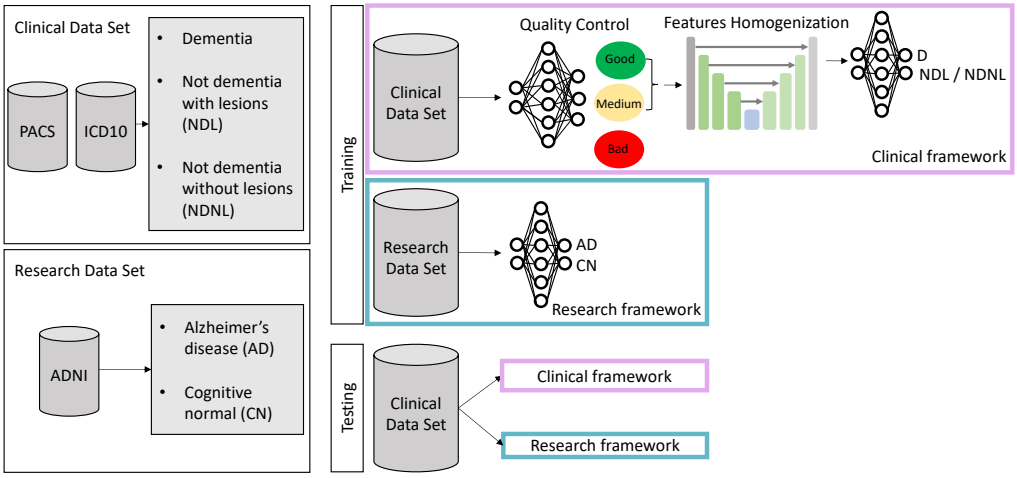
Graphical Abstract

Evaluation of MRI-based machine learning approaches for computer-aided diagnosis of dementia in a clinical data warehouse

Simona Bottani, Ninon Burgos, Aurélien Maire, Dario Saracino, Sebastian Ströer, Didier Dormont, Olivier Colliot, for the Alzheimer's Disease Neuroimaging Initiative[1], and the APPRIMAGE Study Group



[1]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`.

Highlights

Evaluation of MRI-based machine learning approaches for computer-aided diagnosis of dementia in a clinical data warehouse

Simona Bottani, Ninon Burgos, Aurélien Maire, Dario Saracino, Sebastian Ströer, Didier Dormont, Olivier Colliot, for the Alzheimer's Disease Neuroimaging Initiative[2], and the APPRIMAGE Study Group

- We studied the performance of machine learning models for the detection of dementia using anatomical brain MRI on real-life clinical routine data.
- We used images coming from a clinical data warehouse and we identified the population of interest using the 10^{th} revision of the International Classification of Diseases (ICD-10).
- We uncovered that the performance of the classifier is mainly driven by irrelevant characteristics thereby biasing the performance upwards, a phenomenon known as the Clever Hans effect or shortcut learning.
- The performance was considerably lower on real-life clinical routine data compared with that obtained on research data.

[2]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`.

- Our work demonstrates the difficulty of translating computer-aided diagnosis algorithms to clinical routine.

Evaluation of MRI-based machine learning approaches for computer-aided diagnosis of dementia in a clinical data warehouse

Simona Bottani[a], Ninon Burgos[a], Aurélien Maire[b], Dario Saracino[a,c], Sebastian Ströer[d], Didier Dormont[d,e], Olivier Colliot[a], for the Alzheimer's Disease Neuroimaging Initiative[1], and the APPRIMAGE Study Group

[a]*Sorbonne Université, Institut du Cerveau – Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, Paris, 75013, France*
[b]*AP-HP, WIND department, Paris, 75012, France*
[c]*IM2A, Reference Centre for Rare or Early-Onset Dementias, Département de Neurologie, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, 75013, France*
[d]*AP-HP, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, Paris, 75013, France*
[e]*Sorbonne Université, Institut du Cerveau – Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, DMU DIAMENT Paris, 75013, France*

Abstract

A variety of algorithms have been proposed for computer-aided diagnosis of dementia from anatomical brain MRI. These approaches achieve high accuracy when applied to research data sets but their performance on real-life clinical routine data has not been evaluated yet. The aim of this work was to study the performance of such approaches on clinical routine data, based on a hospital data warehouse, and to compare the results to those obtained on a

[1]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`.

research data set. The clinical data set was extracted from the hospital data warehouse of the Greater Paris area, which includes 39 different hospitals. The research set was composed of data from the Alzheimer's Disease Neuroimaging Initiative data set. In the clinical set, the population of interest was identified by exploiting the diagnostic codes from the 10^{th} revision of the International Classification of Diseases that are assigned to each patient. We studied how the imbalance of the training sets, in terms of contrast agent injection and image quality, may bias the results. We demonstrated that computer-aided diagnosis performance was strongly biased upwards (over 17 percent points of balanced accuracy) by the confounders of image quality and contrast agent injection, a phenomenon known as the Clever Hans effect or shortcut learning. When these biases were removed, the performance was very poor. In any case, the performance was considerably lower than on the research data set. Our study highlights that there are still considerable challenges for translating dementia computer-aided diagnosis systems to clinical routine.
Keywords: Clinical Data Warehouse, Dementia, MRI, Neuroimaging, Deep Learning, Shortcut learning

