Automatic motion artefact detection in brain T1-weighted magnetic resonance images from a clinical data warehouse using synthetic data

Sophie Loizillon, Simona Bottani, Aurélien Maire, Sebastian Ströer, Didier Dormont, Olivier Colliot, Ninon Burgos

Automatic motion artefact detection in brain T1-weighted magnetic resonance images from a clinical data warehouse using synthetic data

Sophie Loizillon, Simona Bottani, Aurélien Maire, Sebastian Ströer, Didier Dormont, Olivier Colliot, Ninon Burgos, the APPRIMAGE Study Group, for the Alzheimer's Disease Neuroimaging Initiative

***Abstract*—Containing the medical data of millions of patients, clinical data warehouses (CDWs) represent a great opportunity to develop computational tools. Magnetic resonance images (MRIs) are particularly sensitive to patient movements during image acquisition, which will result in artefacts (blurring, ghosting and ringing) in the reconstructed image. As a result, a significant number of MRIs in CDWs are corrupted by these artefacts and may be unusable. Since their manual detection is impossible due to the large number of scans, it is necessary to develop tools to automatically exclude (or at least identify) images with motion in order to fully exploit CDWs. In this paper, we propose a novel transfer learning method for the automatic detection of motion in 3D T1-weighted brain MRI. The method consists of two steps: a pre-training on research data using synthetic motion, followed by a fine-tuning step to generalise our pre-trained model to clinical data, relying on the labelling of 4045 images. The objectives were both (1) to be able to exclude images with severe motion, (2) to detect mild motion artefacts. Our approach achieved excellent accuracy for the first objective with a balanced accuracy nearly similar to that of the annotators (balanced accuracy>80 %). However, for the second objective, the performance was weaker and substantially lower than that of human raters. Overall, our framework will be useful to take advantage of CDWs in medical imaging and highlight the importance of a clinical validation of models trained on research data.**

***Index Terms*—Clinical Data Warehouse, Deep Learning, Motion, MRI**

I. Introduction

Recently, hospitals have created clinical data warehouses (CDWs) that gather medical images from thousands to millions of patients [1–3]. These resources represent an exceptional opportunity to develop computational tools [4]. In contrast to research datasets where acquisition protocols are well standardised, the quality of CDW images is highly heterogeneous. Images come from different hospitals over several decades and diverse machines are used with no homogenisation on the acquisition parameters [3]. Once an image is acquired at the hospital, it will immediately be saved in the picture archiving and communication system (PACS), meaning that a non negligible number of unusable images will be archived. For instance, if a patient moves during the acquisition, the corrupted image will still be stored in the PACS. Therefore, quality control (QC) is a fundamental first step before developing any machine learning project on a CDW.

Magnetic resonance (MR) images are sensitive to motion induced by patient movement during the acquisition process. As they require a long acquisition time, subjects are likely to move during the examination, which causes artefacts in the reconstructed image. This will appear as blurring, ringing, ghosting or signal loss depending on the timing and spatial changes during acquisition [5]. Thus, motion can be a serious confounding factor for further neuroimaging analyses. This can have dramatic consequences when the presence of motion artefacts is correlated with a diagnosis of interest (e.g. patients with a specific disease have a tendency to move more often) since it would lead to biased models. Previously, based on a CDW gathering data from 39 different hospitals, we found that 25 % of MRI were considered totally unusable for further processing, and almost a third had a very low quality especially due to motion [6]. Another study conducted in a single hospital showed that the prevalence of repeating an MRI examination due to the presence of motion was up to 20 % of all the acquisitions [7]. Beyond the cost that this represents for hospitals, these studies are highlighting the fact that many images present in the PACS are simply unusable, often because corrupted by motion artefacts. Therefore, it is important to be able to automatically exclude such images before conducting any study on a CDW. What's more, several works [8–10] have shown the impact of motion in the use of brain imaging software packages such as Freesurfer [11] or SPM [12]. The presence of motion artefacts induces a constant bias in the morphometric analyses, leading to a reduced estimation of the grey matter volume [9] as well as of the cortical thickness [8].

Quality control is needed to fully exploit the potential of CDWs and important efforts have been made to propose automatic QC tools, including the detection of motion artefacts [13–17]. Esteban et al. [13] introduced MRIQC, a pipeline for the automatic QC of 3D brain T1-weighted (T1w) MRI based on image quality measures (IQMs). It enables the extraction of IQMs such as the signal-to-noise ratio, the contrast-to-noise ratio or the volume of grey and white matter. This method relies on an extensive pre-processing pipeline using neuroimaging software packages such as ANTs [18] and FSL [19], which are only usable on good quality images and therefore incompatible with CDWs. Sadri et al. [14] developed MRQy, a quantitative tool to quickly determine relative differences in MRI volumes within and between large MRI cohorts. As MRIQC, MRQy is based on the extraction of IQMs but it does not require extensive pre-processing thanks to an Otsu thresholding to distinguish the foreground, which includes the whole head, from the background. QC methods based on convolutional neural networks (CNNs) have

The research leading to these results has received funding from the Abeona Foundation (project Brain@Scale), from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAHU-0006 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6)".

Sophie Loizillon, Simona Bottani, Olivier Colliot and Ninon Burgos are with Sorbonne Université, Institut du Cerveau – Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, F-75013, Paris, France.

Aurélien Maire is with AP-HP, WIND department, F-75012, Paris, France.

Sebastian Ströer is with AP-HP, Hôpital Pitié Salpêtrière, Department of Neuroradiology, F-75013 Paris, France.

Didier Dormont is with Sorbonne Université, Institut du Cerveau – Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, DMU DIAMENT, F-75013, Paris, France.

Members of the APPRIMAGE study group can be found at https://www.aramislab.fr/apprimage.

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

Corresponding author: Ninon Burgos (ninon.burgos@cnrs.fr)

also been proposed [15, 20–23]. They have the advantage of learning features without knowing a priori which are the most adapted. Sujit et al. [20] developed an ensemble deep learning model based on CNNs to automatically evaluate the quality of multi-centre structural brain MR images. A limitation of this work is that it relies on images acquired following a well-defined research protocol, which are not representative of the heterogeneity of clinical images. Lei et al. [15] presented a multi-task CNN framework for artefact-based MRI quality assessment, which not only provides a quality score but interprets the cause of the poor image quality. Image rulers, which consists of several versions of the original MRI slices with one type of artefact (noise or motion), are used during inference time. Each of these MRIs will be run through the trained CNN and the different outputs will be compared with the test image. The use of a single image ruler consisting of different versions of a single scan makes this method incompatible with the high heterogeneity that characterises CDWs, where different types of artefacts can coexist in a single image (e.g. noise and motion).

Previously, we proposed a framework for the automatic QC of T1w brain MRI in a CDW using deep learning techniques [6]. 5500 MRIs were manually annotated with a three-level grade for three characteristics: noise, contrast and motion. According to these grades, we determined three tiers corresponding to images of good, medium and bad quality. CNNs were then trained to rate the overall image quality. Our classifier was as efficient as manual rating for the classification of images which are not proper 3D T1w brain MRI (e.g., images of segmented tissues or truncated images). It was also able to recognise low quality images with good accuracy. While the detection of certain features such as noise did not present any particular difficulty for our model, the detection of motion artefacts proved more problematic.

As motion quantification is a complex problem, particularly due to its sensitivity to many cofactors such as contrast, there is a lack of dataset with reliable quantitative ground truths. Some studies thus rely on synthetic motion to detect motion artefacts in a controlled way [24–26]. Despite the excellent results claimed in the literature, only few papers have attempted to validate their performance on data with real motion. And even when they did, their test sets were extremely limited and only composed of research data [26, 27]. It is yet unclear how they would perform on routine clinical data.

In this paper, we propose a transfer learning framework for the automatic detection of motion artefacts in 3D T1w brain MRI from a CDW. We generated synthetic motion in MR images of research databases to train a CNN classifier which was validated on synthetic and real motion artefacts. Our model was then generalised to clinical data with an effective transfer learning technique using 4045 labelled MRIs from a CDW. Preliminary work was accepted for publication in the proceedings of the SPIE Medical Imaging 2023 conference [28]. Contributions specific to this paper include i) a comparison of the two main synthetic motion generation approaches for the detection of motion artefacts in a CDW: the image and k-space based approaches; ii) the implementation of four deep learning architectures (Conv5FC3, ResNet, SE-CNN and ViT) for the detection of motion artefacts; iii) an optimisation of the fine-tuning parameters; iv) a comparison of the proposed framework with MRQy, a QC approach based on IQMs [14].

II. Background

In this work, we focus on the detection of motion artefacts in 3D T1w MRIs in a CDW. As most of the automatic QC tools rely on neuroimaging software packages that are only usable on good quality images, they are incompatible with CDWs. What's more, manual annotation of motion artefacts is a challenging task. When an image is degraded, it may be difficult to properly distinguish motion from noise or bad contrast. Hence the idea of simulating motion, which can be done automatically and provides reliable ground truths.

Head motion can be well approximated as rigid body motion, which requires six degrees of freedom, comprising three translations and three rotations [27]. Lee et al. [29] described the two main approaches for motion simulation in brain MRIs: the image and the k-space based techniques.

