

Investigating the Effect of the Data Partitioning Protocol on Automatic Alzheimer's Disease Detection from Brain MRI Using Deep Learning

Fatima Zahra Salmam*, Sanae Belfrouh, Lamya Ouahmidi, Rahhal Errattahi
Laboratory of Information Technologies (LTI), National School of Applied Sciences,
Chouaib Doukkali University, El Jadida, Morocco

Abstract—Computer vision and deep learning have been widely used to automatically detect and analyze many diseases from medical images. Some of these include tumor detection, diabetic retinopathy classification, nodules classification, and the screening of thoracic pathologies. In this work, we are investigating the application of deep learning techniques to the detection of Alzheimer's disease from brain MRI images. The general purpose was to offer an aided diagnosis system able to distinguish the three diagnostic groups of the ADNI database, namely Alzheimer's disease, cognitively normal subjects and subjects with mild cognitive impairment. An Inception-V3 model pre-trained on ImageNet was adapted to the task and fine-tuned by gradually unfreezing its top convolutional blocks. To address the problem of the imbalance between the diagnostic classes, and given the limited size of the available cohort, we compared three strategies: class weights, data augmentation, and a subject-level version of SMOTE. Then, given the nature of the task, we opted for a double evaluation of the developed models, first with a partition performed at the subject level and then with a partition performed at the slice level. On a cohort of 106 ADNI subjects, the best three-class accuracy under a strict subject-level partition was 67.77% (SMOTE), whereas the same models reached more than 90% with a slice-level partition. The difference between imbalance strategies (up to 13 points) vanished under leakage (less than 0.5 point), showing that, in this setting, the partitioning protocol had a larger influence on the reported accuracy than the strategy used.

Keywords—Alzheimer's disease; brain MRI; ADNI; deep learning; transfer learning; data leakage; class imbalance

I. INTRODUCTION

Neurodegenerative diseases are pathologies in which the neurons of the central nervous system deteriorate progressively and irreversibly. They mainly affect elderly people, and their prevalence increases as the world population ages, which makes them a public health problem of growing importance [1]. Several mechanisms are involved and reinforce each other, among them the abnormal aggregation of misfolded proteins, the failure of the energy metabolism of nerve cells, chronic inflammation of the brain tissue and progressive neuronal death [2].

Alzheimer's disease (AD) is the most common form of dementia. It starts with difficulties in memorising recent events and then affects reasoning, language and executive functions, until the patient becomes dependent for daily activities. At

the level of the brain tissue, the disease is associated with two types of lesion: extracellular amyloid plaques, produced by the accumulation of the beta-amyloid peptide between neurons, and intracellular neurofibrillary tangles, made of abnormally phosphorylated tau proteins [3]. The drugs available today reduce some symptoms, but none of them stops the neurodegeneration. Early identification therefore remains the main lever for patient care.

This is also where the main difficulty lies. Clinical symptoms appear only after a large and irreversible neuronal loss, and the disease can progress silently for fifteen to twenty years before the patient consults for memory complaints. For this reason, research has moved towards biomarkers that can be measured before the symptoms appear [4]. Medical imaging provides several of them. Magnetic resonance imaging (MRI), functional MRI and positron emission tomography (PET) each give a different type of information [5]. Structural T1-weighted MRI is the most widely used, because it makes cerebral atrophy visible, and because the hippocampus, which is the region affected earliest in AD, can be measured on it reliably [6]. PET detects amyloid deposits more than a decade before the first clinical signs [7], and combining modalities gives more robust diagnostic models [8], but PET is expensive and not available in routine clinical practice. This explains why most automatic classification studies start from T1-weighted MRI alone.

Computer vision and deep learning have been applied to many medical imaging tasks, including tumour detection, nodule classification, retinopathy grading and the screening of thoracic diseases [9]. In the case of Alzheimer's disease, convolutional neural networks (CNN) learn the discriminative features directly from the images, which avoids defining in advance the regions or the measurements to analyse [10]. Many studies report accuracies higher than 96% for the AD versus CN task on the ADNI database.

These figures have to be interpreted carefully. Wen et al. [11] have shown that more than half of the published articles in this field contain data leakage, most often because images of the same patient are present in the training set and in the test set at the same time. This happens when the partition is performed at the slice or image level instead of the patient level. In that situation, the network can recognise a patient it has already seen, and the reported accuracy reflects the individual anatomy of the subjects as much as the disease.

*Corresponding author.

The consequence is a literature made of two groups of results that cannot be compared with each other.

In this work we investigate the application of transfer learning to the automatic classification of the three diagnostic groups of ADNI, and we measure the effect of the partitioning protocol on the reported performance. Our contributions are the following. We have built a complete classification pipeline for the AD/CN/MCI task from ADNI pre-processed T1-weighted volumes, with a partition performed at the subject level before any slice extraction. We have compared three strategies to deal with the imbalance between the diagnostic classes, using the same backbone, the same selection procedure and the same test set. Finally, we have repeated the three experiments with a slice-level partition, in order to quantify, on the same data and the same models, the gain that data leakage produces.

The remaining part of the paper is organised as follows. Background and related work are presented in Section II, materials and methods in Section III, results and discussion in Section IV, and a brief summary is given in the last section.

II. BACKGROUND AND RELATED WORK

A. Public Datasets

Research on the automatic detection of Alzheimer's disease relies on a small number of public databases. Medical data are difficult to collect, acquisition protocols differ from one hospital to another, and the clinical annotation requires long neuropsychological assessments. Two databases dominate the literature, ADNI (Alzheimer's Disease Neuroimaging Initiative) and OASIS (Open Access Series of Imaging Studies) [12].