1. Introduction

Dementia is a world-wide syndrome that is becoming more and more important due to population aging. T1-weighted (T1w) brain magnetic resonance imaging (MRI) contributes to the positive diagnosis of dementia by

displaying typical spatial patterns of brain atrophy. A variety of computer-aided diagnosis (CAD) systems using T1w brain MRI data have been developed using machine learning and deep learning (Klöppel et al., 2008; Vemuri et al., 2008; Fan et al., 2008; Gerardin et al., 2009; Cuingnet et al., 2011; Rathore et al., 2017; Wen et al., 2020; Burgos et al., 2021).

So far, CAD systems have been mainly developed and validated using research data sets due to their ease of access (many can directly be downloaded from websites). Several data sets originating from research studies such as the Alzheimer's Disease Neuroimaging Initiative (ADNI)[2], the Open Access Series Of Imaging Studies (OASIS)[3], the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL)[4], and the Frontotemporal lobar degeneration neuroimaging initiative (NIFD)[5] are publicly available and contain various clinical and imaging data, including T1w MRI brain images. They have pushed the research on machine learning and deep learning for CAD using neuroimages: previously published works focusing on Alzheimer's disease (AD) have exploited the ADNI, OASIS or AIBL data sets (Punjabi et al., 2019; Bidani et al., 2019; Spasov et al., 2019; Böhle et al., 2019; Farooq et al., 2017; Wegmayr et al., 2018; Samper-González et al., 2018; Wen et al., 2020; Bron et al., 2021; Cuingnet et al., 2011; Hinrichs et al., 2009; Chupin et al., 2009; Misra et al., 2009), whereas those targeting fronto-temporal

[2] `http://adni.loni.usc.edu/`
[3] `https://www.oasis-brains.org/`
[4] `https://aibl.csiro.au/`
[5] `https://ida.loni.usc.edu/home/projectPage.jsp?project=NIFD`

dementia (FTD) used NIFD (Ma et al., 2020).

Even if all these data sets have proven extremely useful to propel methodological research on machine learning applied to neurological diseases, they are far from the everyday clinical routine for two main reasons. First, in many works, the aim is to differentiate patients with a particular, well-characterised, disease (most often AD), from healthy controls. Such homogeneous diagnostic classes do not reflect the reality of clinical routine. Some works focused on differential diagnosis between different types of dementia but they still use research data sets: Ma et al. (2020) classified patients with AD and FTD using ADNI and NIFD, Koikkalainen et al. (2016) differentiated AD, FTD, dementia with Lewy bodies and vascular dementia using the Amsterdam Dementia Cohort. Second, research images are usually acquired following a standardized protocol whose aim is to guarantee data quality and homogenization. This is obviously not the case in clinical routine.

In order to bring research advances to clinical practice, various groups, including our own, have developed and validated CAD systems using clinical data sets (Morin et al., 2020; Chagué et al., 2021; Platero et al., 2019; Sohn et al., 2015; Klöppel et al., 2015). Nevertheless, the participants, even though the MRI was indeed acquired as part of clinical workup, were retrospectively selected to fit a well defined task of interest. The images were also filtered to remove low quality images. Moreover, the data come from highly specialized centers that are not representative of the overall clinical practice (for instance rare dementias and early-onset cases are over-represented). Furthermore, the

data often come from a single or few hospitals, thus they may not reflect the full spectrum of heterogeneity. Finally, they often restrict themselves to diagnosis of patients with dementia. It is thus unclear what their specificity is when dealing with MRI from patients with other diagnoses. Therefore, the performance reported cannot be considered to reflect those that would be obtained on real-life data.

Clinical data warehouses (CDW), which gather all images acquired in large groups of hospitals, are a better representation of clinical routine and they are thus an important tool for the translation of research to the clinic. Images of a CDW are heterogeneous (i.e. different sites, MRI sequences not harmonized) and they include a very wide range of diagnoses (including not only patients with dementia but also patients with other neurological or psychiatric diseases, as well as patients who underwent a brain MRI for another indication) (Bottani et al., 2022a; Wood et al., 2022).

The aim of this work is to experimentally study the performance of machine learning methods to detect dementia patients in a CDW using T1w brain MRI. Patients with dementia were labeled using diagnostic codes assigned during the hospitalization period. The main machine learning model was a linear support vector machine using gray matter maps as features. It was then compared to several deep learning models. We compared the performance obtained on a research data set to that obtained on the present clinical data set. We studied how results in a clinical data set may be biased by the characteristics of the training data set (in particular by the injection

of gadolinium and the presence of images of different quality). We used an image translation approach to change the appearance of images for which gadolinium was injected in order to mitigate bias associated to this factor.