As illustrated in Fig. 1 (left), the image-based approach assumes that the subject takes Nt different positions during the acquisition. First, Nt rigid transformations of the motion free image are applied before computing the fast Fourier transform (FFT). A new k-space is then built by concatenating blocks for the Nt different simulated positions. Finally, in order to obtain the final synthetic image corrupted by motion, an inverse FFT is applied [26, 30]. While in the image-based approach, motion parameters are applied on the motion clean MRI, the k-space based method directly uses the raw k-space to perform the simulation of motion (Fig. 1, right). In their algorithm, Loktyushin et al. [31] start by applying the rotation to the k-space grid and perform a non-uniform FFT. Then, a linear phase shift proportional to the amplitude of the translation is added before the final FFT. More recently, Al-masni et al. [32] introduced a new approach by combining the image and the k-space based methods, where the translation was directly performed in the k-space domain, whereas the rotation was applied on the image. This method has the advantage of preserving the uniformity of the k-space sampling as the rotation is applied in the image domain.

Motion detection in MRI with deep learning techniques has been studied in [24, 27, 33] using datasets of images corrupted with synthetic motion obtained from motion free MRI. Mohebbian et al. [24] developed a stacked ensemble model to classify motion artefacts into five severity levels in brain MRIs. While their model was perfectly able to predict, across different sequences (T1w and T2w), synthetic motion artefact (balanced accuracy $>$90 %), their approach was not validated on MRIs with real motion. Oksuz [33] introduced a dense CNN to detect motion in brain MRIs and successfully validated their binary algorithm using 28 MRIs from a research dataset (balanced accuracy: 97.8 %). This method was also only validated on research MRIs corrupted with synthetic motion artefacts. Recently, Sagawa et al. [34] presented a CNN trained using images corrupted with synthetic motion labelled with their full-reference image quality assessment (FR–IQA) metrics to predict with a high accuracy these metrics. Their approach enables a quantitative assessment of motion artefacts without any reference image. The model classified real motion artefacts from research MRIs with an AUC of 0.928.

The requirement for accurate ground truths encourages researchers to develop solutions using synthetic data, where labels are easily available. However, the anatomical complexity and diversity of healthy and pathological brain tissues makes it difficult to generate an appropriate spectrum of synthetic MRIs, which leads to poor performance of the classifiers at the stage of inference on real data [35, 36]. To benefit from the use of synthetic data, it is thus important to bridge the gap between synthetic and real data.

Transfer learning applies knowledge learned from one domain and one task to another related domain and task. In the case of motion detection using synthetic data, if we have labels for both synthetic and real data, we can resort to the use of fine-tuning (inductive transfer learning). Fine-tuning involves transferring the weights from a pre-trained network to the network to be trained. In a classification context, a common practice is to replace some of the last fully connected layers of the pre-trained CNN with new fully connected layers to target the new application. Tajbakhsh et al. [37] demonstrated that the use of a pre-trained CNN with fine-tuning

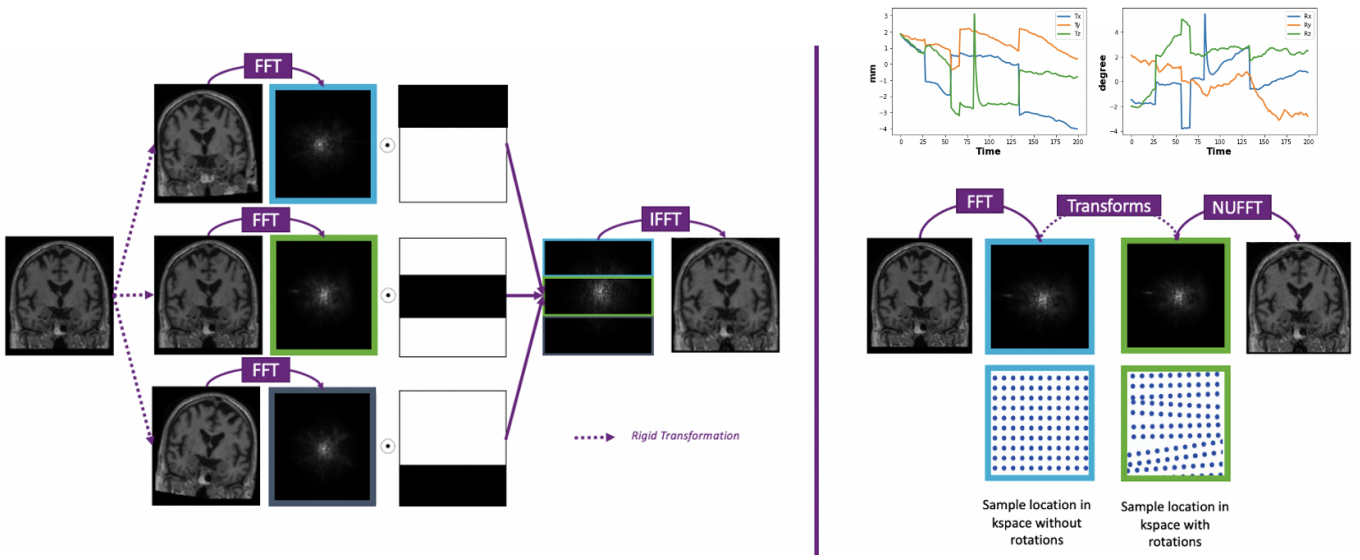

Fig. 1. Left: Image-based motion simulation. (1) Nt rigid transformations of the motion free image are applied (here $Nt = 2$), (2) Fast Fourier transform (FFT) of the original and Nt transformed images, (3) Concatenation of the $Nt + 1$ blocks to create a new k-space, (4) Inverse FFT (IFFT) to obtain the motion corrupted image. Right: k-space based motion simulation. A time course example of the six parameters is displayed at the top of the figure (left: translation parameters, right: rotation parameters). (1) FFT of the motion free image. (2) Transformation (rotation + translation) for each point of the time course. (3) Non uniform FFT (NUFFT) to reconstruct the corrupted image (because of the non uniform sample spacing due to the rotation in the k-space).

outperformed or, in the worst case, performed as well as a CNN trained from scratch for four distinct medical imaging applications. Although the distance between natural images and medical images is considerable, Ahmed et al. [38] also showed that fine-tuning a CNN, initially trained on ImageNet, by transferring the learned feature representations to the MRI-based survival time prediction task, performed better than training from scratch.

III. Materials and Methods

We developed an approach based on the generation of synthetic motion to improve the detection of motion artefacts in clinical T1w brain MR images. We used T1w images, which were acquired with scanners from different manufacturers and different magnetic fields, from publicly available research data sources as well as from a CDW. Motion artefacts were synthetically generated by applying both image and k-space based approaches using rigid body transformations to simulate different severity degrees of artefacts. CNNs were first trained on research databases to recognise synthetic motion, and their performance was evaluated on real motion. We generalised our model to the CDW by applying an efficient transfer learning technique.

A. Datasets

To detect motion artefacts in routine clinical images, we first used three publicly available research datasets to pre-train a CNN on images with synthetic motion artefacts. Then, images from our CDW were exploited for transfer learning and validation.

1) Research databases: We worked with the ADNI, MSSEG, and MNI BITE research databases to cover a wide spectrum of pathologies that can be found in a CDW. A special attention was paid to the search for contrast-enhanced T1w MRI as the CDW includes images acquired with and without injection of a gadolinium-based contrast agent.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a multi-site study of elderly individuals with normal cognition, mild cognitive impairment, or Alzheimer's disease [39]. The ADNI-1 phase included T1w MRIs acquired on 1.5 T scanners from different manufacturers (GE, Siemens, and Philips). A two-level quality control was performed, one related to the adherence to the protocol parameters and one to the series-specific quality. Part of the metadata, the IPMOTION score indicates the absence of motion (0), or the presence of mild (1), moderate (2) or severe (3) motion artefacts. This score was used to select motion free T1w MRI. Our selection procedure resulting in 1143 MR images for 70 subjects is detailed in Appendix (Fig. A.1). We also created a small test set with MRIs corrupted by motion artefacts based on the IPMOTION and the comments section of the corresponding metadata file.

The MSSEG MICCAI challenge, which aim is to perform the segmentation of multiple sclerosis lesions, includes 53 patients across four different sites [40]. Four different MRI scanners were used: GE Discovery 3 T, Philips Ingenia 3 T, Siemens Aera 1.5 T and Siemens Verio 3 T. Each scan included four MRI sequences: 3D FLAIR, 3D T1w, 3D contrast-enhanced T1w and 2D T2w. In our study, we only considered the 3D contrast-enhanced T1w.

The Montreal Neurological Institute's Brain Images of Tumors (MNI BITE) database made available pre and postoperative MR and ultrasound images acquired from brain tumour patients [41]. The study includes 13 patients with gliomas, who underwent a pre- and post-operative contrast-enhanced T1w MRI using the 1.5 T GE Signa EXCITE scanner.

Demographic information for each of these databases is reported in Table I.