ADNI is a multicentre North American project, active since 2004, which gathers 63 clinical sites in the United States and Canada. Its acquisitions are standardised across the sites and the data are made available without embargo to qualified researchers. The subjects are divided into three clinical groups: cognitively normal (CN), mild cognitive impairment (MCI) and Alzheimer's disease (AD). The MCI group can be divided further into early and late MCI, but the three-class CN/MCI/AD setting remains the standard framework [12], [13]. The collection was carried out in successive phases. ADNI-1 (2004–2010) is the founding phase, with acquisitions at 1.5 tesla. ADNI-GO and ADNI-2 (2009–2016) extended the cohort to the earliest stages of MCI and introduced acquisitions at 3 tesla. ADNI-3 (2016–2022) added tau PET and higher resolution MRI, and ADNI-4 has been running since 2022 with more diverse cohorts. The MRI protocol uses the MP-RAGE sequence, a three-dimensional T1-weighted acquisition that produces isotropic volumes with a spatial resolution of about 1 mm³. Since the acquisition sites operate scanners of different brands and generations, the raw images contain technical variability that has no diagnostic value. ADNI therefore also distributes pre-processed images, which have already been corrected for gradient non-linearities, for the B1 field and for intensity scaling. For deep learning applications, these pre-processed images are the usual starting point [14], [15].

OASIS is a project of Washington University in St. Louis, which provides neuroimaging data without access restriction and without any particular accreditation. It exists in three

versions. OASIS-1, released in 2007, is cross-sectional and gathers 416 subjects aged 18 to 96, each of them with three or four T1-weighted images, among whom 100 subjects over 60 received a clinical diagnosis of Alzheimer's disease ranging from very mild to moderate [16]. OASIS-2, released in 2010, is longitudinal and contains 150 subjects aged 60 to 96, some of whom developed a dementia during the follow-up period [13]. OASIS-3, described by LaMontagne et al. [17], covers 1378 participants aged 42 to 95, among them 755 cognitively normal subjects and 622 subjects at different stages of cognitive decline, with 2842 MRI sessions including several sequences and more than 1500 PET sessions. Depending on the study, OASIS is used for training jointly with ADNI or as a separate cohort [23], [22].

Two other databases appear in the literature. NACC, active since 1999, exceeds 40,000 subjects, but its acquisition protocols are not standardised between the participating centres [18]. AIBL, conducted in Australia since 2006, gathers more than 1000 participants and covers a population that is geographically and ethnically different from the North American cohorts [9].

B. Slice-Based Approaches

Most studies do not use the whole three-dimensional volume, but a set of two-dimensional slices extracted from it. The choice of the plane has been studied directly. Ramalho et al. [19] have compared the axial, coronal and sagittal planes on ADNI-1 and have found that the axial plane gives the best results, with an accuracy of 97.4% for the AD versus CN task. Mohsin and Abdulameer [15] report the same value in the axial plane and 89.9% in the coronal plane for MCI versus CN. This preference has an anatomical explanation, since the hippocampus, the entorhinal cortex and the medial temporal cortex have an orientation that axial sections capture well. The coronal plane is preferred when the medial temporal lobe is targeted specifically, as in the work of Bae et al. [20], who used 20 to 30 coronal slices and obtained an AUC of 0.94.

The number of slices per subject varies considerably in the literature, from one central slice in Shahid et al. [13] to 120 slices distributed over three planes in Turanli [21]. Carcagni et al. [16] have carried out the most systematic study on this question, testing from 1 to 128 axial slices on ADNI-2. Their results show that the performance improves quickly up to 32 slices, and then stops improving. Sener et al. [14] used 54 slices centred on the hippocampal region and obtained 98.70% with Inception-V3, which indicates that the gains beyond 32 slices exist but remain small when the slices are well chosen.

Transfer learning is used in almost all of these works, because medical datasets are too small to train a deep network from scratch. The usual approach is to start from a network pre-trained on ImageNet, to replace its classification part by one adapted to the number of target classes, and to fine-tune the top convolutional layers on the medical images [9].

C. Evaluation Protocols and Data Leakage

The literature is clearly divided according to the level at which the data are partitioned. Wen et al. [11] have carried out the first systematic analysis of data leakage in this field. With a strict subject-level partition, the accuracies obtained with CNN

TABLE I. OVERVIEW OF ALZHEIMER'S DISEASE CLASSIFICATION WORKS ON ADNI, WITH THE LEVEL OF THE TRAIN/TEST PARTITION * PROBABLY USE IMAGE-LEVEL PARTITIONING, ALTHOUGH NOT STATED

Reference	Architecture	Dataset	Task	Partition level	Accuracy	AUC
Wen et al. [11]	Standard CNN	ADNI (about 400 subj.)	Multi-class	Subject	55–65%	–
Bron et al. [12]	SVM + CNN	ADNI (854 subj.)	AD/CN	Subject	–	0.940
Bron et al. [12]	SVM + CNN	ADNI (854 subj.)	MCI conversion	Subject	–	0.665
Liu et al. [18]	3D CNN	ADNI (1000)	CN vs. rest	Subject	–	0.851
Carcagni et al. [16]	DenseNet201	ADNI	Multi-class	Subject	69%	–
Ashwini et al. [30]	DenseNet169	ADNI	4 grades	Subject	–	0.82
AlSaeed and Omar [25]	ResNet50	ADNI	AD/CN	Slice	99%	–
Sener et al. [14]	Inception-V3	ADNI (627 subj.)	Multi-class	Slice	98.70%	–
Mandal and Mahto [26]	Multi-branch CNN	ADNI	Multi-class	Not specified *	99.05%	–
Abbas et al. [28]	ConvMixer + RF	ADNI	Multi-class	Not specified *	99.5%	–
Alqahtani et al. [29]	DBN + RF	ADNI	Multi-class	Not specified *	97.46%	–
Akan et al. [27]	AlzFormer	ADNI	Multi-class	Not specified *	97%	–

models on the three-class task are between 55 and 65%. Bron et al. [12] obtained an AUC of 0.940 for the binary AD versus CN task with a subject-level partition on 854 subjects, but the same pipeline reached only 0.665 for the prediction of MCI conversion. Liu et al. [18] report an AUC of 0.851 on ADNI with a 3D CNN and a patient-level partition, and Carcagni et al. [16] did exceed 69% on three classes with a DenseNet201 architecture.