2. Materials

To compare the performance of CAD systems to detect dementia in a research and a clinical setting, two data sets were used.

2.1. Research data set

The research data set used in this work was composed of subjects from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease.

We considered subjects from ADNI 1/2/Go/3 diagnosed as cognitive normal (CN) or AD at baseline and only kept subjects whose diagnosis did not change over time. This resulted in 800 subjects with a T1w MR image at the first session including imaging data (CN: 410 subjects, 54.87 % F, age 73.20 $\pm$ 6.15 in range [55.1, 89.6]; AD: 390 subjects, 44.0 % F, age 74.88 $\pm$ 7.76 in range [55.1, 90.1]). Two hundred subjects (100 CN and 100 AD) composed the independent test set and the remaining subjects (310 CN

and 290 AD) were used for the training/validation of the models using a 5-fold cross-validation (CV).

2.2. Clinical routine data set

The clinical data set comes from the data warehouse of AP-HP (Assistance Publique-Hôpitaux de Paris) which represents data from 39 hospitals of the Greater Paris area (Daniel and Salamanca, 2020). The study was approved by the Ethical and Scientific Board of the AP-HP data warehouse. All details regarding the ethics approval and the procedure and regulations allowing the access and use of patient data for research purposes are described in Supplementary material section 1. The data were only accessed within the AP-HP network and it was strictly forbidden to export any kind of data.

All the data, both imaging and clinical, were pseudonymized by the AP-HP data warehouse and they always remained within the hospital network. The DICOM were pseudonymized as follows: information about the patient such as name, age, sex, weight as well as information about the physicians who requested and analysed the results of the examination are erased, the examination date is shifted of a random amount of time (from 1 to 10 years). Note that, for a given patient, the same shift is applied to the examination date and to the date of birth (part of clinical data in the ORBIS system, see below). As the age is calculated as the difference between these two dates, this pseudonymization process does not affect the computation of the

age. The images were not defaced. However, the identification from the images would be very difficult because no 3D image viewer (with 3D rendering or mutiple plane visualization) was available within the platform. Only a JupyterLab instance was available (Bottani et al., 2022a). In order to visualize a snapshot of the image on a Jupyter Notebook, we developed a tool available at: (`https://github.com/SimonaBottani/image_synthesis`, commit number 98710ed).

2.2.1. Imaging and clinical data collection

Images from this clinical data warehouse are very heterogeneous (Bottani et al., 2022a): they include 3D T1w brain MR images of patients with a wide range of ages (from 18 to more than 90 years old) and diseases, acquired with different scanners (more than 30 different models). Imaging data were gathered in a central hospital picture archiving and communication system (PACS) and images relevant to our research project were copied to the research PACS where they were pseudonymized. The selection process to obtain images of interest is described in (Bottani et al., 2022a): a neuroradiologist manually selected all the DICOM header attributes (in particular the acquisition protocol, the series description and the body part) referring to a 3D brain T1w MRI.

At the same time, clinical data corresponding to the patients of our query are stored in a database managed by the ORBIS clinical information system. Clinical data gather all the information connected to the patients, i.e. date of

birth, sex, diagnostic codes, medications, biological tests, electronic health reports. As explained in (Daniel and Salamanca, 2020), ORBIS has been installed progressively in the AP-HP hospitals since 2009. Among all the patients aged more than 18 years old who undertook a 3D T1w brain MRI examination at AP-HP ($\sim$130.000 patients), only $\sim$25% were registered in ORBIS. Among them, 23,688 patients were hospitalized. Note that for non-hospitalized patients, only sociodemographic data (sex and age) are available and not clinical data. As for the imaging data, the data warehouse provided the pseudonymized clinical data.

For our work, we were interested in two sociodemographic items (age and sex) as well as one clinical item (diagnostic codes). Codes from the 10^{th} revision of the International Classification of Diseases (ICD-10) (World Health Organization et al., 2007) were used to associate a diagnosis to each T1w brain MRI. Images were labeled according to the ICD-10 codes assigned to the visit corresponding to the acquisition of the image. We defined a visit as a period of plus or minus three months from the acquisition date of the image. As clinical data can be entered by the medical staff at different moments during hospitalization, this time window ensures that all pieces of information regarding brain disorders related to the need of a brain MRI exam are collected.

In conclusion, the initial clinical data set of interest was composed of 23,688 patients, which corresponds to 32,348 visits and 43,418 3D T1w brain MR images.

2.2.2. Definition of the different diagnostic categories from ICD-10 codes

On average, 60 ICD-10 codes were assigned to each visit. Since we did not know the reason of a patient's hospitalization (which may be different from the reason why they were prescribed an MRI examination), we considered principal diagnoses, secondary diagnoses and comorbidities at the same level. First, we identified all the ICD-10 codes that could refer to dementia (denoted as D). Note that we use the term "dementia" in a broad sense, i.e. we consider mild cognitive impairment as belonging to this category. However, we restricted this category to the two most common causes of dementia (i.e. neurodegenerative and vascular dementias) and we did not include the more atypical causes such as dementia in HIV disease (F02.4) or psychotic disorder due to alcohol (F10.7), or dementia whose cause was undefined (F03).