2) Clinical data warehouse: The clinical routine data come from a large CDW containing all the T1w brain MRIs of adult patients scanned in hospitals of the Greater Paris area (Assistance Publique-Hôpitaux de Paris [AP-HP]). The large number of hospitals being part of the AP-HP consortium (39 hospitals) and the huge number of images collected every day is making this CDW a good representation of 3D T1w brain MRI that may be acquired in other hospitals.

We used the same dataset as in our previous study, where we randomly selected 5500 images, corresponding to 4177 patients that were acquired on various scanners: Siemens Healthineers (n = 3752), GE Healthcare (n = 1710), Philips (n = 33) and Toshiba (n = 5) [6].

Motion artefacts were manually annotated as a three-grade level by

TABLE I
DISTRIBUTION OF THE SEX AND AGE OVER THE RESEARCH (ADNI, MSSEG AND MNI BITE) AND THE CLINICAL (AP-HP) DATASETS.

	Database	N patients	N images	Age in years [range]	Sex (%F)
Research databases	ADNI	70	1143	74.31 ± 7.11 [55,90]	41.43 %
	MSSEG	53	53	45.42 ± 10.27 [24,66]	71.70 %
	MNI BITE	13	26	52.00 ± 17.70 [31,76]	35.71 %
Clinical data warehouse	AP-HP	4177	5500	55.15 ± 7.89 [18, 95]	55.39 %

two annotators. A score of 0 was given when no motion was seen, 1 when the structures of the brain was distinguishable despite the presence of motion and 2 when the cortical and sub-cortical structures were difficult to distinguish (Fig. A.2). Some of the 5500 images did not correspond to 3D T1w brain MRI (e.g. because of truncation) and were therefore not labelled with a motion score (SR: straight reject, n=1455). If the users labelled differently a given MRI, the consensus grade was chosen as the maximum of the two grades. The weighted Cohen's kappa was used to evaluate the inter-rater agreement between the annotators and a moderate agreement was found with a score of 0.68. Among the 4045 images that were not labelled as straight reject, 2319 had a consensus motion score of 0, 1196 a score of 1 and 530 a score of 2. Patients' demographics are reported in Table I.

B. Image pre-processing

To make the annotation process easier, MRIs were pre-processed using Clinica [42] and its `t1-linear` pipeline. First a bias field correction was applied using the N4ITK method [43]. An affine registration to the MNI space was then performed [18]. Next, images were cropped to remove background resulting in images of size 169×208×179, with 1 mm isotropic voxels [44]. The Z-score method, which consists of subtracting the mean intensity of the entire image from each voxel value and dividing it by the corresponding standard deviation, was used to normalise the voxel intensities. Our initial aim was to obtain a rough alignment and intensity rescaling to facilitate annotation but this pre-processing is also useful when training CNNs.

C. Proposed approach

We developed a transfer learning approach to detect motion artefacts in clinical images based on motion simulated on research images. Our method is composed of two steps: (1) A pre-training task using synthetic motion to distinguish motion-free from motion-corrupted images, (2) A fine-tuning task to improve the generalisation ability of our pre-trained network on clinical datasets.

1) Motion generation: Because of the lack of research dataset with quantitative assessment of motion artefacts in T1w brain MRI, we adopted an approach based on synthetic motion. We compared the two main motion simulation techniques described in Section II: the image and the k-space based approaches. We used the open-source Python library TorchIO and its function `RandomMotion` described in [30] for the image based approach and the `RandomMotionTimeCourseAffines` implemented in [45] for the k-space method. The `RandomMotion` function was used with a limited number of rigid transformations ($Nt = 4$) due to its computation time, whereas the `RandomMotionTimeCourseAffines` was applied using 200 points of the simulated time course ($nT = 200$).

By selecting different translation and rotation range parameters, several degrees of motion severity can be generated. Different values have been tested in this study to simulate motion ranges of [2 mm,8 mm] for translation and [2°,8°] for rotation.

2) Network architectures: To classify motion artefacts, we used a CNN composed of five convolutional blocks and of three fully connected layers (denoted as Conv5FC3) that proved successful in our previous work [6]. Each convolutional block is made of a convolutional layer, a batch normalisation layer, a ReLU activation function and a max pooling layer. The weighted binary cross-entropy was used as loss function. The learning rate of the Adam optimiser was set to 1e-4 and the batch size equals to 16. The model with the lowest loss on the validation set was saved as final model. The architecture was implemented using Pytorch and is available through the ClinicaDL software on GitHub (https://github.com/aramis-lab/clinicaDL) [46].

We compared this network to more sophisticated architectures such as a 3D ResNet, a Squeeze and Excitation CNN (SE-CNN) and a 3D Vision Transformer (ViT). The 3D ResNet inspired by [47] was previously used to predict brain age from 3D T1w MRI and outperformed the Conv5FC3 [48]. The combination of a ResNet with Squeeze and Excitation blocks was successfully tested on brain tumour classification [49]. SE blocks are composed of a squeeze and an excitation step. The squeeze operation is obtained through an average pooling layer and provides a global understanding of each channel. The excitation part consists of a two-layer feed-forward network that outputs a vector of n values corresponding to the weights of each channel of the feature maps. Whereas traditional CNNs weight each of their channels equally when creating feature maps, SE-CNNs weight each channel adaptively through this content-aware mechanism. Transformers, which have become the model of choice in natural language processing, have recently been applied to computer vision tasks. Even if applications in medical imaging remain limited, vision transformers have been used to perform classification (Alzheimer's disease detection) as well as segmentation tasks (brain tumour segmentation) [50, 51]. The different architectures are detailed in Appendix (Fig. A.3, A.4, A.5, and A.6). The same hyperparameters as for the Conv5FC3 were used for the training of these networks.

3) Model generalisation using transfer learning: To detect motion artefacts in clinical MRIs, we pre-trained a classifier on research datasets by simulating motion. We now need to close two gaps, one between research and clinical datasets and one between real and synthetic motion. To do so, we chose a transfer learning method based on fine-tuning.

The key idea of fine-tuning is to transfer knowledge learnt from one domain to another one. We first trained a classifier to learn features for the research datasets' domain using motion simulation. Then, the network was optimised again for a new domain (clinical dataset with real motion) by allowing the re-training several layers of the pre-training model and freezing the weights of the other layers. Thus, we were able to generalise our model from synthetic to routine clinical data.

D. Experiments

We performed two sets of experiments that correspond to the two steps of the proposed approach. The first set focuses on the network

pre-training step with research data using synthetic motion and the second set concerns the network fine-tuning step with clinical data and real motion.

1) Network pre-training on research data: At first, we aimed to test the ability of the different deep learning models to detect motion in research-quality images using only synthetic motion while training. We performed a series of experiments on research datasets, where we corrupted motion-free MRIs with different motion severity degrees to study the influence of the translation and rotation ranges. The four different architectures as well as the two motion simulation techniques presented in Section III-C2 and Section III-C1 were tested to determine the best approach for motion detection.

Before starting the experiments, we defined an independent test set by randomly selecting 184 images over the three public datasets and corrupting half of them with different motion severity degrees (rotation: [2°, 8°]; translation: [2 mm, 8 mm])). The remaining 1040 images (520 corrupted with synthetic motion and 520 with no motion) were split into training and validation using a 5-fold cross validation (CV) as shown in Table A.I . The separation between training, validation and test sets was made at the patient level to avoid data leakage. Our model was also validated on a second small test set with ADNI MRIs corrupted by real motion as explained in Section III-A1.

2) Network fine-tuning on routine clinical data: The second set of experiments aims to evaluate the performance of our transfer learning approach. This step is required because of the quality gap that exists between research, where strict acquisition protocols are respected, and clinical data, which suffer from a lack of homogenisation of the acquisition parameters.

To generalise our pre-trained network on clinical datasets, we used a transfer learning technique that consists in fine-tuning our model on two target tasks:

- the detection of severe motion (Mov01vs2): being able to detect these MRIs in CDWs is of great importance as subsequent processing steps are likely to fail on these images,
- the detection of moderate motion (Mov0vs1): MRIs labelled as motion 1 may lead to unreliable diagnostic predictions.

Before starting the experiments, we defined a test set by selecting the same MRIs as in [6] as well as the same training and validation splits with a 5-fold CV (Table A.II). Because of the presence of straight reject MRIs, the different models trained in the CV were evaluated on respectively 328 and 385 images of the test set for the Mov0vs1 and Mov01vs2 tasks.

We also studied the influence of the number of layers to freeze for the different architectures. To evaluate the impact of the proposed method, we compared the results obtained with the use of fine-tuning and by training a model from scratch.

3) Comparison with state-of-the-art QC: Several quality check tools for T1w brain MRI based on IQMs have been developed in the last years [13, 14]. Whereas MRIQC cannot be applied on CDWs because of the need for extensive image pre-processing designed for T1w brain MRI of good quality without gadolinium, MRIQY is compatible with CDW MRIs thanks to its automatic extraction and separation of the background from the foreground with an Otsu thresholding. Thus, we were able to extract 13 IQMs such as noise ratios, variation metrics, entropy, and energy criteria from our clinical images. Thanks to these IQMs, we trained a random forest (RF) classifier to detect the presence of motion artefacts in MRI. We performed a random search in order to optimise the hyper-parameters and particularly analysed: the number of decision trees, the maximum tree-depth, the minimum number of samples per split, and the minimum leaf samples. 300 different combinations were tested using a 5-fold CV. To evaluate the impact of our method over a machine learning approach, we compared the RF model with the deep learning models fine-tuned on the clinical dataset.