The studies that partition the data at the image or slice level report much higher values. AlSaeed and Omar [25] obtained 99% with ResNet50 on the AD versus CN task, Mandal and Mahto [26] report 99.05% on three classes with a multi-branch CNN, Akan et al. [27] 97% with AlzFormer, Abbas et al. [28] 99.5% with ConvMixer combined with a random forest, Alqahtani et al. [29] 97.46% with a deep belief network, and Sener et al. [14] 98.70% with Inception-V3 on ADNI-3.

Table I presents an overview of the most relevant works, with the level of the partition indicated for each of them. It can be seen from the table that the two groups of studies produce results that differ by more than thirty accuracy points on comparable tasks, and that the difference follows the partitioning protocol rather than the architecture. This observation is the starting point of the present work.

D. Pre-Processing Practices

The pre-processing chain reported in the literature aims at removing the technical variability introduced by the acquisition, so that the network learns anatomical features and not the signature of a scanner. The correction of the field inhomogeneity is the step with the most constant benefit. The N3 algorithm and its improved version N4ITK, proposed by Tustison et al. [33], estimate the bias field iteratively and remove it from the volume. Intensity normalisation follows. Min-max normalisation is used with models trained from scratch, while Z-score normalisation is preferred with networks pre-trained on ImageNet, whose internal layers were optimised on normalised distributions [14], [24].

Skull stripping was applied systematically for a long time, but this practice has been questioned recently. Tinauer et al. [31] have shown that removing the skull can induce shortcut learning in Alzheimer's disease classification models, because the shape of the resulting cerebral contour may be correlated

with the diagnostic group and gives the network a signal that is not related to atrophy.

III. MATERIALS AND METHODS

The pipeline developed in this work is organised in four successive steps: preparation of the MRI volumes, extraction of the informative slices, partition of the data at the subject level, and fine-tuning of the Inception-V3 network. Each step is described below.

A. Dataset

The dataset used in this work was extracted from the ADNI database (adni.loni.usc.edu), phases ADNI-1, ADNI-GO and ADNI-2. We selected pre-processed MP-RAGE images, which is the three-dimensional T1-weighted sequence used as a standard in this database. This choice is consistent with the works listed in Table I, which all rely on the same sequence.

The subset we have collected contains 106 subjects distributed in three diagnostic classes: 13 AD subjects, 36 CN subjects and 57 MCI subjects. The distribution of the subjects among the classes and the three subsets is given in Table II. The imbalance between the classes, and in particular the small number of AD subjects, is one of the main constraints of this work. It reflects the composition of the ADNI-1 and ADNI-2 cohorts, in which patients with an established dementia are less numerous than MCI or CN subjects, because the collection protocol was designed to over-represent the early stages of the disease.

TABLE II. DISTRIBUTION OF THE 106 SUBJECTS AMONG THE CLASSES AND THE TRAINING, VALIDATION AND TEST SETS

	Train	Validation	Test	Total
AD	9	2	2	13
CN	25	5	6	36
MCI	40	9	8	57
Total	74	16	16	106

B. Data Pre-Processing

The volumes downloaded from ADNI have already been corrected for the inhomogeneity of the B1 field with the N3

algorithm, for the geometric distortions caused by the non-linearity of the gradients, and for the intensity scale. Three additional operations were applied in this work.

Each volume was first reoriented to the RAS convention with the nibabel library. This step harmonises the axes between scanners and avoids the axis inversions that would corrupt the extraction of the axial slices. A complementary bias correction was then applied with N4ITK [33] through the SimpleITK library, using an Otsu threshold mask, two resolution levels and 20 iterations per level. Although the ADNI images have already been corrected with N3, we observed a residual inhomogeneity on part of the volumes, and N4ITK converges better than N3 on these difficult cases. Finally, each extracted slice was normalised with a Z-score computed over the whole volume. This normalisation is preferable to the min-max normalisation for a network pre-trained on ImageNet, since it reduces the difference between the distribution of the MRI intensities and the distribution on which the pre-trained weights were obtained.

The selected slices were converted into three-channel images, and resized to 299×299 pixels with a bilinear interpolation, which is the input format of Inception-V3.

C. Extraction of the Informative Slices

An MRI volume contains between 160 and 200 axial slices, and a large part of them carries no useful information. The slices located outside the brain contain only noise or air, and the slices at the extremities of the volume show the skull with little or no cerebral tissue. Keeping them would dilute the training signal.

For each volume, the slices were therefore ranked automatically according to the quantity of cerebral tissue they contain and to their entropy, which measures the amount of anatomical detail present in the image. Khan et al. [32] have used the entropy in a similar way to select the most informative ADNI slices before transfer learning. The slices containing less than 1% of tissue, or with a standard deviation below three intensity units, were discarded. The 32 best ranked slices of each subject were kept, following the benchmark of Carcagni et al. [16]. The final dataset contains 3392 slices.

D. Data Partitioning and Leakage Protocol

The partition was performed at the subject level, before any slice extraction. As shown by Wen et al. [11], a partition performed at the slice level artificially increases the accuracy, because the slices of the same patient share the same individual characteristics. The partition was stratified by class, with a fixed seed for reproducibility, and 70% of the subjects were assigned to the training set, 15% to the validation set and 15% to the test set. Since no subject appears in more than one subset, the resulting distribution is the one given in Table II.

The test set was isolated at the beginning of the experiments and used only once, for the final evaluation. No hyperparameter was chosen by looking at it, and the validation set alone was used to select the fine-tuning strategy.

In order to measure the effect of data leakage, we also applied a second and deliberately incorrect partition, performed at the slice level. In that protocol, slices of the same patient

may appear in the training set and in the test set at the same time. This second partition is used only to reproduce experimentally the error documented in the literature, and to measure its effect under otherwise identical conditions.

E. Handling the Class Imbalance

With only 9 AD subjects in the training set, a three-class classifier trained without any correction ignores the AD class completely, since predicting MCI for every input already maximises the accuracy. Three strategies were compared, one per experiment.

