function, $\lambda(t)$, given as input, were defined according to:

Algorithm 1 Self-paced curriculum learning algorithm

$training\_samples = [s_1, s_2, ..., s_N]$
$\gamma = [\gamma_{s_1}, \gamma_{s_2}, ..., \gamma s_{s_N}]$ ▷ Predetermined curriculum
$\lambda(t)$ ▷ Growing function
for t in $[0, E]$ **do**:
    Train the model using $training\_samples$
    $losses = [ls_1, ls_2, ..., l_N]$ ▷ Normalized loss
    $\gamma = \gamma \odot losses$ ▷ Update curriculum
    $threshold = \lambda(t)$ ▷ Update threshold
    $updated = [\,]$
    foreach $x \in [0, ..., N]$ **do**:
        if $\gamma_{s_x} <= threshold$ **then**:
            $updated = updated + [s_x]$
        end if
    end for
    $training\_samples = updated$
end for

- γ is an array with values in [0,1], as the normalized $losses$ vector, where each instance γ_{si}, corresponds to the weight of each training sample, s_i. The easier samples have lower γ_{si} values, since they are the ones that should be learnt first in the training process. Two SPCL strategies were implemented, differing only on the predetermined curriculum. In SPCL 1, each entry of the predefined curriculum vector was defined as: $\gamma_{s_i} = 0.33$ if s_i is an easy AD or NC sample; $\gamma_{s_i} = 0.66$ if s_i is a MCI sample or $\gamma_{s_i} = 0.99$ if s_i is a hard AD sample. This follows the same curriculum used in the manual CL strategy described in Figure 2: first training with easy AD and NC samples, then MCI samples are added and afterwards hard AD samples are also added. In the other strategy, SPCL 2, the predetermined curriculum follows the curriculum of the task strategy and γ is defined as: $\gamma_{s_i} = 0.33$ if s_i is an AD or NC sample and $\gamma_{s_i} = 0.99$ if s_i is a MCI sample.
- The growing function, $\lambda(t)$, dictates how the threshold grows. Similarly to [31], $\lambda(t)$ was defined so training would start with only 2% of samples at the first iteration, t=0, and then exponentially increase to include all samples in 3/4 of the maximum epoch, in epoch t=75.

3.4. Baseline methods

The CL strategies were compared to two baseline methods, Simple model and Focal loss. Although none of the baseline methods use curriculum learning, the focal loss model takes into account the model's feedback and Sample weights model takes into account medical knowledge prior to training, such as MMSE scores.

Simple model: The same CNN architecture trained without CL, i.e. the entire dataset was presented to the network at every training epoch.

Focal loss: In this method the model was trained like the Simple model, but the loss function used was a balanced focal loss (FL) function. The FL function was introduced to deal with class imbalance and is described by equation 1 [32]:

$$FL(y, \hat{p}_y) = -\alpha(1 - \hat{p}_y)^{\delta} * log(\hat{p}_y) \quad (1)$$

where $y = [0, ..., K-1]$ is an integer class label (K denotes the number of classes), $\hat{p}_y = [\hat{p}_0, ..., \hat{p}_{K-1}]$ is a vector representing an estimated probability distribution over the K classes and α represents the balance factor. FL, according to δ, smoothly adjusts the rate at which easy examples (correctly classified) are down weighted. In our implementation we used $\alpha = 0.25$ and $\delta = 2$.

Sample weights: In the Sample weights (SW) strategy the model was trained in the same way as the Simple model, only rather than using class weights, sample weights were implemented. Each sample was associated with a specific weight during training. This weight specifies how much influence each sample in a batch should have, in the computation of the total loss. Here, easier samples are associated with higher weights in the beginning so that they are given more relevance. This corresponds to being the first ones to be added to train in the SPCL strategies. Then, as we evolve through the epochs, their weight decreases. Contrarily, harder samples are associated with lower weights in the first epochs, which increase as the model progresses through the epochs. Similarly to SPCL 1, the samples were divided into 3 groups: A which comprises easy AD and NC samples, B which comprises MCI samples and C with comprises hard AD samples. The weight value, $weight_{s_i}$, of each sample, s_i, is defined according to table 3, where $i \in [1, N]$ and N corresponds to the total number of training samples.