IV. Results

A. Validation on research data

The ability of deep learning models to detect motion was first assessed using images from research datasets corrupted with synthetic motion. Fig. 2 displays three corrupted images obtained with the image and the k-space approach using different translation and rotation ranges, and the original image without any motion.

We started by evaluating the performance of our Conv5FC3 model trained on synthetic motion when applied to our synthetic independent test set corrupted with different motion severity degrees (rotation: [2°, 8°]; translation: [2 mm, 8 mm]). We studied the influence of the translation and rotation ranges by performing several experiments with different motion severity degrees. We first trained a model with synthetic severe motion by applying a large rotation and translation range ([6°,8°]; [6 mm, 8mm]). The balanced accuracy (BA) on our independent test set is excellent with both motion simulation techniques (>98 %). We also obtained very good results for smaller ranges of rotation and translation as reported in Table II.

Then, we evaluated the ability of these models to detect real motion. As mentioned in Section III-A1, we defined a test set according to the IPMOTION score. Our models were perfectly able to detect motion on these images. No notable differences were noted in terms of performance between the two simulation techniques (Table II).

We also compared the performance obtained by different architectures on the same test set corrupted with synthetic motion. In Table III, we report the results of the four architectures trained using k-space based motion simulation with the following parameters: rotation: [2°, 4°]; translation: [2 mm, 4 mm], as these led to the best results on synthetic and real motion for the Conv5FC3 architecture. The results of the ResNet and the SE-CNN were comparable to that of the Conv5FC3 with a BA>99 %, whereas the ViT BA was lower (BA=97.69 %). Thus, more complex networks did not provide any notable improvement. The same conclusion was reached for the image-based motion simulation technique (Table A.III).

B. Application to routine clinical data

The first set of experiments performed with the routine clinical data consisted in fine-tuning the Conv5FC3 network pre-trained on the research dataset with synthetic image-based motion to detect severe motion (mov01vs2) by unfreezing one to five layers of the Conv5FC3 architecture. We used the same training and validation split as for the mov01vs2 task (Table A.II). The different models trained in the CV were evaluated on the 385 images composing the test set for the mov01vs2 task. Best results were obtained by freezing all the layers except the three fully connected ones (Table A.IV). All the fine-tuning results presented below were trained using this configuration.

The results obtained with the proposed transfer learning framework on our independent clinical test set are presented in Table IV. For the detection of severe motion (mov01vs2), the classifier BA is almost as good as that of the annotators, which is defined as the average of the BA between each rater and the consensus (classifier: 84.52 %; annotators: 86.29 %). For the detection of mild motion (mov0vs1) the classifier BA is low (62.61 %) and lower than that of the raters (73.21 %).

We compared the results obtained with and without fine-tuning to measure the impact of our approach. When applying the network trained on the synthetic research data directly to the clinical data, we observed a large drop in performance with a particularly low

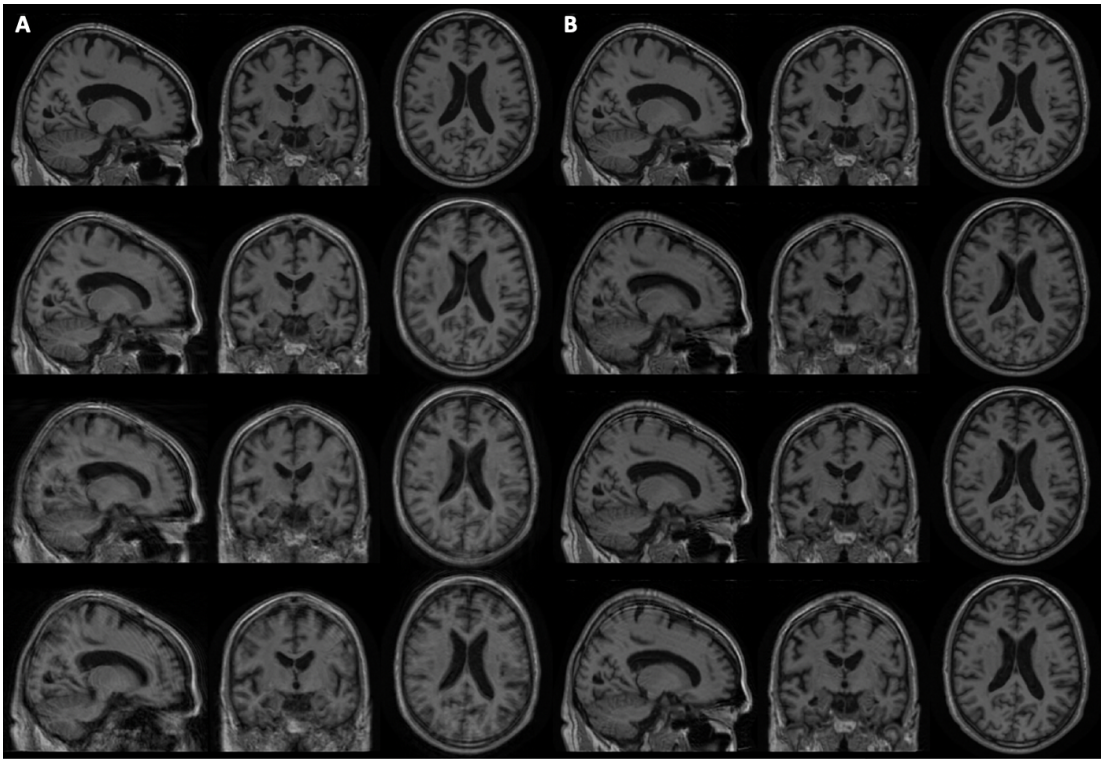

Fig. 2. Example of motion simulation on brain MRI using the k-space (A) and the image (B) based approach with different translation and rotation ranges. From top to bottom: motion free MRI, MRIs corrupted with a rotation of 3°, 5° and 7° and a translation of 3 mm, 5 mm and 7 mm.

TABLE II
RESULTS FOR THE DETECTION OF SYNTHETIC AND REAL MOTION IN T1W BRAIN MRI FROM RESEARCH DATASETS. FOR THE VALIDATION ON SYNTHETIC MOTION, WE REPORT THE MEAN AND THE EMPIRICAL STANDARD DEVIATION ACROSS THE FIVE FOLDS FOR THE BALANCED ACCURACY, SPECIFICITY AND SENSITIVITY. FOR THE DETECTION OF REAL MOTION, ONLY THE ACCURACY OBTAINED BY THE BEST MODEL OF THE 5-FOLD CV WAS REPORTED AS OUR INDEPENDENT TEST SET CONTAINED ONLY IMAGES WITH MOTION. RESULTS ARE DETAILED FOR BOTH SIMULATION APPROACHES: IMAGE AND K-SPACE BASED.

			Cross-validation on synthetic motion			Test on real motion
Motion simulation	Rotation range	Translation range	Balanced accuracy	Specificity	Sensitivity	Accuracy
Image	[6°, 8°]	[6 mm, 8 mm]	98.22 ± 1.39	99.29 ± 0.86	97.14 ± 2.33	100 %
	[4°, 6°]	[4 mm, 6 mm]	97.06 ± 1.47	98.25 ± 1.92	95.87 ± 1.27	100 %
	[2°, 4°]	[2 mm, 4 mm]	95.51 ± 2.47	98.94 ± 2.11	92.06 ± 5.76	98.41 %
k-space	[6°, 8°]	[6 mm, 8 mm]	98.44 ± 0.05	97.22 ± 0.12	99.70 ± 0.01	100 %
	[4°, 6°]	[4 mm, 6 mm]	97.77 ± 0.03	95.56 ± 0.06	100 ± 0.00	100 %
	[2°, 4°]	[2 mm, 4 mm]	99.17 ± 0.03	98.33 ± 0.06	100 ± 0.00	100 %

TABLE III
RESULTS OF FOUR DIFFERENT CNN CLASSIFIERS (CONV5 FC3, RESNET, SE-CNN AND VIT) TRAINED AND TESTED ON MRIS FROM OUR RESEARCH DATASET CORRUPTED WITH K-SPACE BASED MOTION SIMULATION.

Architectures	Balanced accuracy	Specificity	Sensitivity	Training time
Conv5FC3	99.17 ± 0.03	98.33 ± 0.06	100 ± 0.00	3 h 52 min
ResNet	99.72 ± 0.03	99.44 ± 0.07	100 ± 0.00	6 h 27 min
SE-CNN	100 ± 0.00	100 ± 0.00	100 ± 0.00	6 h 18 min
ViT	97.69 ± 0.07	98.61 ± 0.03	96.77 ± 0.04	4 h 31 min

specificity for both tasks. A second comparison was performed between the proposed transfer learning framework and when training with the clinical data from scratch. Our transfer learning method achieved a gain of more than 10 percent points for the detection of severe motion. A much smaller improvement was observed for the detection of mild motion. The receiver operating characteristic curves (ROC) for the detection of severe (mov01vs2) and moderate (mov0vs1) motion in 3D T1w MRIs are shown in Fig. 3. The AUC of the proposed approach for severe motion detection (AUC: 0.85) outperformed that of training from scratch by 5 percent points and that of training on research datasets by 24 percent points. For moderate motion detection, the AUC of the proposed approach (AUC:

0.61) is only 3 percent points higher than that of learning from scratch (AUC: 0.58).