In the first experiment, the imbalance is corrected with class weights inversely proportional to the frequency of each class, computed with the balanced mode of scikit-learn. Only the loss function is modified, the data and their distribution remain unchanged, which makes this the least intrusive strategy and a natural baseline.

In the second experiment, data augmentation is applied to the training set only, after the subject-level partition. The transformations were chosen to remain anatomically plausible: translation up to 4% of the image size, zoom between 0.92 and 1.08, and adjustment of the contrast and of the brightness between 0.90 and 1.10. Large rotations and vertical flips were excluded, because an axial brain slice turned upside down does not correspond to any real anatomy. The synthetic copies were generated and stored before training, and each class was brought to the size of the majority class.

In the third experiment, the imbalance is corrected with SMOTE [35], which does not duplicate the minority examples but creates new ones by interpolating between two real examples of the same class. We applied it at the level of the subject. Two real subjects of the same class are selected and their slice stacks are interpolated slice by slice, as given in Eq. (1):

$$x_{\text{synth}} = x_1 + \alpha(x_2 - x_1), \quad \alpha \in [0.25, 0.75] \quad (1)$$

where, x_1 and x_2 are the slices of the two parent subjects and α is the interpolation coefficient. The interval keeps the synthetic subject away from both parents, so that the procedure does not produce near duplicates. As for the augmentation, SMOTE was applied to the training set only, and the validation and test sets contain real subjects exclusively.

F. Model Architecture and Proposed Approach

The classification model used in this work is based on the Inception-V3 architecture [34] pre-trained on the ImageNet dataset. This architecture is organised in two parts. The first one is the convolutional part, which contains a set of convolution and pooling layers followed by ten inception modules, denoted mixed0 to mixed9, in which convolutions of different sizes are applied in parallel and their outputs concatenated. This design captures information at several spatial scales at the same time, which is useful for medical images where the relevant structures do not have the same size. The second part is the classification part, made of fully connected and softmax layers.

To adapt Inception-V3 to our task, we removed the classification part of the model and built a customised one. The customised part is composed of a global average pooling layer, followed by a fully connected layer of 256 units with a ReLU activation, a dropout layer with a probability of 0.5, and a softmax layer that outputs the probability of each class. The model was optimised with AdamW [36] and a categorical cross-entropy loss with a label smoothing of 0.02. The architecture is presented in Fig. 1.

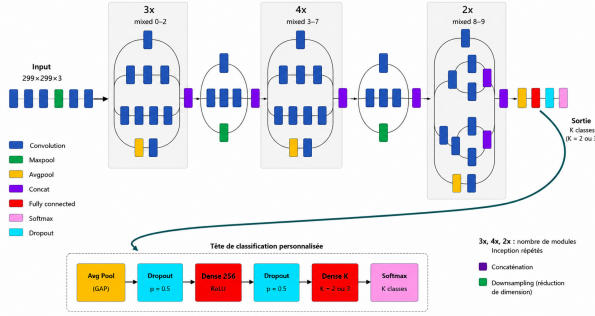


Fig. 1. The customised inception-v3 architecture.

The model was trained following a gradual unfreezing strategy. Our baseline is the pre-trained network used as a fixed feature extractor, where the convolutional part is frozen and only the customised classification part is trained on the new task, for 8 epochs with a learning rate of 3×10^{-4} . Starting from these weights, we then explored the fine-tuning of the model by gradually unfreezing the top convolutional blocks, from mixed9 down to mixed0, and finally the whole network. Each of these configurations was trained with a learning rate reduced to 10^{-5} , which is low enough to preserve the representations learned on ImageNet, and evaluated on the validation set. The configuration giving the best validation accuracy was retained for the final evaluation.

The reason for this progressive approach is that we do not have enough data to fine-tune all the layers of the model. The first layers of a CNN learn generic features such as contours and textures, which are common to many domains, while the top layers learn features that are specific to the training domain. Unfreezing too many layers on 74 subjects makes the model memorise the individual characteristics of the patients, while unfreezing too few layers prevents it from adapting to MRI images, which are very different from natural images.

G. Performance Evaluation

In addition to the classification accuracy, we used the sensitivity, the specificity and the macro F1 score as evaluation metrics. The sensitivity, also called recall, refers to the ability of the model to correctly classify the subjects of a given class c , and is given by Eq. (2):

$$Sensitivity_c = \frac{TP_c}{TP_c + FN_c} \quad (2)$$

The specificity relates to the ability of the model not to assign the class c to subjects that do not belong to it, and is given by Eq. (3):

$$Specificity_c = \frac{TN_c}{TN_c + FP_c} \quad (3)$$

where, FN , FP , TN and TP denote the false negative, false positive, true negative and true positive, respectively. The sensitivity of the AD class is particularly important in our context, since a low value means that patients with Alzheimer's disease are classified as MCI or CN, which is the most costly error in a diagnostic support system.

The F1 score is the harmonic mean of the precision and the sensitivity, and its macro average, given by Eq. (4), gives the same weight to each class whatever its size:

$$F1_{macro} = \frac{1}{K} \sum_{c=1}^K \frac{2 Precision_c Sensitivity_c}{Precision_c + Sensitivity_c} \quad (4)$$

where, K is the number of classes. This metric is better suited than the accuracy to an imbalanced problem such as AD/CN/MCI, in which a minority class must not be hidden by the majority class. The sensitivity and the specificity are also macro-averaged over the classes. We report in addition the confusion matrices, which show which confusions dominate.

All the metrics are computed at the slice level, each slice of a test subject being classified independently, while the partition itself remains at the subject level.

IV. RESULTS AND DISCUSSION

A. Selection of the Training Strategy

The first set of experiments is aimed at investigating the training strategy of the Inception-V3 model, in order to determine which one performs best for each of the three imbalance strategies. Table III gives the validation accuracy obtained for each unfreezing point on the three-class task.