Table 3. Value of $weight_{s_i}$ with respect to s_i and the epoch number (t).

$\boldsymbol{weight_{s_i}}$		s_i		
		$s_i \in A$	$s_i \in B$	$s_i \in C$
Epoch number (t)	t < 30	1.33	1	0.77
	30 < t < 60	1	1.33	0.77
	t > 60	0.77	1	1.33

4. RESULTS

The overall results are presented in Figures 4 and 5. The results per class, for all methods implemented, are presented in Figure 6. They show that the use of the CL strategies improve the overall accuracy and F1-score of the classifications, up to 4.5% and 4.3%, respectively, when comparing

the Simple model and the best CL strategy, SPCL 1. Regarding the baseline models, the Simple model presents the poorest overall accuracy, F1-score and MCI accuracy. It can be considered the least suitable for the selected dataset and for early AD diagnosis. The results show that taking the model's feedback into account (Focal loss) or incorporating medical knowledge into the models (Sample weights) is advantageous for improving the overall and MCI accuracy, being that the later had a higher contribution for such improvements. However, in the Sample weights strategy, the improvement of MCI accuracy is achieved at the cost of AD and NC accuracy, which suffer a considerable decrease (See Figure 6).

Regarding the manual strategies, the Replicate is the one that presents highest F1-score and MCI accuracy. However, the ROI strategy using ROI 5+9+10 is the one with highest overall accuracy, directly followed by the Replicate strategy and the Task strategy. On the one hand, comparing the strategies that incorporate the scores of the cognitive tests in the process of building the curriculum, the MMSE has proven to be the best regarding both overall and MCI accuracy. On the other hand, comparing the strategies that incorporate ROI information, the use of ROI 5+9+10 has shown to be advantageous, improving the overall accuracy by 1%, when compared to All ROI. Regarding the mixed strategies, we can see that there is no advantage in combining both information of cognitive tests and ROI to build the curriculum, since the results were the poorest out of all manual curriculum learning strategies.

Regarding the automatic strategies, we can observe that SPCL 1 yields the best results. SPCL strategies, in comparison with SPL, require the extra work of building the curriculum. Nevertheless, they complement SPL and prove that incorporating medical knowledge prior to training into the models brings an added advantage. SPCL 1 improves the overall accuracy of SPL by 1.3%, achieving the best overall performance. Comparing the two SPCL strategies, the first one shows higher overall and MCI accuracy and F1-score. This shows that combining information about the MMSE scores and task complexity (SPCL 1 proceeding) for building the training curriculum is more advantageous than using only information about task complexity (SPCL 2 proceeding).

Comparing the manual and automatic strategies, by analysing Figure 6, it can be verified that the manual ones, despite achieving higher MCI accuracy, they also have higher uncertainty and higher discrepancy between MCI and AD/NC accuracies, making the automatic ones the most robust. Regarding Figures 4 and 5, the SPCL 1 strategy (automatic) has yielded the best results in terms of overall accuracy and F1-score, followed by five strategies, four of them manual, Task, ROI 5+19+10, Replicate and MMSE, and one automatic, SPCL 2.

The incorporation of the model's feedback and medical knowledge into the models has shown to be effective to improve overall and MCI accuracy. This is true for baseline methods (Focal loss and Sample weights) as well for all curriculum learning strategies applied. However, the later show less discrepancy between MCI and AD/NC accuracies, achieving similar MCI accuracy to the Sample weights model, but always better accuracy for AD and NC (See Figure 6). Thereby, CL strategies can be considered superior to all baseline methods, even those which incorporate extra information. Moreover, the incorporation of medical knowledge into the process of building the curriculum have proven to be advantageous for early AD diagnosis, since all strategies that incorporate it yield better MCI accuracy results than SPL, Focal loss and Simple model, which do not take it into account (Figure 6).