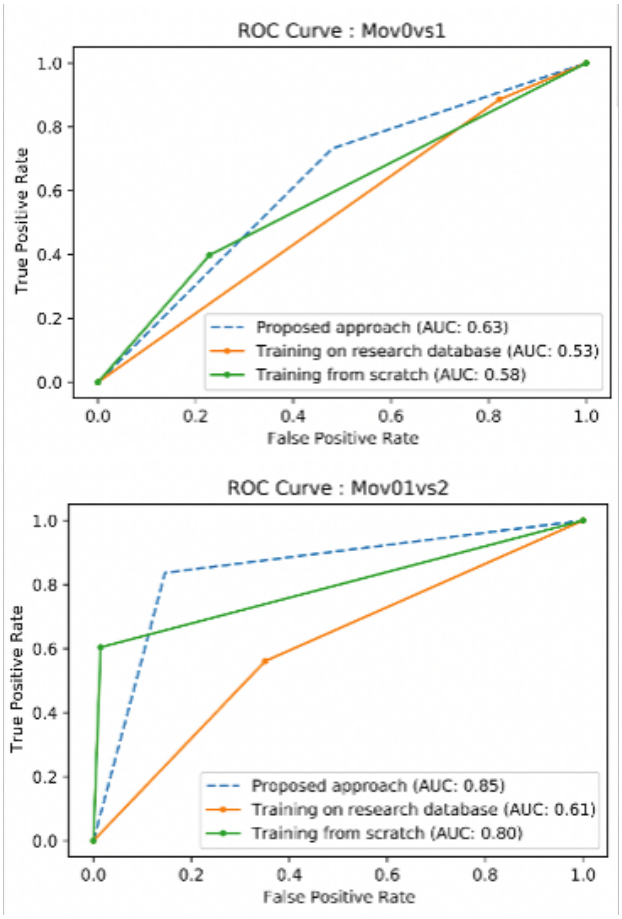

Fig. 3. Receiver operating characteristic curves (ROC) for the detection of severe (mov01vs2) and moderate (mov0vs1) motion in 3D T1w MRIs.

The same set of experiments was performed using the model pre-trained on the research dataset with k-space based synthetic motion (Table A.IV). As with the model pre-trained on image-based motion, our transfer learning method led to a substantial performance improvement when detecting severe motion, with a gain of 6.85 percent points in BA compared with training from scratch on clinical data, but the improvement was limited when detecting mild motion (gain of 1.7 percent point in BA).

Finally, we compared our deep learning method with an IQM based approach (RF classifier trained on IQMs) [14], which achieved a 61.94 % BA for the detection of severe motion and 54.72 % for mild motion detection. Our proposed method outperformed the IQM approach by 8 percent points for the mild motion detection and by 23 percent points for the severe motion.

V. Discussion

In this work, we developed a transfer learning framework for the automatic detection of motion in 3D T1w brain MRI from a CDW. After a pre-training phase using synthetic motion to distinguish images with and without motion artefacts, we enabled the generalisation of our pre-trained network on clinical data using fine-tuning. We validated our approach on labelled clinical MRIs of the AP-HP CDW. To the best of our knowledge, we are the first to propose a very large-scale validation using a CDW for motion artefact detection using synthetic motion.

For the detection of synthetic motion on research datasets, our model achieved excellent results with a BA over 98 % for the image and the k-space based simulation approaches. Trained using only synthetic motion, the model had no difficulty generalising to real motion artefacts and was able to detect every image corrupted with motion on a small independent research dataset. Despite the performance obtained on research datasets, the model was not able to generalise to the CDW (BA: 60.26 %). This poor result, with a low specificity, does not come as a surprise. It highlights the critical importance of validating models trained on research datasets to clinical ones, but also the quality gap that exists between research, where strict acquisition protocols are respected, and clinical data, which suffer from a lack of homogenisation of acquisition parameters. To overcome these issues, we proposed a transfer learning framework that achieved very good results for the detection of severe motion with a BA of 84.52 %, which is nearly as good as that of the annotators (86.29 %) and 10 percent points higher than when training the model from scratch on clinical data (BA: 73.75 %). The performance is significantly different from that of a network trained from scratch ($p = 6.60 \times 10^{-17} < 0.05$, McNemar's test). However, the result for the more difficult task of mild motion detection was less successful with a BA of 62.61 % compared to the 73.21 % obtained by the annotators. Our two-step approach with pre-training and fine-tuning still allowed an increase of almost 5 percent points of BA over training from scratch on this task and the difference between the two approaches was significant ($p = 1.39 \times 10^{-8} < 0.05$, McNemar's test). Our method also outperformed the one proposed by MRQy [14] that consists in training a random forest classifier with IQMs. The proposed approach reached a BA 8 percent points higher for the detection of mild motion and 23 percent points higher for severe motion detection. These results highlight the limitations of MRQY that computes IQMs between the head and the background thanks to an Otsu thresholding. IQMs based on noise measurements such as the coefficient of joint variation or the contrast-to-noise ratio are much more relevant when computed between brain tissues to evaluate how separated their distributions are and thus conclude on the overall image quality. This is why most of the IQMs in MRIQC [13] are evaluated between grey and white matter thanks to a substantial pre-processing requiring good quality data that is incompatible with a CDW.

In the scope of our work, we have been comparing the two main approaches of motion simulation: the image and the k-space based methods. The main limitation of the image based approach is the restriction to a small number of positions due to the computation time needed to reslice each of them and to compute the corresponding FFT. For example, the time required to corrupt MRIs with motion using 4, 16 and 128 positions is respectively 3.12, 8.93 and 63.25 seconds. In contrast, the k-space based approach using motion time courses simulates more realistic motion by considering slow, sudden and swallowing motion with respectively a Perlin noise, a step displacement and a transient motion. Despite these considerations, our proposed transfer learning framework obtained the best performance with the image-based approach. As we only used motion simulation to pre-train our model, it appears that we can limit ourselves to a very simplified motion simulation with the image-based technique considering four positions. What's more, the latter allows corrupting in a simplified way an MRI in 3 seconds where the k-space based approach, which generates a much more complex motion, takes 20 seconds.

We also compared the proposed architecture, Conv5FC3, with the SE-CNN, ResNet and ViT architectures. The BA obtained with the first four networks is comparable: the SE-CNN performance (100.00 ± 0.00) was slightly higher than the ResNet (99.72 ± 0.03), the

TABLE IV
DETECTION OF MOTION ARTEFACTS WITHIN BRAIN T1W MR IMAGES OF THE CDW. FOR BOTH THE DETECTION OF SEVERE MOTION (MOV01VS2) AND MILD MOTION (MOV0VS1), WE REPORT: THE AGREEMENT BETWEEN HUMAN RATERS AND THE CONSENSUS (MANUAL ANNOTATIONS), RESULTS OF THE PROPOSED APPROACH (PRE-TRAINING ON IMAGE-BASED SYNTHETIC MOTION FROM RESEARCH DATA AND FINE-TUNING ON CDW), RESULTS WHEN TRAINING ON IMAGE-BASED SYNTHETIC MOTION FROM RESEARCH DATASETS WITHOUT FINE-TUNING, RESULTS WHEN TRAINING FROM SCRATCH ON CDW, AND RESULTS OF THE RANDOM FOREST TRAINED ON IQMS EXTRACTED FROM THE RESEARCH AND CDW DATA.

		BA	Specificity	Sensitivity
Mov01vs2	Manual annotations	86.29 %	–	–
	Conv5FC3 fine-tuned on CDW (proposed)	84.52 %	85.37 %	83.67 %
	Conv5FC3 trained on research dataset	60.26 %	33.33 %	87.19 %
	Conv5FC3 trained from scratch on CDW	73.75 %	49.58 %	97.93 %
	Random forest trained on research dataset	62.13 %	85.67 %	38.60 %
	Random forest trained on CDW	61.94 %	97.56 %	26.32 %
Mov0vs1	Manual annotations	73.21 %	–	–
	Conv5FC3 fine-tuned on CDW (proposed)	62.61 %	52.00 %	73.23 %
	Conv5FC3 trained on research dataset	53.18 %	17.96 %	88.57 %
	Conv5FC3 trained from scratch on CDW	58.93 %	28.81 %	89.05 %
	Random forest trained on research dataset	50.06 %	85.71 %	14.40 %
	Random forest trained on CDW	54.72 %	85.71 %	23.72 %

Conv5 FC3 (99.17 ± 0.03) and the ViT (97.69 ± 0.07) architectures. As the performance of the different classifiers was not statistically significantly different ($p > 0.05$, McNemar test), the time needed to train the different networks allowed us to select the best model for our transfer learning framework, namely the Conv5FC3, which needs less than 4 hours to be trained on our research databases for the pre-training task (Table III). By performing the comparison of architectures on an easy task (the detection of synthetic motion in the research dataset), we were not able to highlight the usefulness of more complex models. To analyse their potential impact on the performance, we could explore the optimisation of frozen layers during the fine-tuning step on clinical data for each architecture.