TABLE III. VALIDATION ACCURACY IN % OBTAINED WITH INCEPTION-V3 FOR THE DIFFERENT TRAINING STRATEGIES, FOR THE THREE EXPERIMENTS

Strategy	EXP1 Class weights	EXP2 Augmentation	EXP3 SMOTE
Baseline (head only)	49.80	50.20	55.27
mixed9	50.98	55.86	56.64
mixed8	52.54	55.86	58.01
mixed7	52.73	56.25	56.84
mixed6	56.45	60.16	55.47
mixed5	60.45	60.35	57.42
mixed4	53.13	58.59	54.10
mixed3	53.52	59.57	52.34
mixed2	54.49	58.40	54.69
mixed1	53.71	57.81	52.15
mixed0	51.95	57.81	53.32
All unfrozen	53.52	59.57	53.32

The results show that the baseline model, in which the convolutional part is used as a fixed feature extractor, gives an accuracy close to chance in the first two experiments, 49.80% and 50.20%. The features learned on ImageNet are therefore not sufficient by themselves to separate the AD, CN and MCI volumes. The performance increases when the top

convolutional blocks are unfrozen, and reaches its maximum at mixed5 for the first two experiments, with 56.45% and 60.35%. Beyond that point, unfreezing more blocks does not help, and unfreezing the whole network is not better than unfreezing half of it. The third experiment behaves differently and reaches its best value at mixed8, which unfreezes fewer layers than mixed5. Since SMOTE balances the classes in the training set, the pull towards the majority class is reduced and the high level features of the pre-trained model are reused more efficiently, so that a shallower unfreezing is enough. Unfreezing more blocks then only adds variance on a training set whose size is fixed. The baseline of this experiment, 55.27%, is already higher than the baselines of the other two, which confirms that the effect of the balancing appears before the convolutional part is adapted at all.

B. Classification Results

The retained configurations were then evaluated on the 16 test subjects, which were never used during the development. Table IV gives the accuracy, the macro sensitivity, the macro specificity and the macro F1 obtained for the three binary tasks and for the three-class task.

It can be seen from Table IV that the binary tasks involving the AD class are the easiest ones. In the first experiment, AD versus CN reaches 87.11% and AD versus MCI reaches 84.66%, which is expected since the AD and CN profiles are the most different from each other. The CN versus MCI task is by far the most difficult, with 54.91% and a macro F1 of 0.442, which is barely better than a random assignment. The three-class task reaches 63.48%, with a macro specificity of 79.56% clearly higher than the macro sensitivity of 53.30%. The model is therefore more able to reject a class than to detect it.

The second experiment gives mixed results. The binary tasks improve, and AD versus MCI reaches 92.61% with a macro F1 of 0.868, which is the best binary result of the three experiments. The CN versus MCI task also improves slightly, from 54.91% to 57.37%. The three-class task, on the contrary, decreases to 54.30%, which is nine points below the baseline experiment. The augmentation increases the diversity of the training examples for each class taken separately, but it does not add real anatomical information, and the balance found by the model between the three classes becomes less stable. The analysis of the confusion matrices, presented in the next subsection, shows this effect in detail.

The third experiment gives the best result on the three-class task, with 67.77% of accuracy and a macro F1 of 0.577, as well as the best macro sensitivity and specificity on this task. Its binary results are slightly lower than those of the second experiment, in particular AD versus MCI which decreases from 92.61% to 85.80%. The pseudo-subjects created by interpolation make the boundaries between the classes smoother, which is useful when the three decisions have to be taken at the same time, and slightly harmful on a binary task where one class can legitimately dominate.

C. Analysis of the Confusion Matrices

For further analysis, we plot the confusion matrices of the three experiments on the three-class task in Fig. 2. The test

set contains 512 slices, which correspond to 16 subjects with 32 slices each, distributed in 64 AD slices, 192 CN slices and 256 MCI slices.

In the first experiment, the AD class concentrates most of the errors. Only 17.2% of the AD slices are correctly classified, while 32.8% are predicted as CN and 50.0% as MCI. The model did not learn a discriminative representation for this class, which is difficult to avoid with only 9 AD subjects in the training set. The CN and MCI classes reach 80.2% and 62.5% respectively, with a large confusion between them in both directions.

In the second experiment, the situation changes but does not improve globally. The AD class rises to 56.2%, which is the best value obtained for this class in the three experiments, but the MCI class falls to 41.8%, and 39.5% of its slices are classified as CN. One class gains and another loses more, so the overall accuracy decreases. This is a good illustration of why the accuracy has to be interpreted together with the confusion matrix.

The third experiment gives the most balanced behaviour on the two largest classes. The MCI class reaches 70.7%, its best value in the three experiments, and the CN class 78.6%, with only 2.1% of the CN slices predicted as AD. The AD class improves from 17.2% to 23.4% compared with the first experiment, but it remains the weak point of the model, with half of its slices still classified as MCI. The gain obtained by SMOTE on the three-class task therefore comes from a cleaner separation between CN and MCI, and from a strong reduction of the false AD predictions, rather than from a real improvement of the AD class.

D. Effect of the Data Partitioning Protocol

The next experiment looks at the effect of the partitioning protocol. We re-evaluated the three retained configurations with a partition performed at the slice level, all the rest being identical. Table V compares the two protocols, and Fig. 3 gives the corresponding confusion matrices.

The results confirm the suspicions we had raised in the state of the art section. With a slice-level partition, the three configurations reach an accuracy of about 91%, whatever the quality of the model measured under the correct protocol. In the first experiment, the AD class is recognised perfectly, with 63 slices out of 63, the CN class reaches 95.4% and the MCI class 86.1%. In the second experiment the corresponding values are 98.4%, 84.4% and 93.0%, and in the third one 90.5%, 85.0% and 95.2%. Reported without any indication of the protocol, any of these results would be presented as an excellent performance, and would be in the same range as the works listed in the lower part of Table I.