Then, we divided the remaining codes into two groups: ICD-10 codes referring to diseases (for instance cancer, demyelinating diseases, stroke, hydrocephalus) that lead to lesions that visibily alter T1w brain MRI (referred to as "no dementia but with lesions" - NDL) and ICD-10 codes corresponding to diseases that, in principle, do not lead to lesions visibly altering T1w brain MRI (referred to as "no dementia and no lesions" - NDNL). We considered two different classification tasks in which dementia patients had to be differentiated from these two classes (NDL and NDNL), which have very different characteristics.

In Table 1, we list the three classes mentioned above (D, NDL, NDNL). For each of them, we provide a brief description and a list of all the associ-

ated ICD-10 codes. Sixteen diseases were associated to the category dementia. Four families of diseases were associated to the NDL category (which are defined by grouping different ICD-10 codes). The NDNL category corresponded to all the other codes. According to the standard structure of the ICD-10 codes, we considered just the first letter and the first two numbers, indicating the category, to identify the diseases belonging to the NDL category. The third number, indicating the etiology, was used to identify the diseases corresponding to the dementia category as we wanted to be more specific.

2.2.3. Selection of patients belonging to the dementia category

Dementia is the principal category we consider since our aim is to study how well this category can be distinguished from the others. We thus started by selecting patients labeled as dementia. In the workflow displayed in Figure 1 we report the different choices made to create this population. For each step, we report the number of patients, visits and images.

Starting from 2441 patients with at least one ICD-10 code in the dementia category, corresponding to 2671 visits and 3633 images (considering only 3D T1w brain MRI), the final population is composed of 1255 patients, corresponding to 1255 visits and 1415 images. We first excluded patients that had multiple ICD-10 codes belonging to the dementia category at the same visit to have a unique label per visit. We then excluded patients with an ICD-10 code belonging to the NDL category with the aim that lesions visible on T1w

Table 1: Description of the three categories of interest with the corresponding ICD-10 codes. Details about dementia codes: "/" indicates that the two codes refer to the same diagnosis, "+" means that the diagnosis of dementia is defined by the presence of both codes. ".*" in the NDL category indicates that all the sub-categories of the code were considered.

Category	**ICD-10 codes**
D: Dementia associated to a neurodegenerative disease or a vascular disease that causes atrophy visible on T1w MRI.	• Dementia in AD with early onset (F00.0/G30.0) • Dementia in AD with late onset (F00.1/G30.1) • Dementia in AD, atypical or mixed type (F00.2/G30.8) • Dementia in AD, unspecified (F00.9/G30.9) • Dementia in Pick disease (F02.0/G31.0) • Dementia in Creutzfeldt-Jakob disease (F02.1/A81.0) • Dementia in Huntington disease (F02.2 + G10) • Vascular dementia of acute onset (F01.0) • Multi-infarct dementia (F01.1) • Subcortical vascular dementia (F01.2) • Mixed cortical & subcortical vascular dementia (F01.3) • Other vascular dementia (F01.8) • Vascular dementia, unspecified (F01.9) • Mild cognitive disorder (F06.7) • Dementia in Parkinson's disease (F02.3 + G20) • Lewy bodies dementia (G02.8 + G31.8)
NDL: No dementia but diagnosis that suggests presence of lesions that modify the anatomical structure of the brain visible on T1w MRI.	• Cancer (C70.*, C71.*, C72.*, D32.*, D33.*, D42.*) • Demyelination (G35.*, G36.*, G37.*) • Stroke (G45.*, G46.*) • Hydrocephalus (G91.*)
NDNL: No dementia and no diagnosis suggesting the presence of lesions on T1w brain MRI.	All the other codes

brain MRI originate only from dementia. Patients were further excluded if the ICD-10 code in the dementia category was changing over time (i.e. over the different visits) as this may be due to an error in coding. Patients aged more than 90 years old were excluded because there were very few patients above this age across the different diagnostic groups (and thus it was not possible to find patients with the same age/sex). Patients labeled F067 (mild cognitive disorder) aged less than 45 years old were excluded because the diagnosis may correspond to a transient mild cognitive impairment and not to a prodromal stage of dementia. Some images were also excluded after the pre-processing step: if they had less than 40 DICOM slices or if they were labeled as straight reject by the quality control step (see Section 3.1). This pre-processing step was applied to all the images of the different categories.