One of the limitations of our study is the pre-processing needed before applying our model, which might prevent its direct application to new data. Our motion detection model was trained using T1w MRIs pre-processed with the `t1-linear` pipeline of Clinica [42]. This pre-processing step includes an affine registration to the MNI space that facilitates the manual annotation for the graders. Such spatial normalisation could also be beneficial when training neural networks as it reduces the variability between the images. Another limitation is the annotation process of the CDW images. As the IT environment is extremely limited and data cannot be downloaded locally, the annotations performed in our previous study had to rely on only three slices (a central slice in each plane). Thus the annotators may have missed some artefacts [6]. Finally, it would be interesting to study in the future the potential generalisation of such QC models to other MRI sequences available in CDWs, such as FLAIR.

VI. CONCLUSION

In this study, we proposed a transfer learning framework for the automatic detection of motion artefacts of 3D T1w brain MRI which was validated on a large clinical data warehouse. We trained and validated different CNNs on three research datasets using motion simulation and we successfully tested them on an independent test set with synthetic and real motion. We were able to generalise our pre-trained model to clinical images thanks to the motion labelling of 4045 MRIs. Our deep learning classifier was almost as reliable as manual rating for the detection of severe motion artefacts. Our work demonstrated the usefulness of synthetic motion to improve the detection of motion artefacts in MRI, as well as the crucial need of transfer learning to generalise model trained on research to routine clinical data.

APPENDIX

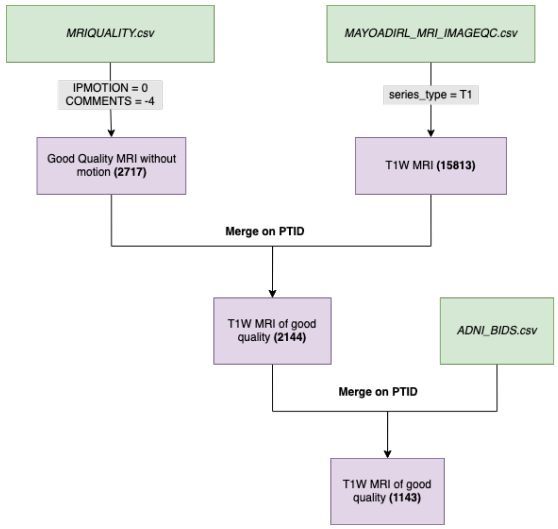

Fig. A.1. Selection process for ADNI MRIs based on the IPMOTION and on the comments section. MRIQUALITY.csv and MAYODRIL_MRI_IMAGEQC.csv are provided by ADNI whereas ADNI_BIDS.csv was obtained with the Clinica software [42]. The ADNI Participant ID (PTID) consists of 3-digit site number, a single character identifier, followed by a sequential 4-digit subject number reflecting the chronological order in which the ID's are assigned across sites.

TABLE A.I
DISTRIBUTION OF THE TRAINING, VALIDATION AND TEST SETS SEPARATELY FOR THE ADNI, MSSEG AND MNI BITE COMPOSING OUR RESEARCH DATASET.

	Label	ADNI	MSSEG	MNI BITE	Total
Training	Motion	400	19	9	428
	No motion	400	19	9	428
Validation	Motion	86	4	2	92
	No motion	86	4	2	92
Testing	Motion	85	3	2	90
	No motion	86	4	2	92

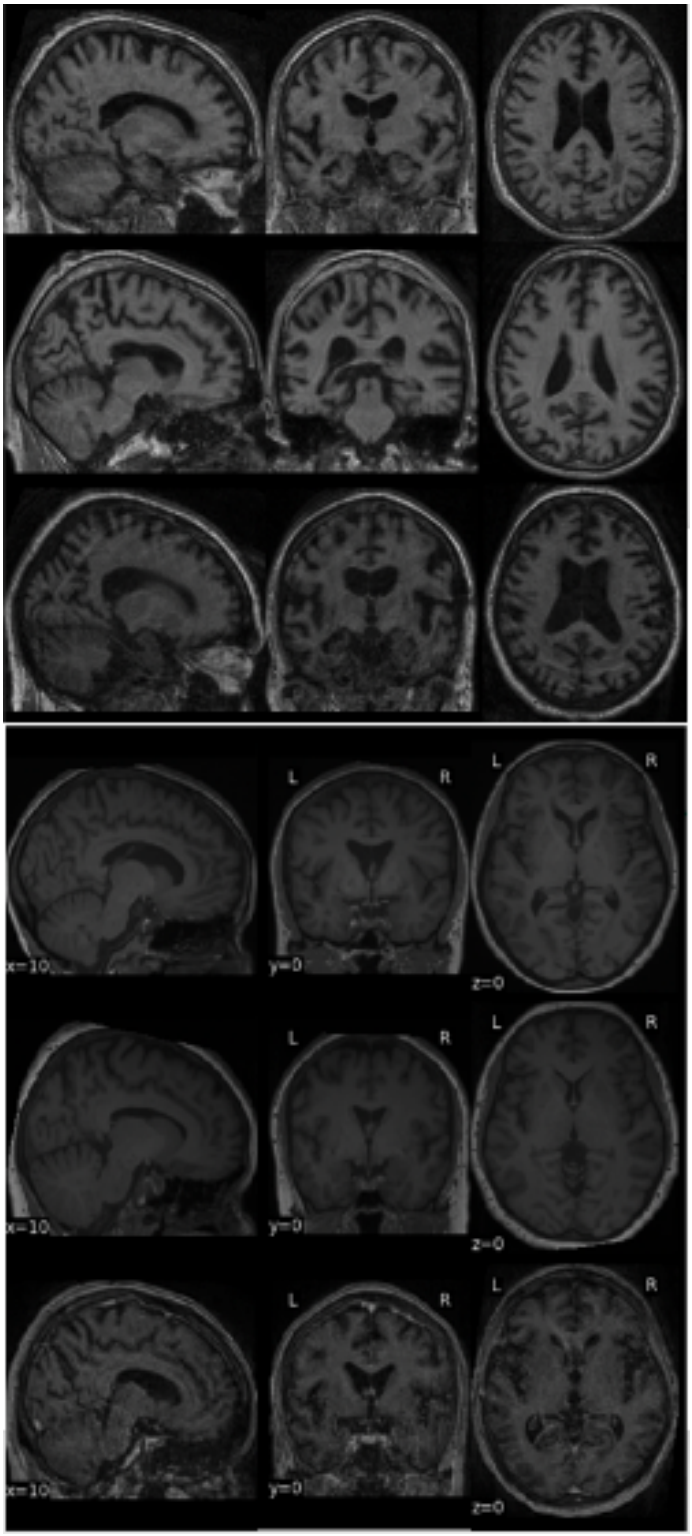

Fig. A.2. Top : Example of three T1w MRI with "real" motion from the ADNI dataset. All MRIs were labelled with an IPMOTION of 3. Bottom : Example of three T1w MRI from the AP-HP CDW. From top to bottom: motion-free MRI (mov0), mild motion (mov1), severe motion (mov2).

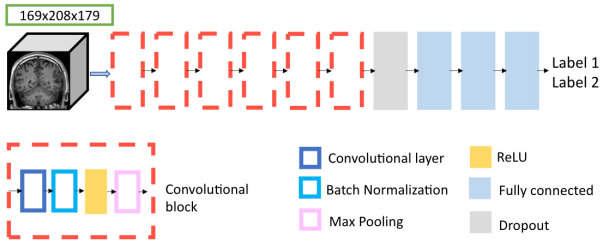

Fig. A.3. Architecture of the 3D CNN called Conv5 FC3. Five convolutional blocks (composed sequentially of a convolutional layer, a batch normalisation layer, a ReLU and a max pooling layer) are followed by a dropout and three fully connected layers.

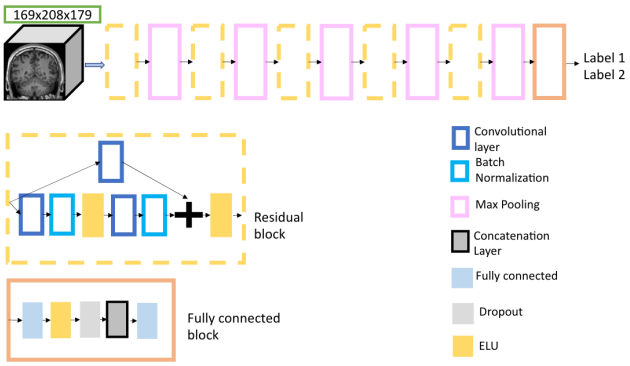

Fig. A.4. Architecture of the ResNet 3D CNN.