Two observations can be made on these numbers. The first one is the size of the gain, between 23.39 and 36.47 accuracy points depending on the experiment. The second one, which is less expected, is its direction. The largest gain is obtained by the second experiment, which is the weakest one under the correct protocol with 54.30%, and the smallest gain by the third one, which is the best with 67.77%. The leakage therefore does not measure the quality of the model, but the internal coherence of the subject. The slices of a same patient

TABLE IV. CLASSIFICATION RESULTS OBTAINED ON THE TEST SET WITH A STRICT SUBJECT-LEVEL PARTITION

Experiment	Task	Accuracy (%)	Macro Sensitivity (%)	Macro Specificity (%)	F1 macro
EXP1: class weights, mixed5	AD vs. CN	87.11	83.07	83.07	0.829
	CN vs. MCI	54.91	49.67	49.67	0.442
	AD vs. MCI	84.66	57.81	57.81	0.592
	AD vs. CN vs. MCI	63.48	53.30	79.56	0.528
EXP2: augmentation, mixed5	AD vs. CN	88.28	76.56	76.56	0.811
	CN vs. MCI	57.37	54.23	54.23	0.533
	AD vs. MCI	92.61	84.55	84.55	0.868
	AD vs. CN vs. MCI	54.30	56.12	77.02	0.514
EXP3: SMOTE, mixed8	AD vs. CN	83.20	69.01	69.01	0.722
	CN vs. MCI	53.57	49.61	49.61	0.472
	AD vs. MCI	85.80	63.98	63.98	0.675
	AD vs. CN vs. MCI	67.77	57.60	81.82	0.577

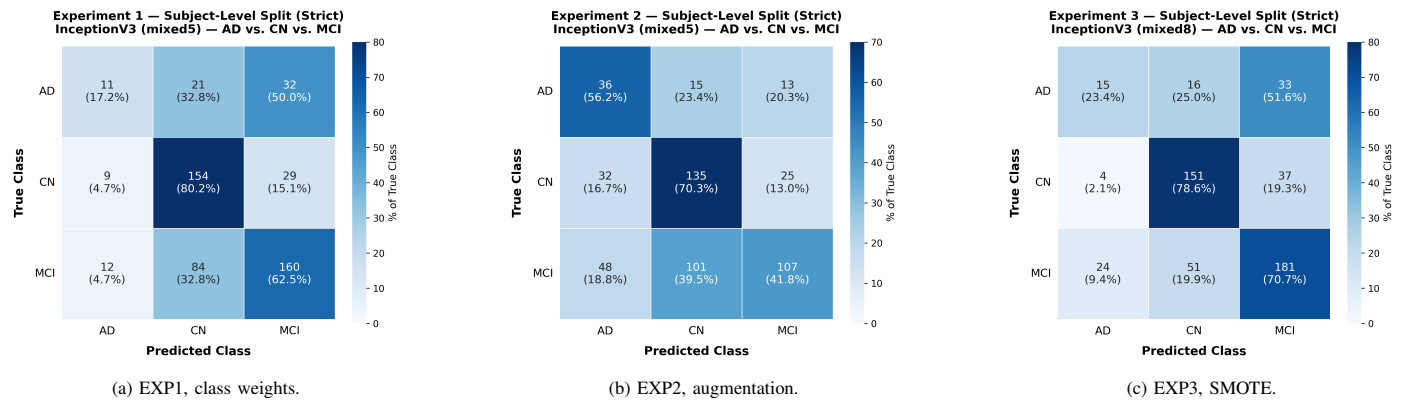


Fig. 2. Confusion matrices of the three experiments on the AD/CN/MCI task, with a strict subject-level partition.

TABLE V. EFFECT OF THE PARTITIONING PROTOCOL ON THE AD/CN/MCI TASK. THE SAME MODELS AND THE SAME DATA ARE EVALUATED WITH A STRICT SUBJECT-LEVEL PARTITION AND WITH A SLICE-LEVEL PARTITION THAT ALLOWS SLICES OF THE SAME SUBJECT TO APPEAR IN THE TRAINING SET AND IN THE TEST SET

Experiment	Protocol	Accuracy (%)	Macro Sensitivity (%)	Macro Specificity (%)	F1 macro	Apparent gain
EXP1: class weights	Subject-level	63.48	53.30	79.56	0.528	+27.48 points
	Slice-level	90.96	93.82	95.31	0.917	
EXP2: augmentation	Subject-level	54.30	56.12	77.02	0.514	+36.47 points
	Slice-level	90.77	91.95	94.67	0.901	
EXP3: SMOTE	Subject-level	67.77	57.60	81.82	0.577	+23.39 points
	Slice-level	91.16	90.23	94.53	0.905	

TABLE VI. COMPARISON OF THE OBTAINED RESULTS WITH THE WORKS OF THE LITERATURE USING A STRICT SUBJECT-LEVEL PARTITION

Reference	Architecture	Dataset	Accuracy (%)
Wen et al. [11]	Standard CNN	ADNI (~400)	55–65
Carcagni et al. [16]	DenseNet201	ADNI	69
This work, EXP1	Inception-V3	ADNI (106)	63.48
This work, EXP2	Inception-V3	ADNI (106)	54.30
This work, EXP3	Inception-V3	ADNI (106)	67.77

share the shape of the skull, the thickness of the cortex and the artefacts of a same acquisition, and these characteristics allow the network to recognise them without having learned anything about the disease. Since the interpolated pseudo-subjects of the

third experiment have already diversified the training set, the advantage given by the leakage is slightly smaller in that case.

The practical consequence is that the three strategies, which are separated by more than thirteen accuracy points under the correct protocol, are all located within an interval of less than half a point under the leaked protocol. A comparative study conducted in these conditions would not only report optimistic values, it would also be unable to tell which of the compared methods is the best one.

E. Comparison With the Literature

Table VI compares our results with the works of the literature that respect a strict subject-level partition. The works that use a slice-level partition, listed in Table I, are not comparable and are not included here.