2.2.4. Selection of patients belonging to the no dementia with lesions (NDL) and no dementia and no lesions (NDNL) categories

The aim of this work is to assess whether patients with dementia can be distinguished from patients with other brain diseases, no matter if these diseases result in the presence (NDL category) or absence (NDNL category) of lesions visible on T1w brain MRI. To define the cohorts for the NDL and NDNL categories, we matched each patient belonging to the dementia category with a patient in the NDL category and with a patient in the NDNL category that had the same age (± 1 year) and sex. We first created the NDL cohort, which is composed of patients with one of these four diseases

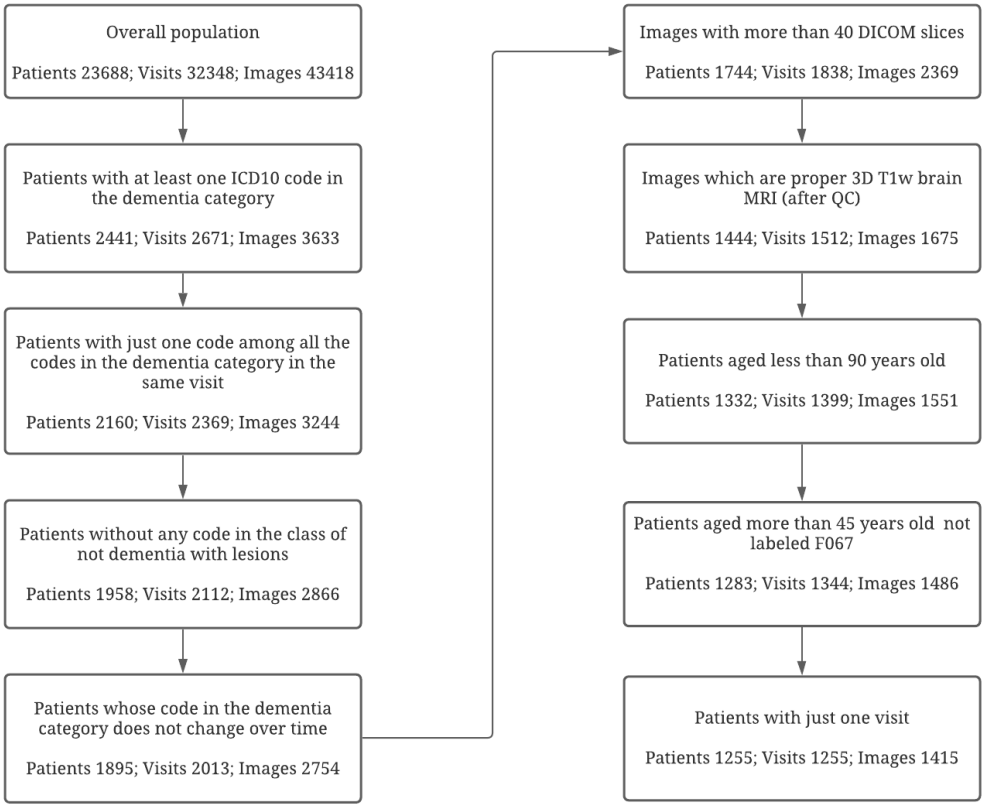


Figure 1: Workflow describing the selection of patients belonging to the dementia category. For each selection step, we report the corresponding number of patients, visits and images.

potentially leading to brain lesions visible on the T1w MRI: cancer, stroke, demyelination and hydrocephalus (see Table 1). We selected all the patients having at least one ICD-10 code in this category, resulting in 3843 patients corresponding to 6598 visits and 9615 images. We then matched these patients with those composing the dementia cohort following several criteria. For each patient with dementia:

- We selected all the patients with the same age (± 1 year) and the same sex having at least one code in the NDL category.
- We excluded all the patients having different NDL codes at the same visit.
- We considered only one visit for each patient when there were multiple visits available with the same diagnosis. The visit was selected randomly.
- Among all the patients with one visit matching these criteria, we randomly selected one of them.

We iterated this selection process twice since some images were discarded after the pre-processing steps (i.e. images with fewer than 40 DICOM slices or flagged as straight reject at the quality control step). In total we matched 808 patients (corresponding to 808 visits and 978 images).

The NDNL class is composed of all the patients having no code in the dementia nor NDL categories. For each patient with dementia:

- We selected all the patients with the same age (± 1 year) and the same sex having no ICD-10 code in the dementia nor NDL categories.
- In case of multiple visits for a patient, we randomly selected one of them.
- Among all the patients with one visit matching these criteria, we randomly selected one of them.

We iterated this selection process twice since some images were discarded after the pre-processing steps. In total we matched 1144 patients (corresponding to 1144 visits and 1343 images).

2.2.5. Final cohorts

The final cohorts were created by taking the intersection of the NDL patients matching with dementia patients and of the NDNL patients matching with dementia patients. This resulted in three cohorts each of 756 patients for a total number of 2268 patients (corresponding to 2268 visits and 2823 images). Note that this number of 756 patients is lower than the initial number of patients in the dementia class because some of them could not be matched for age and sex with a patient of the two other classes.

In Table 2 we report the number of subjects, visits and images for each category. In addition, we report the percentage of females and the average age of the patients as well as the percentage of images with and without injection of gadolinium, and of images of good or medium quality (tier 1/2). The presence of gadolinium and the quality of the images were determined through the automatic approach described in (Bottani et al., 2022a), which will be detailed in the Methods section.