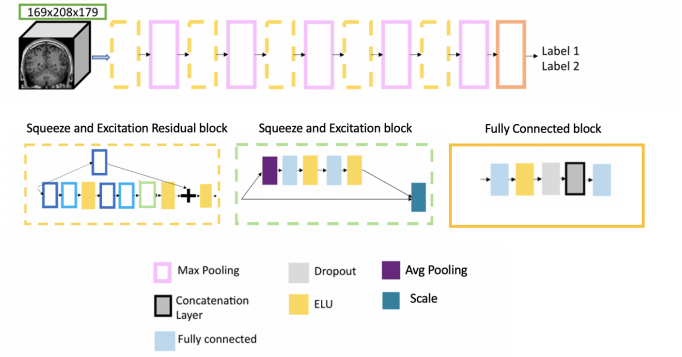

Fig. A.5. Architecture of the SECNN with the Squeeze and Excitation Residual blocks.

ACKNOWLEDGMENT

The research was done using the Clinical Data Warehouse of the Greater Paris University Hospitals. The authors are grateful to the members of the AP-HP WIND and URC teams, and in particular Stéphane Bréant, Florence Tubach, Jacques Ropers, Antoine Rozès, Camille Nevoret, Christel Daniel, Martin Hilka, Yannick Jacob, Julien Dubiel, Cyrina Saussol and Rafael Gozlan. They would also like to thank the "Collégiale de Radiologie of AP-HP" as well as, more generally, all the radiology departments from AP-HP hospitals. The authors would also like to thank Romain Valabregue and Ghiles Reguig for their help implementing the k-space motion simulation and their feedback.

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

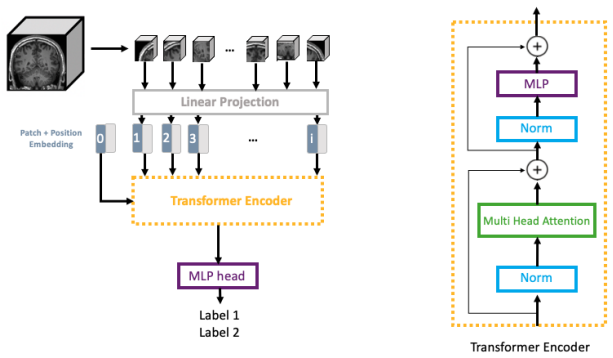

Fig. A.6. Architecture of the Vision Transformer (ViT) with the transformer encoder introduced by [52].

TABLE A.II
DISTRIBUTION OF THE TRAINING VALIDATION AND TEST SETS SEPARATELY FOR THE TWO FINE-TUNING TASKS ON THE AP-HP CDW: THE DETECTION OF SEVERE (MOV01VS2) AND MODERATE (MOV0VS1) MOTION.

	Label	N° of images	
		Mov0vs1	Mov01vs2
Training	Motion 0	1681	1681
	Motion 1	859	859
	Motion 2	-	379
Validation	Motion 0	428	428
	Motion 1	219	219
	Motion 2	-	94
Testing	Motion 0	210	210
	Motion 1	118	118
	Motion 2	-	57

TABLE A.III
RESULTS OF FOUR DIFFERENT CNN CLASSIFIERS (CONV5 FC3, RESNET, SE-CNN AND VIT) TRAINED AND TESTED ON MRIS FROM OUR RESEARCH DATASET CORRUPTED WITH IMAGE-BASED MOTION SIMULATION. WE REPORT THE PERFORMANCE OBTAINED BY THE BEST MODEL OF THE 5-FOLD CV.

Architectures	Balanced Accuracy	Specificity	Sensitivity
Conv5FC3	100	100	100
ResNet	100	100	100
SE-CNN	99.21	98.41	100
ViT	96.88	98.61	95.16

from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; 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 (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

REFERENCES

[1] B. Nordlinger, C. Villani, and D. Rus, *Healthcare and Artificial Intelligence*. Springer International Publishing, 2020.

[2] M. Karami, A. Rahimi, and A. H. Shahmirzadi, "Clinical data warehouse: an effective tool to create intelligence in disease management," *The Health Care Manager*, vol. 36, no. 4, pp. 380–384, 2017.

[3] M. R. Mia, A. S. M. L. Hoque, S. I. Khan, and S. I. Ahamed, "A privacy-preserving National Clinical Data Warehouse: Architecture and analysis," *Smart Health*, vol. 23, p. 100238, 2022.

[4] A.-S. Jannot, E. Zapletal, P. Avillach, M.-F. Mamzer, A. Burgun, and P. Degoulet, "The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience," *Int. J. Med. Inform.*, vol. 102, pp. 21–28, 2017.

[5] M. L. Wood and R. M. Henkelman, "Truncation Artifacts in Magnetic Resonance Imaging," *Magn Reson Med*, 1985.

[6] S. Bottani, N. Burgos, A. Maire, A. Wild, S. Strer, D. Dormont, and O. Colliot, "Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse," *Med. Image Anal.*, p. 102219, 2021.

[7] J. B. Andre, B. W. Bresnahan, M. Mossa-Basha, M. N. Hoff, C. P. Smith, Y. Anzai, and W. A. Cohen, "Toward Quantifying the Prevalence, Severity, and Cost Associated With Patient Motion During Clinical MR Examinations," *J. Am. Coll. Radiol.*, vol. 12, no. 7, pp. 689–695, 2015.

[8] E. P. Hedges, M. Dimitrov, U. Zahid, B. Brito Vega, S. Si, H. Dickson, P. McGuire, S. Williams, G. J. Barker, and M. J. Kempton, "Reliability of structural MRI measurements: The effects of scan session, head tilt, inter-scan interval, acquisition sequence, FreeSurfer version and processing stream," *NeuroImage*, vol. 246, p. 118751, 2022.

[9] M. Reuter, M. D. Tisdall, A. Qureshi, R. L. Buckner, A. J. W. van der Kouwe, and B. Fischl, "Head motion during MRI acquisition reduces gray matter volume and thickness estimates," *NeuroImage*, vol. 107, pp. 107–115, 2015.

[10] A. Alexander-Bloch, L. Clasen, M. Stockman, L. Ronan, F. Lalonde, J. Giedd, and A. Raznahan, "Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI," *Hum. Brain Mapp.*, vol. 37, no. 7, pp. 2385–2397, 2016.

[11] B. Fischl, "FreeSurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, 2012.

[12] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[13] O. Esteban, D. Birman, M. Schaer, O. O. Koyejo, R. A. Poldrack, and K. J. Gorgolewski, "MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites," *PLoS ONE*, vol. 12, no. 9, p. e0184661, 2017.

[14] A. R. Sadri, A. Janowczyk, R. Zou, R. Verma, N. Beig, J. Antunes, A. Madabhushi, P. Tiwari, and S. E. Viswanath, "MRQy: An Open-Source Tool for Quality Control of MR Imaging Data," *Med. Phys.*, vol. 47, no. 12, pp. 6029–6038, 2020.

[15] K. Lei, A. B. Syed, X. Zhu, J. M. Pauly, and S. S. Vasanawala, "Artifact- and content-specific quality assessment for MRI with image rulers," *Med. Image Anal.*, vol. 77, p. 102344, 2022.

[16] D. Ravi, F. Barkhof, D. C. Alexander, G. J. Parker, and A. Eshaghi, "An efficient semi-supervised quality control system trained using physics-based MRI-artefact generators and adversarial training," arXiv:2206.03359, 2022.

TABLE A.IV
FINE-TUNING OF THE CONV5FC3 MODEL PRE-TRAINED ON OUR RESEACH DATASET FOR THE DETECTION OF SEVERE MOTION (MOV01VS2). FIVE EXPERIMENTS WERE PERFORMED BY UNFREEZING ONE TO FIVE LAYERS OF THE CONV5FC3 ARCHITECTURE. THE SAME MODEL PRE-TRAINED ON THE RESEARCH DATASET WITH SYNTHETIC IMAGE-BASED MOTION WAS USED FOR ALL THE EXPERIMENTS.

Architecture	Unfreezed Layers	Balanced Accuracy	Specificity	Sensitivity
Conv5FC3	FC3	64.81	63.16	66.46
	FC3 ; FC2	65.84	64.91	66.77
	FC3 ; FC2 ; FC1	**84.52**	**85.37**	**83.67**
	FC3 ; FC2 ; FC1 ; Conv5	75.53	92.68	58.37
	FC3 ; FC2 ; FC1 ; Conv5 ; Conv4	60.61	52.63	68.60

TABLE A.V
DETECTION OF MOTION ARTEFACTS WITHIN BRAIN T1W MR IMAGES OF THE CDW. FOR BOTH THE DETECTION OF SEVERE MOTION (MOV01VS2) AND MILD MOTION (MOV0VS1), WE REPORT: THE AGREEMENT BETWEEN HUMAN RATERS AND THE CONSENSUS (MANUAL ANNOTATIONS), RESULTS OF THE TRANSFER-LEARNING APPROACH (PRE-TRAINING ON K-SPACE BASED MOTION SIMULATION FROM RESEARCH DATA AND FINE-TUNING ON CDW), RESULTS WHEN TRAINING ON SYNTHETIC MOTION, USING THE K-SPACE BASED APPROACH, FROM RESEARCH DATASETS WITHOUT FINE-TUNING AND RESULTS WHEN TRAINING FROM SCRATCH ON CDW.