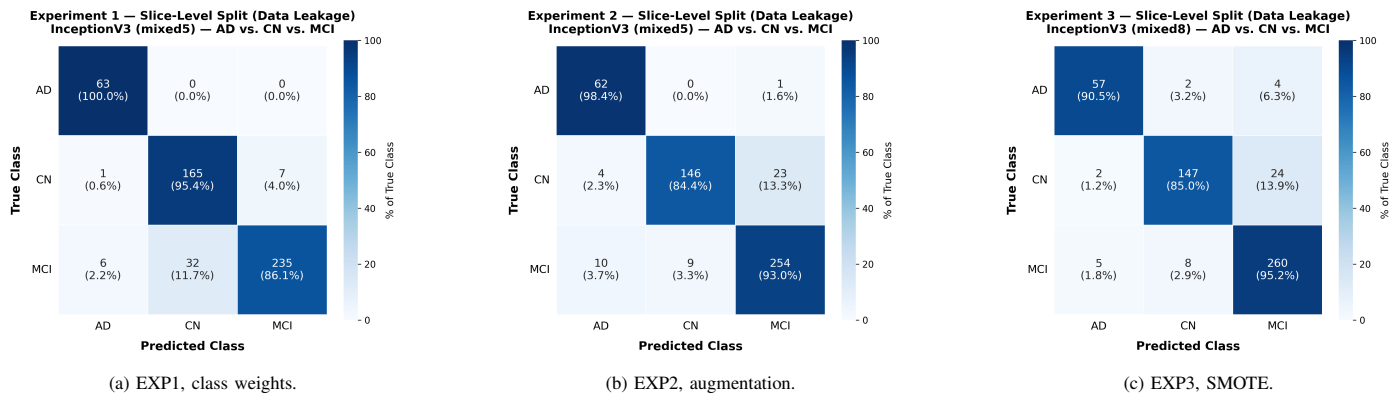


Fig. 3. Confusion matrices of the three experiments on the AD/CN/MCI task, with a slice-level partition producing data leakage.

The accuracies obtained with the first and third experiments, 63.48% and 67.77%, are located in the upper part of the range reported by the studies using a rigorous protocol. Wen et al. [11] report between 55.2% and 65% on a larger subset of ADNI with a standard CNN, and Carcagni et al. [16] do exceed 69% with a DenseNet201 based architecture on ADNI. Given the size of our cohort, 106 subjects among which only 13 are AD, these values are credible and consistent with the real difficulty of the task.

The CN versus MCI task remains the limiting factor in the three experiments, with accuracies between 53.57% and 57.37%. This result agrees with what the literature has been reporting for several years. The MCI group is clinically heterogeneous, some of its subjects being close to a normal profile and others close to an AD profile, and T1-weighted MRI alone does not always contain enough information to separate them. The limitation comes from the definition of the class and from the modality, more than from the architecture.

V. CONCLUSION

The use of transfer learning has been investigated for the classification of the three diagnostic groups of the ADNI database from brain MRI. We used as a baseline the Inception-V3 model pre-trained on ImageNet, where the convolutional part is used as a fixed feature extractor, and then explored the fine-tuning of the model by gradually unfreezing its top convolutional blocks. Three strategies were compared to deal with the imbalance between the classes. The best result on the AD/CN/MCI task was obtained with a subject-level version of SMOTE and an unfreezing from the block mixed8, with an accuracy of 67.77%, against 63.48% with class weights and 54.30% with data augmentation. Separation between CN and MCI remained the most difficult part of the task in the three experiments.

The second part of the work concerns the evaluation protocol. When the same three models are evaluated with a partition performed at the slice level, their accuracy rises to between 90.77% and 91.16%, which corresponds to an apparent gain of 23.39 to 36.47 points. The largest gain was obtained by the weakest model, so that the three strategies, separated by more than 13 points under the correct protocol, all fall within half a point of each other under the leaked one.

This shows that the level at which the data are partitioned has more influence on the reported accuracy than the choice of the architecture and that a result published without this information cannot be interpreted.

ACKNOWLEDGMENT

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<https://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 https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (<https://www.fnih.org/>). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California,