2.2.6. Training, validation and testing subsets

Before starting the experiments, we defined a test set by randomly selecting 20% of the patients of the dementia class and the corresponding matched patients of the other two classes (NDL and NDNL). While for the train-

ing/validation set, if there were several images at the same visit all were kept to increase the number of training samples, for the test set, we selected only one image per visit (the selection was made randomly). This resulted in a test set composed of 152 patients/images for each of the three classes (D, NDL, NDNL). The training/validation set was composed of 604 patients and 719 images for D, 604 patients and 799 images for NDL, 604 patients and 756 images for NDNL.

We respected the same distribution of image quality and presence of gadolinium between the test and the training/validation sets. We also checked that the distribution of the ICD-10 codes between the test and the training/validation sets among the dementia and NDL categories was the same.

For each task, the images of the training/validation set were further split using a 5-fold CV. The splits were the same for all the experiments and the distribution of image quality and presence of gadolinium respected the overall distribution.

2.2.7. Training subsets

In order to study potential biases related to the presence of gadolinium or the quality of the images, we created different training subsets:

- $T^{172}_{\text{no gado}}$ includes only matched dementia, NDL and NDNL patients with images acquired without gadolinium injection. This results in a training subset of 172 patients per class.

- $T^{181}_{\text{tier } 1/2}$ includes only matched dementia, NDL and NDNL patients with

Table 2: For each category, we report the number of patients and images, the age, the percentage of females, of images in tier 1/2 (i.e. images of good and medium quality) and the percentage of images with gadolinium-based contrast agent. Results with ** mean that the distributions between the overall population and a specific category were statistically significantly different (χ^2 test corrected for multiple comparisons using the Bonferroni procedure, corrected p-value <0.05). Age and sex were computed at the patient level, while the tiers and the gadolinium injection were computed at the image level.

Category	**N patients**	**N images**	**Age (mean ± std [range])**	**Sex (%F)**	**%Tier 1/2**	**With gadolinium**
D	756	887	71.17 ± 11.58 [18,90]	50.34%	57.72%**	24.80%**
NDL	756	997	71.17 ± 11.58 [18,90]	50.34%	52.25%	63.59%**
NDNL	756	939	71.17 ± 11.58 [18,90]	50.34%	36.42%**	66.13%**
Total	2268	2823	71.17 ± 11.58 [18,90]	50.34%	48.71%	52.24%

images of good or medium quality (tier 1/2). This results in a training subset of 181 patients per class.

- T^{172} includes 172 patients per class with the same distribution of image quality and gadolinium injection than the overall data set.

- $T^{88}_{\text{no gado, tier 1/2}}$ includes only matched dementia, NDL and NDNL patients with images of medium or good quality acquired without gadolinium injection. This results in a training subset of 88 patients per class.

- $T^{88}_{\text{tier 1/2}}$ includes 88 patients per class of only images of good or medium quality.

- T^{88} includes 88 subjects per class with the same distribution of image quality and gadolinium injection as the overall data set.

3. Methods

3.1. Image pre-processing

The T1w MR images were converted from DICOM to NIfTI using the software `dicom2niix` (version tag v1.0.20190902, commit number f54be46) (Li et al., 2016) and organized following the Brain Imaging Data Structure (BIDS) standard (Gorgolewski et al., 2016). Images with a voxel dimension smaller than 0.9 mm were resampled using a 3^{rd}-order spline interpolation to obtain 1 mm isotropic voxels. Two different pre-processing pipelines were applied to the T1w MR images in the BIDS format.

Most of the pre-processing was performed using Clinica (Routier et al., 2021) (version tag 0.3.5, commit number 06fdbc5). The first pre-processing consisted in applying the `t1-linear` pipeline of Clinica, which is a wrapper of the ANTs software (Avants et al., 2014) (version tag 2.3.1). Bias field correction was applied using the N4ITK method (Tustison et al., 2010). An affine registration to MNI space was performed using the SyN algorithm (Avants et al., 2008). N4ITK and SyN algorithms are implemented in the ANTs software. The registered images were further rescaled based on the min and max intensity values. Images were then cropped to remove background resulting in images of size 169×208×179, with 1 mm isotropic voxels (Wen et al., 2020).

This pre-processing was used to assess the quality of the images with an automatic approach proposed in (Bottani et al., 2022a). The automatic quality control (QC) approach first identified if a given image was or not a

straight reject (i.e. segmented or cropped image). If it was not a straight reject, it was further labeled by the automatic QC tool according to the tiers of quality, i.e. tier 1 (good quality), tier 2 (medium quality) or tier 3 (bad quality). In addition, the automatic QC tool determined the presence or the absence of gadolinium-based contrast agent.