To assess the statistical significance of the results obtained, the Wilcoxon signed-rank test was used. Table 4 summarized the p-values of the statistical tests performed between the results of the curriculum learning strategies and each of the baselines methods. The p-values were lower than the threshold (0.05) for four out of the six best curriculum strategies, such as Task, MMSE, Replicate and SPCL 1. This indicates that the differences between their results and those of the baseline methods are statistically relevant. Moreover, regarding the SPCL 1 results, they show statistical relevance when compared to the results of the Simple model and Focal loss, but not when compared to the Sample weights model. This can be explained by the fact that these two models use equivalent curricula. In general, it was verified that the difference between the results of curriculum learning and baseline methods are statistically relevant, which contributes for the robustness of CL strategies.

Additionally, Table 5 presents the p-values between the results of comparable curriculum learning strategies, whether by the fact that they use equivalent curricula or whether to compare the effect of incorporating medical knowledge *vs* not incorporating it. In order to further compare the results of automatic strategies, the p-values between those that do not incorporate medical knowledge (SPL) and those that do (SPCL), were obtained. It was verified that there is no statistically relevance between these results, despite the fact that their accuracy and F1-scores differ up to 1.3% and 1.2%, respectively. Additionally, to compare the results of Task strategy, which only incorporates the knowledge of task complexity, and MMSE and CDR, which incorporate both knowledge of task complexity and the cognitive test results, the p-values between their results were obtained. They reflect that there is a significant difference in the results. Moreover, to fairly compare automatic strategies with manual ones, we obtained the p-values between the results of these two types of strategies using equivalent curricula: SPL *vs* Replicate, SPCL 1 *vs* MMSE and SPCL 2 *vs* Task. The p-values are bellow the threshold in all these three cases, allowing us to conclude

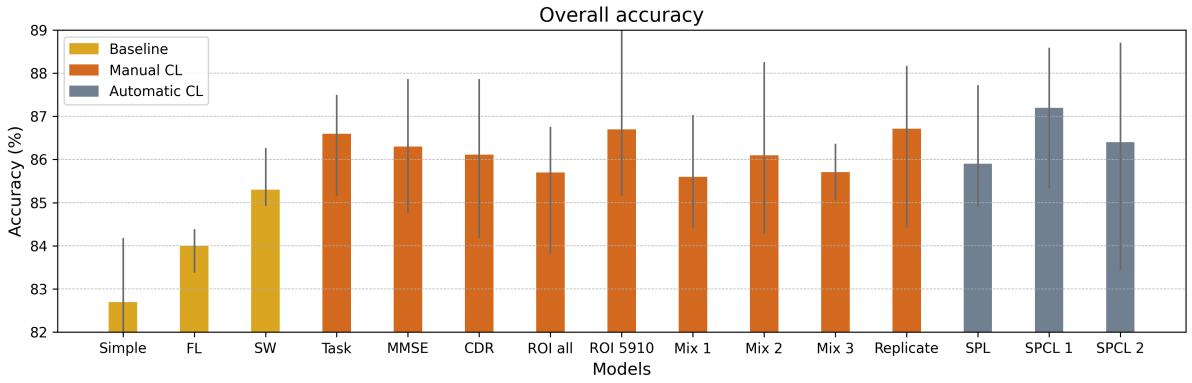

Fig. 4. Overall accuracy results for all strategies implemented with error lines indicating the variability of data (minimum and maximum value). FL: Focal loss; SW: Sample weights.