San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

REFERENCES

- [1] D. G. Gadhav, V. V. Sugandhi, S. K. Jha, *et al.*, "Neurodegenerative disorders: mechanisms of degeneration and therapeutic approaches with their clinical relevance," *Ageing Research Reviews*, vol. 99, p. 102357, 2024.
- [2] D. M. Wilson, M. R. Cookson, L. Van Den Bosch, *et al.*, "Hallmarks of neurodegenerative diseases," *Cell*, vol. 186, no. 4, pp. 693–714, 2023.
- [3] M. Sharma, P. Pal, S. K. Gupta, "Advances in Alzheimer's disease: a multifaceted review of potential therapies and diagnostic techniques for early detection," *Neurochemistry International*, vol. 177, p. 105761, 2024.
- [4] C. Toader, N. Dobrin, F.-M. Brehar, *et al.*, "From recognition to remedy: the significance of biomarkers in neurodegenerative disease pathology," *International Journal of Molecular Sciences*, vol. 24, no. 22, p. 16119, 2023.
- [5] C. Dang, Y. Wang, Q. Li, Y. Lu, "Neuroimaging modalities in the detection of Alzheimer's disease-associated biomarkers," *Psychoradiology*, vol. 3, p. kkad009, 2023.
- [6] P. Huang, M. Zhang, "Magnetic resonance imaging studies of neurodegenerative disease: from methods to translational research," *Neuroscience Bulletin*, vol. 39, no. 1, pp. 99–112, 2023.
- [7] A. Bronte, E. Prieto, G. Quincoes, *et al.*, "The basics of PET molecular imaging in neurodegenerative disorders with dementia and/or parkinsonism," *European Radiology*, vol. 35, pp. 4621–4634, 2025.
- [8] J. R. Loftus, S. Puri, S. P. Meyers, "Multimodality imaging of neurodegenerative disorders with a focus on multiparametric magnetic resonance and molecular imaging," *Insights into Imaging*, vol. 14, p. 8, 2023.
- [9] M. G. Alsubaie, S. Luo, K. Shaukat, "Alzheimer's disease detection using deep learning on neuroimaging: a systematic review," *Machine Learning and Knowledge Extraction*, vol. 6, pp. 464–505, 2024.
- [10] L. Zhang, M. Wang, M. Liu, D. Zhang, "A survey on deep learning for neuroimaging-based brain disorder analysis," *arXiv preprint arXiv:2005.04573*, 2020.
- [11] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-Gonzalez, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, "Convolutional neural networks for classification of Alzheimer's disease: overview and reproducibility," *Medical Image Analysis*, vol. 63, p. 101694, 2020.
- [12] E. E. Bron, S. Klein, J. M. Pappa, *et al.*, "Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease," *NeuroImage: Clinical*, vol. 31, p. 102712, 2021.
- [13] S. B. Shahid, M. Kaikaus, M. H. Kabir, *et al.*, "Novel deep learning for multi-class classification of Alzheimer's in disability using MRI datasets," *Frontiers in Bioinformatics*, vol. 5, p. 1567219, 2025.
- [14] B. Sener, K. Acici, E. Sumer, "Improving early detection of Alzheimer's disease through MRI slice selection and deep learning techniques," *Scientific Reports*, vol. 15, no. 1, p. 29260, 2025.
- [15] N. A. Mohsin, M. H. Abdulameer, "Evaluating the impact of 2D MRI slice orientation and location on Alzheimer's disease diagnosis using a lightweight convolutional neural network," *Journal of Imaging*, vol. 11, no. 8, p. 260, 2025.
- [16] P. Carcagni, M. Leo, M. Del Coco, C. Distanto, A. De Salve, "Convolution neural networks and self-attention learners for Alzheimer dementia diagnosis from brain MRI," *Sensors*, vol. 23, no. 3, p. 1694, 2023.
- [17] P. J. LaMontagne, T. L. S. Benzinger, J. C. Morris, *et al.*, "OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease," *medRxiv*, 2019.
- [18] S. Liu, A. V. Masurkar, H. Rusinek, *et al.*, "Generalizable deep learning model for early Alzheimer's disease detection from structural MRIs," *Scientific Reports*, vol. 12, no. 1, p. 17106, 2022.
- [19] B. A. C. Ramalho, L. R. Bortolato, N. D. Gomes, *et al.*, "The impact of the orientation of MRI slices on the accuracy of Alzheimer's disease classification using convolutional neural networks," *Journal of Medical Artificial Intelligence*, vol. 7, 2024.
- [20] J. B. Bae, S. Lee, W. Jung, *et al.*, "Identification of Alzheimer's disease using a convolutional neural network model based on T1-weighted magnetic resonance imaging," *Scientific Reports*, vol. 10, no. 1, p. 22252, 2020.
- [21] M. Turanli, "Multi-orientation hippocampus-centered 3D CNN with attention mechanism for Alzheimer's disease classification from MRI scans," *medRxiv*, 2025.
- [22] S. Bekhet, N. Saad, S. Farag, "Alzheimer disease predicting from clinical and MRI data using DeepALZNET dual pathway framework," *Scientific Reports*, vol. 15, no. 1, p. 43250, 2025.
- [23] M. Sethi, S. Rani, A. Singh, J. L. V. Mazon, "A CAD system for Alzheimer's disease classification using neuroimaging MRI 2D slices," *Computational and Mathematical Methods in Medicine*, vol. 2022, p. 8680737, 2022.
- [24] S. Fathi, A. Ahmadi, A. Dehnad, M. Almasi-Dooghaee, M. Sadegh, "A deep learning-based ensemble method for early diagnosis of Alzheimer's disease using MRI images," *Neuroinformatics*, vol. 22, no. 1, pp. 89–105, 2024.
- [25] D. AlSaeed, S. F. Omar, "Brain MRI analysis for Alzheimer's disease diagnosis using CNN-based feature extraction and machine learning," *Sensors*, vol. 22, no. 8, p. 2911, 2022.
- [26] P. K. Mandal, R. V. Mahto, "Deep multi-branch CNN architecture for early Alzheimer's detection from brain MRIs," *Sensors*, vol. 23, no. 19, p. 8192, 2023.
- [27] T. Akan, S. Akan, S. Alp, C. R. Ledbetter, M. A. N. Bhuiyan, "AlzFormer: video-based space-time attention model for early diagnosis of Alzheimer's disease," *Neuroscience*, vol. 585, pp. 133–143, 2025.
- [28] Q. Abbas, A. Hussain, A. R. Baig, "CAD-ALZ: a blockwise fine-tuning strategy on convolutional model and random forest classifier for recognition of multistage Alzheimer's disease," *Diagnostics*, vol. 13, p. 167, 2023.
- [29] N. Alqahtani, S. Alam, I. Aqeel, *et al.*, "Deep belief networks (DBN) with IoT-based Alzheimer's disease detection and classification," *Applied Sciences*, vol. 13, p. 7833, 2023.
- [30] K. Ashwini, K. Valarmathi, B. Vanipriya, M. Devi, "Alzheimer's disease detection using deep neural network in DenseNet 169," in *Proc. 6th Int. Conf. on Intelligent Computing*, Atlantis Press, 2023, pp. 20–25.
- [31] C. Tinauer, M. Sackl, R. Stollberger, R. Schmidt, S. Ropele, C. Langkammer, "Skull-stripping induces shortcut learning in MRI-based Alzheimer's disease classification," *Insights into Imaging*, vol. 16, no. 1, p. 283, 2025.
- [32] N. M. Khan, M. Hon, N. Abraham, "Transfer learning with intelligent training data selection for prediction of Alzheimer's disease," *arXiv preprint arXiv:1906.01160*, 2019.
- [33] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, J. C. Gee, "N4ITK: improved N3 bias correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.
- [34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the Inception architecture for computer vision," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.
- [35] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [36] I. Loshchilov, F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2019.