The second pre-processing consisted in applying the `t1-volume-tissue-segmentation` pipeline of Clinica (Routier et al., 2021; Samper-González et al., 2018) to obtain probability gray matter maps from the T1w MR images in the BIDS format. This wrapper of the Unified Segmentation procedure implemented in SPM12 (Ashburner and Friston, 2005) simultaneously performs tissue segmentation, bias correction and spatial normalization. SPM12 was installed with the Matlab standalone version. This results in probability gray matter maps in the MNI space that have a size of 121×145×121, with 1.5 mm isotropic voxels.

3.2. Synthesis of images without gadolinium

To attenuate a potential bias due to the presence or absence of gadolinium, all the images pre-processed with the `t1-linear` pipeline went through the *Att-U-Net* described in (Bottani et al., 2022b) that translates contrast-enhanced images into non-contrast-enhanced images. The code can be found at: `https://github.com/SimonaBottani/image_synthesis` (commit number 98710ed). To prevent introducing a potential bias because of differences in smoothness between the real and synthetic images, all the images were fed

to the network no matter the initial presence or absence of gadolinium. The synthetic images were then pre-processed with the `t1-volume-tissue-segmentation` pipeline, as done for the real images.

3.3. Machine learning models used for classification

3.3.1. Linear support vector machine

A linear support vector machine (SVM) using probability gray matter maps as features was used for the binary classification tasks. We followed the implementation of (Samper-González et al., 2018) using Scikit-learn (Pedregosa et al., 2011). The Gram matrix $K = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j}$ was pre-calculated using a linear kernel k for each pair of images $(\mathbf{x}_i, \mathbf{x}_j)$ for the provided subjects and was used as input for the generic SVM. When using a pre-computed Gram matrix, computing time depends on the number of subjects, and not on the number of features and it can speed up calculations. The implementation of the linear SVM models, including the computation of the Gram Matrix and the integration with the features pre-processed with Clinica can be found here: `https://github.com/aramis-lab/AD-ML` (branch 2018_NeuroImage, commit number 4d07049). We optimized the penalty parameter C of the error term. The optimal value of C was chosen using nested CV, with an inner k-fold (k=10). For each fold of the outer CV, the value of C that led to the highest balanced accuracy in the inner k-fold was selected. The test set was never used for hyper-parameter optimization and was only used to estimate the final performance results.

3.3.2. Convolutional neural networks

We used three different 3D convolutional neural networks (CNN) for the binary classification tasks to have a comparison with the linear SVM. Note that the input of the CNNs are the images pre-processed with `t1-linear` as this procedure was validated in (Wen et al., 2020).

The three 3D CNN architectures considered in the paper are denominated as follows: Conv5_FC3, ResNet, InceptionNet. The first is composed of five convolutional layers and three fully connected layers as implemented in (Wen et al., 2020; Thibeau-Sutre et al., 2022b), the ResNet contains residual blocks inspired from (Jónsson et al., 2019) and the InceptionNet is a modified version of the Inception architecture implemented by (Szegedy et al., 2016). The ResNet and the InceptionNet were implemented and used for the work of (Couvy-Duchesne et al., 2020). All the details of the architectures can be found in (Bottani et al., 2022a). The deep learning models developed for classification are available within the ClinicalDL (Thibeau-Sutre et al., 2022b) software, which repository is `https://github.com/aramis-lab/clinicadl` (version tag 0.0.2, commit number 8286513) and within the repository `https://github.com/aramis-lab/pac2019` (commit number 200681e).

The models were trained using the cross entropy loss. We used the Adam optimizer with a learning rate of 10^{-5} for the ResNet and of 10^{-4} for the InceptionNet and Conv5_FC3 architectures. We implemented early stopping and all the models were evaluated with a maximum of 50 epochs. The batch

size was set to 2. The model with the lowest loss, determined on the validation set, was saved as final model. As mentioned above, the test set was never used for hyper-parameter optimization and was only used to estimate the final performance results. Implementation was done using Pytorch through the ClinicaDL platform (Thibeau-Sutre et al., 2022b).

3.4. Computing environment

All computations were performed using the computing infrastructure of the hospital data warehouse running the operating system Linux Ubuntu 14.04 LTS.

4. Results

We first classified AD vs CN subjects using the ADNI data set in order to obtain baseline results on a research data set. Then we performed two tasks using the clinical data sets: dementia vs no dementia with lesions (D vs NDL) and dementia vs no dementia no lesions (D vs NDNL).

4.1. Performance in a research data set

Results for classification of AD vs CN on ADNI are reported in Table 3. The best balanced accuracy was reached using the linear SVM with gray matter maps as input (86.4%), followed by the ResNet (85.3%), the Conv5_FC3 (84.1%), and the InceptionNet (82.1%) using minimally pre-processed T1w MR images as input. These results are in line with the literature (Samper-González et al., 2018; Wen et al., 2020) even though higher performance has

been reported with more sophisticated approaches (e.g. (Lian et al., 2018; Li et al., 2018; Hett et al., 2018; Coupé et al., 2012; Liu et al., 2012; Tong et al., 2014; Suk et al., 2017; Basaia et al., 2019; Wee et al., 2019; Hett et al., 2021)). Note that the objective of our work was not to improve the state of the art of classification of AD on research data but to have a baseline for further comparison with results on clinical routine data. This is why we used standard classification approaches. As training linear SVMs is less computationally expensive than CNNs and since the objective of our work is not to compare different machine learning approaches, for the subsequent experiments we will mostly report results obtained with the linear SVM.