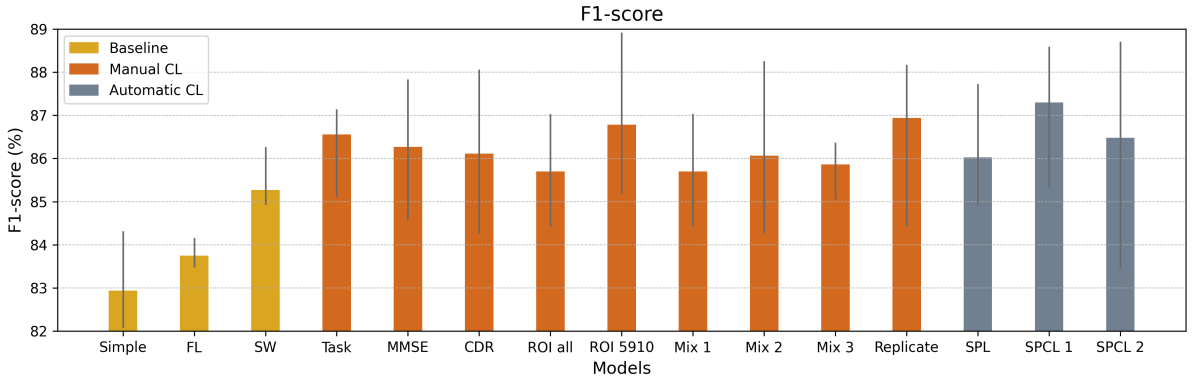

Fig. 5. F1-score results for all strategies implemented with error lines indicating the variability of data (minimum and maximum value).

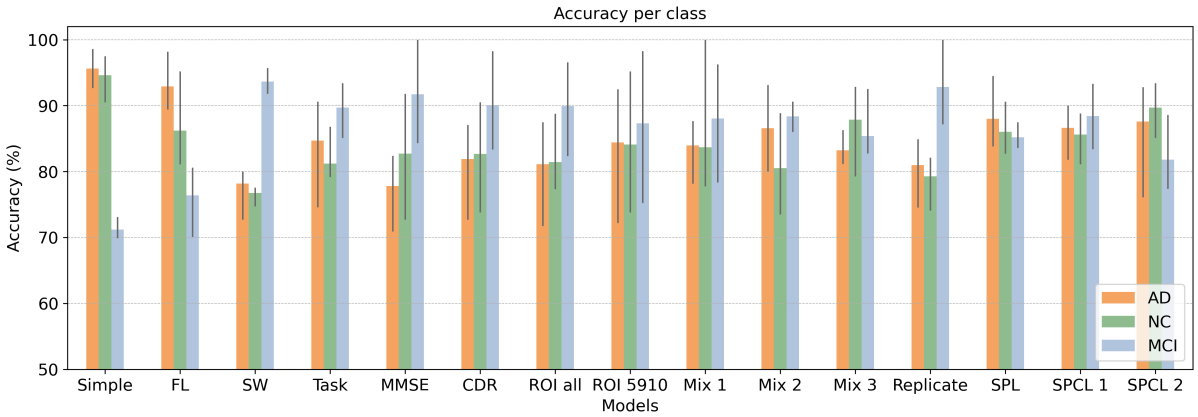

Fig. 6. Bar plots representing the accuracy per class (AD, NC and MCI), for all the implemented models with error lines indicating the variability of data (minimum and maximum value).

that the differences in performance between the results of manual and automatic strategies are statistically relevant.

Table 4. P-value between the predictions of the baseline methods and curriculum learning strategies. The p-values below the threshold (0.05) are highlighted in gray.

p-value	Baseline methods		
	Simple model	Focal loss	Sample weights
Task	**0.006**	**0.011**	**0.020**
MMSE	**0.007**	**0.003**	**1.56e−10**
CDR	0.098	0.073	**0.012**
All ROI	0.061	0.062	0.065
ROI 5+9+10	0.061	0.067	0.065
MMSE+ROI	**0.024**	**0.025**	0.064
CDR+ROI	**0.046**	**0.037**	**0.027**
MMSE+CDR+ROI	**0.026**	**0.005**	**0.003**
Replicate	**0.004**	**0.019**	**0.014**
SPL	0.052	0.065	0.061
SPCL 1	**0.038**	**0.028**	0.075
SPCL 2	0.056	0.052	0.05

Table 5. P-value between the predictions of curriculum learning strategies. The p-values below the threshold (0.05) are highlighted in gray.

p-value	SPL	Task	SPCL 1	SPCL 2
SPCL 1	0.087	—	—	—
SPCL 2	0.12	—	—	—
Replicate	**0.009**	—	—	—
Task	—	—	—	**7.33e−5**
MMSE	—	**5.10e−11**	**2.41e−6**	—
CDR	—	**3.04e−4**	—	—

5. CONCLUSION AND FUTURE WORK

This thesis was, as far as we know, the first work investigating the use of curriculum learning for early AD diagnosis from neuroimaging. Twelve different CL strategies, nine manual and three automatic, incorporating different kinds of medical knowledge were implemented. The knowledge could be in the form of task complexity (Task and SPCL 2), ROI information (ROI all and ROI 5+9+10) or a combination of different types of information, such as mixing cognitive test scores with task complexity (MMSE, CDR, SPCL 1) or with ROI information (Mix 1, Mix 2 and Mix 3). Moreover, one automatic strategy (SPL) and one manual (Replicate) were also defined, for comparison purposes, since none of them incorporated any kind of medical knowledge.