		BA	Specificity	Sensitivity
Mov01vs2	Manual annotations	86.29 %	–	–
	Conv5FC3 fine-tuned on CDW	80.60 %	77.00 %	84.21 %
	Conv5FC3 trained on research dataset	60.60 %	24.56 %	96.64 %
	Conv5FC3 trained from scratch on CDW	73.75 %	49.58 %	97.93 %
Mov0vs1	Manual annotations	73.21 %	–	–
	Conv5FC3 fine-tuned on CDW	60.02 %	38.13 %	81.90 %
	Conv5FC3 trained on research dataset	52.15 %	7.63 %	96.67 %
	Conv5FC3 trained from scratch on CDW	58.31 %	33.39 %	83.24 %

[17] R. Shaw, C. H. Sudre, S. Ourselin, M. J. Cardoso, and H. G. Pemberton, "A Decoupled Uncertainty Model for MRI Segmentation Quality Estimation," arXiv:2109.02413, 2021.

[18] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image Anal.*, vol. 12, no. 1, pp. 26–41, 2008.

[19] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.

[20] S. J. Sujit, I. Coronado, A. Kamali, P. A. Narayana, and R. E. Gabr, "Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks," *J. Magn. Reson. Imaging*, vol. 50, no. 4, pp. 1260–1267, 2019.

[21] I. Fantini, C. Yasuda, M. Bento, L. Rittner, F. Cendes, and R. Lotufo, "Automatic mr image quality evaluation using a deep cnn: A reference-free method to rate motion artifacts in neuroimaging," *Comput. Med. Imaging Graph.*, vol. 90, p. 101897, 2021.

[22] T. Küstner, A. Liebgott, L. Mauch, P. Martirosian, F. Bamberg, K. Nikolaou, B. Yang, F. Schick, and S. Gatidis, "Automated reference-free detection of motion artifacts in magnetic resonance images," *Magn. Reson. Mater. Phy.*, vol. 31, no. 2, pp. 243–256, 2018.

[23] J. E. Iglesias, G. Lerma-Usabiaga, L. C. Garcia-Peraza-Herrera, S. Martinez, and P. M. Paz-Alonso, "Retrospective head motion estimation in structural brain MRI with 3D CNNs," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 314–322.

[24] M. Mohebbian, E. Walia, M. Habibullah, S. Stapleton, and K. A. Wahid, "Classifying MRI motion severity using a stacked ensemble approach," *J. Magn. Reson. Imaging*, vol. 75, pp. 107–115, 2021.

[25] K. Pawar, Z. Chen, N. J. Shah, and G. F. Egan, "Suppressing motion artefacts in MRI using an Inception-ResNet network with motion simulation augmentation," *NMR Biomed.*, vol. 35, no. 4, p. e4225, 2022.

[26] R. Shaw, C. Sudre, S. Ourselin, and M. J. Cardoso, "MRI k-space motion artefact augmentation: model robustness and task-specific uncertainty," in *Medical Imaging with Deep Learning*, 2018, p. 10.

[27] B. A. Duffy, W. Zhang, H. Tang, L. Zhao, M. Law, A. W. Toga, and H. Kim, "Retrospective correction of motion artifact affected structural MRI images using deep learning of simulated motion," in *Medical Imaging with Deep Learning*, 2018, p. 8.

[28] S. Loizillon, S. Bottani, A. Maire, S. Ströer, D. Dormont, O. Colliot, and N. Burgos, "Transfer learning from synthetic to routine clinical data for motion artefact detection in brain T1-weighted MRI," in *SPIE Medical Imaging 2023*, 2023, p. 6.

[29] S. Lee, S. Jung, K.-J. Jung, and D.-H. Kim, "Deep Learning in MR Motion Correction: A Brief Review and a New Motion Simulation Tool (view2Dmotion)," *Investigative Magnetic Resonance Imaging*, vol. 24, no. 4, p. 196, 2020.

[30] F. Pérez-García, R. Sparks, and S. Ourselin, "TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Comput. Meth. Prog. Bio.*, vol. 208, p. 106236, 2021.

[31] A. Loktyushin, H. Nickisch, R. Pohmann, and B. Schölkopf, "Blind retrospective motion correction of MR images," *J. Magn. Reson. Imaging*, vol. 70, no. 6, pp. 1608–1618, 2013.

[32] M. A. Al-masni, S. Lee, J. Yi, S. Kim, S.-M. Gho, Y. H. Choi, and D.-H. Kim, "Stacked U-Nets with self-assisted priors towards robust correction of rigid motion artifact in brain MRI," *NeuroImage*, vol. 259, p. 119411, 2022.

[33] I. Oksuz, "Brain MRI artefact detection and correction using convolutional neural networks," *Comput. Meth. Prog. Bio.*, vol. 199, p. 105909, 2021.

[34] H. Sagawa, K. Itagaki, T. Matsushita, and T. Miyati, "Evaluation of motion artifacts in brain magnetic resonance images using convolutional neural network-based prediction of full-reference image quality assessment metrics," *J. Med. Imaging*, vol. 9, no. 1, p. 015502, 2022.

[35] N. Karani, K. Chaitanya, C. Baumgartner, and E. Konukoglu, "A lifelong learning approach to brain MR segmentation across scanners and protocols," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 476–484.

[36] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, and J. E. Iglesias, "SynthSeg: Domain randomisation for segmentation of brain mri scans of any contrast and resolution," arXiv:2107.09559, 2021.

[37] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for med. image anal.: Full training or fine tuning?" *IEEE T. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[38] K. B. Ahmed, L. O. Hall, D. B. Goldgof, R. Liu, and R. A. Gatenby, "Fine-tuning convolutional deep features for MRI based brain tumor classification," in *SPIE Medical Imaging 2017*, vol. 10134, 2017, pp. 613–619.

[39] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga, J. Q. Trojanowski, and M. W. Weiner, "Alzheimer's Disease Neuroimaging Initiative (ADNI)," *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.

[40] O. Commowick, F. Cervenansky, and R. Ameli, Eds., *MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, 2016. [Online]. Available: https://www.hal.inserm.fr/inserm-01397806

[41] L. Mercier, R. F. Del Maestro, K. Petrecca, D. Araujo, C. Haegelen, and D. L. Collins, "Online database of clinical MR and ultrasound images of brain tumors," *Med. Phys.*, vol. 39, no. 6Part1, pp. 3253–3261, 2012.

[42] A. Routier, N. Burgos, M. Díaz, M. Bacci, S. Bottani, O. El-Rifai, S. Fontanella, P. Gori, J. Guillon, A. Guyot, R. Hassanaly, T. Jacquemont, P. Lu, A. Marcoux, T. Moreau, J. Samper-González, M. Teichmann, E. Thibeau-Sutre, G. Vaillant, J. Wen, A. Wild, M.-O. Habert, S. Durrleman, and O. Colliot, "Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies," *Front. Neuroinform.*, vol. 15, p. 689675, 2021.

[43] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: Improved N3 Bias Correction," *IEEE T. Med. Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[44] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, and O. Colliot, "Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation," *Med. Image Anal.*, p. 101694, 2020.

[45] G. Reguig, M. Lapert, S. Lehericy, and R. Valabregue, "Global displacement induced by rigid motion simulation during MRI acquisition," arXiv:2204.03522, 2022.

[46] E. Thibeau-Sutre, M. Díaz, R. Hassanaly, A. Routier, D. Dormont, O. Colliot, and N. Burgos, "ClinicaDL: An open-source deep learning software for reproducible neuroimaging processing," *Comput. Meth. Prog. Bio*, vol. 220, p. 106818, 2022.

[47] B. A. Jonsson, G. Bjornsdottir, T. E. Thorgeirsson, L. M. Ellingsen, G. B. Walters, D. F. Gudbjartsson, H. Stefansson, K. Stefansson, and M. O. Ulfarsson, "Brain age prediction using deep learning uncovers associated sequence variants," *Nature Communications*, vol. 10, no. 1, p. 5409, 2019.

[48] B. Couvy-Duchesne, J. Faouzi, B. Martin, E. Thibeau–Sutre, A. Wild, M. Ansart, S. Durrleman, D. Dormont, N. Burgos, and O. Colliot, "Ensemble Learning of Convolutional Neural Network, Support Vector Machine, and Best Linear Unbiased Predictor for Brain Age Prediction: ARAMIS Contribution to the Predictive Analytics Competition 2019 Challenge," *Front. Psychiatry*, vol. 11, 2020.

[49] P. Ghosal, L. Nandanwar, S. Kanchan, A. Bhadra, J. Chakraborty, and D. Nandi, "Brain Tumor Classification Using ResNet-101 Based Squeeze and Excitation Deep Neural Network," in *International Conference on Advanced Computational and Communication Paradigms*, 2019, pp. 1–6.

[50] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal Brain Tumor Segmentation Using Transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 109–119.

[51] X. Xing, G. Liang, Y. Zhang, S. Khanal, A.-L. Lin, and N. Jacobs, "ADViT: Vision Transformer on Multi-Modality PET Images for Alzheimer Disease Diagnosis," in *IEEE International Symposium on Biomedical Imaging*, 2022, pp. 1–4.

[52] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv:2010.11929, 2020.