The results show that all the proposed CL strategies improve both overall and MCI classification (early AD) performances. SPCL 1 has obtained the highest overall accuracy and F1-score, making the automatic strategies the preferred ones. The incorporation of medical information (Task complexity information, ROI information and MMSE/CDR scores) in the CL strategies has proven to be advantageous in all cases, improving the overall accuracy, F1-score and MCI accuracy.

The results obtained in this paper show that the order in which data is fed into the CNNs, for early AD diagnosis, is meaningful. That said, CL strategies incorporating medical knowledge when building the curriculum allow for a better earlier AD diagnosis, which can contribute to the ongoing search for treatments to delay the onset or prevent this devastating disease. Even though the results obtained were distinctly positive, there is still a lot of room for improvements. For example, other types of external information, such as medical imaging reports or evidence maps obtained during training, could be used for developing different curricula for CL strategies. Moreover, to make a more accurate early AD prediction, these strategies could be applied to a dataset that allows for MCIc *vs* MCInc distinction, allowing to distinguish early AD from other unrelated dementia cases. These strategies could also be applied to other type of input images, different from PET, such as MRI or others, or yet adapted to the diagnosis of other neurogenerative disorders, like Parkinson's or Huntington's disease.

6. ACKNOWLEDGMENTS

I would like to thank Professor Margarida Silveira, for the guidance and motivation provided during this thesis, and my family and friends for their love and support.

7. REFERENCES

[1] C. L. Masters, R. Bateman, K. Blennow, C. c. Rowe, R. A. Sperling, and J. L. Cummings, "Alzheimer's disease," *Nature Reviews Disease Primers*, vol. 1, pp. 15056, 2015.

[2] M.A. Myszczynska, P.N. Ojamies, A.M.B Lacoste, D. Neil, A. Saffari, R. Mead, G.M Hautbergue, J. Holbrook, and L. Ferraiuolo, "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," *Nature Reviews Neurology*, vol. 16, pp. 440–456, 2020.

[3] A. Ebrahimighahnavieh, , S. Luo, and R. Chiong, "Deep learning to detect Alzheimer's Disease from neuroimaging: A systematic literature review," *Computer Methods and Programs in Biomedicine*, vol. 187, 2020.

[4] Alzheimer's Association, "2008 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 4, no. 2, pp. 110–133, 2008.

[5] L. Zhang, M. Wang, M. Liu, and D. Zhang, "A survey on deep learning for neuroimaging-based brain disorder analysis," *Frontiers in neuroscience*, vol. 14, 2020.

[6] M. Weiner and Z. Khachaturian, "The use of MRI and PET for clinical diagnosis of dementia and investigation of cognitive impairment: a consensus report," *Alzheimer's Assoc Chicago, IL*, vol. 1, pp. 1–15, 2005.

[7] Tapan Kammar Khan, *Biomarkers in Alzheimer's Disease*, Academic Press, 2017.

[8] A. Chopra, T. Cavalieri, and D. J. Libon, "Dementia screening tools for the primary care physician," *Clinical Geriatrics*, vol. 15, no. 1, pp. 38, 2007.

[9] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," *Classification in BioApps*, pp. 323–350, 2018.

[10] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[11] S. Basaia, F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo, M. Filippi, ADNI, et al., "Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks," *NeuroImage: Clinical*, vol. 21, pp. 101645, 2019.

[12] S. Spasov, L. Passamonti, A. Duggento, P. Lio, N. Toschi, ADNI, et al., "A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease," *Neuroimage*, vol. 189, pp. 276–287, 2019.

[13] K. Bäckström, M. Nazari, I. Y. Gu, and A. S. Jakola, "An efficient 3D deep convolutional network for Alzheimer's disease diagnosis using MR images," *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 149–153, 2018.

[14] H. Karasawa, C. Liu, and H. Ohwada, "Deep 3d convolutional neural network architectures for Alzheimer's disease diagnosis," *Asian conference on intelligent information and database systems*, pp. 287–296, 2018.

[15] Danni Cheng and Manhua Liu, "Classification of Alzheimer's disease by cascaded convolutional neural networks using PET images," *International Workshop on Machine Learning in Medical Imaging*, pp. 106–113, 2017.

[16] Arwa Mohammed Taqi, Ahmed Awad, Fadwa Al-Azzo, and Mariofanna Milanova, "The impact of multi-optimizers and data augmentation on tensorflow convolutional neural network performance," *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 140–145, 2018.

[17] Jianping Qiao, Yingru Lv, Chongfeng Cao, Zhishun Wang, and Anning Li, "Multivariate deep learning classification of Alzheimer's disease based on hierarchical partner matching independent component analysis," *Frontiers in aging neuroscience*, vol. 10, pp. 417, 2018.

[18] Yosra Kazemi and Sheridan Houghten, "A deep learning pipeline to classify different stages of Alzheimer's disease from fMRI data," *2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–8, 2018.

[19] Danni Cheng and Manhua Liu, "Combining convolutional and recurrent neural networks for Alzheimer's disease diagnosis using PET images," *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–5, 2017.

[20] Manhua Liu, Danni Cheng, Weiwu Yan, Alzheimer's Disease Neuroimaging Initiative, et al., "Classification of Alzheimer's disease by combination of convolutional and recurrent neural networks using FDG-PET images," *Frontiers in neuroinformatics*, vol. 12, pp. 35, 2018.

[21] Torgyn Shaikhina and Natalia A Khovanova, "Handling limited datasets with neural networks in medical applications: A small-data approach," *Artificial intelligence in medicine*, vol. 75, pp. 51–63, 2017.

[22] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu, "A survey on incorporating domain knowledge into deep learning for medical image analysis," *Medical Image Analysis*, 2021.

[23] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48, 2009.

[24] Y. Tang, X. Wang, A.P. Harrison, L. Lu, J. Xiao, and R.M. Summers, "Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs," *International Workshop on Machine Learning in Medical Imaging*, pp. 249–258, 2018.

[25] C. Haarburger, M. Baumgartner, D. Truhn, M. Broeckmann, H. Schneider, S. Schrading, C. Kuhl, and D. Merhof, "Multi scale curriculum CNN for context-aware breast MRI malignancy classification," *Medical Image Computing and Computer Assisted Intervention*, p. 495–503, 2019.

[26] G. Maicas, A. P. Bradley, J. C. Nascimento, I. Reid, and G. Carneiro, "Training medical image analysis systems like radiologists," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 546–554, 2018.

[27] M. Grassi, D. A. Loewenstein, D. Caldirola, K. Schruers, R. Duara, and G. Perna, "A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion: further evidence of its accuracy via a transfer learning approach," *International psychogeriatrics*, vol. 31, no. 7, pp. 937–945, 2019.

[28] J. Rondina, L. Ferreira, F. de Souza Duran, R. Kubo, C. R. Ono, C. C. Leite, J. Smid, R. Nitrini, C. A. Buchpiguel, and G. F. Busatto, "Selecting the most relevant brain regions to discriminate Alzheimer's disease patients from healthy controls using multiple kernel learning: A comparison across functional and structural imaging modalities and atlases," *NeuroImage: Clinical*, vol. 17, pp. 628–641, 2018.

[29] M. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," *Advances in neural information processing systems*, vol. 23, pp. 1189–1197, 2010.

[30] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, "Self-paced curriculum learning," *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[31] K. Ghasedi, X. Wang, C. Deng, and H. Huang, "Balanced self-paced learning for generative adversarial clustering network," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4391–4400, 2019.

[32] T. Lin, P. Goyal, R. Girshick, H. He, and P. Dollár, "Focal loss for dense object detection," *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.