Table 3: Dementia classification performance (AD vs CN) on the research data set (ADNI). Results were obtained with different machine learning models: a linear SVM using as input gray matter maps and three CNN models (Conv5_FC3, ResNet and InceptionNet) using as input minimally pre-processed T1w MR images. We present results on the independent test set using the average performance and the empirical standard deviation (SD) of the five models corresponding to the five folds. Note that the empirical SD just provides a rough estimate of the variability of the performance across folds but is not an unbiased estimator of the SD of the performance.

AD vs CN

Metric	**SVM**	**Conv5_FC3**	**ResNet**	**InceptionNet**
Balanced accuracy	86.80 ± 0.40	84.10 ± 1.59	85.30 ± 1.03	82.10 ± 1.77
Sensitivity	82.80 ± 0.40	79.80 ± 4.45	83.00 ± 4.52	75.80 ± 8.68
Specificity	90.80 ± 0.40	88.40 ± 7.26	87.60 ± 4.67	88.40 ± 5.16

Table 4: Dementia classification performance (D vs NDNL and D vs NDL) in the clinical data set. Results were obtained with a linear SVM using as input gray matter maps.

Metric	**D vs NDNL**	**D vs NDL**
Balanced accuracy	68.75 ± 0.36	73.09 ± 0.32
Sensitivity	66.97 ± 0.64	75.92 ± 0.89
Specificity	70.53 ± 0.49	70.26 ± 0.49

4.2. Performance in the clinical data set

Classification results on the clinical data set (for both D vs NDNL and D vs NDL) using all the training samples available are reported in Table 4. We observed an important drop in balanced accuracy compared with that obtained on the research data set: 68.8% for D vs NDNL and 73.1% for D vs NDL compared with 86.4% for AD vs CN in ADNI. This may be due to the heterogeneity of the classes in the clinical data set, where many diagnoses coexist, but also to differences in image characteristics.

4.2.1. Influence of gadolinium injection and image quality on the classification performance

As shown in Table 2, the proportions of images with and without gadolinium injection and of good/medium vs low quality differ in the dementia, NDL and NDNL categories. In the dementia class, 25% images were acquired with gadolinium injection. In NDL and in NDNL, this proportion is around 65%. In the dementia and NDL categories, the majority of the images are

of good/medium quality (58% and 52%, respectively), while in the NDNL category only 36% of images are of good/medium quality. Since these acquisition characteristics are correlated with the diagnostic class, it is possible that the classifier uses this information characteristic, thereby biasing the performance upwards, a phenomenon often referred to as the Clever Hans effect (Lapuschkin et al., 2019) or shortcut learning (Geirhos et al., 2020).

To test this hypothesis, we used the training subsets $T^{172}_{\text{no gado}}$, $T^{181}_{\text{tier 1/2}}$ and T^{172}. The order of magnitude of patients per class among the training subsets is equivalent, meaning that differences observed in the classification score should not depend on the training sample size but on the characteristics of the training subset. We assume that if gadolinium or image quality has no impact, the performance will not vary when using the different training subsets. On the other hand, if results differ between training subsets, this will be the sign of a Clever Hans effect. Results of these experiments are displayed in Table 5. Note that the test set never changed across all the experiments of the work: it is composed of 152 patients/images per class. The balanced accuracy when using T^{172} was substantially higher than when using $T^{172}_{\text{no gado}}$ or $T^{181}_{\text{tier 1/2}}$. This indicates that results are biased by the presence of gadolinium and by the differences in image quality. The classifiers exploit these characteristics to determine the diagnosis.

The training subset $T^{172}_{\text{no gado}}$ still contains images of different quality and $T^{181}_{\text{tier 1/2}}$ images with and without gadolinium. The classifier may thus still be exploiting biases in the image characteristics. To evaluate the performance

Table 5: Influence of gadolinium injection and image quality on the classification performance. Results were obtained for the D vs NDNL and D vs NDL classification tasks with a linear SVM using as input gray matter maps and trained on different clinical data subsets ($T^{172}_{\text{no gado}}$, $T^{181}_{\text{tier 1/2}}$ and T^{172}).

A. D vs NDNL

Metric	$T^{172}_{\textbf{no gado}}$	$T^{181}_{\textbf{tier 1/2}}$	T^{172}
Balanced accuracy	60.33 ± 0.26	61.32 ± 2.83	68.16 ± 0.38
Sensitivity	52.76 ± 0.26	79.87 ± 2.72	73.95 ± 2.41
Specificity	67.89 ± 0.26	42.76 ± 12.70	62.37 ± 2.18

B. D vs NDL

Metric	$T^{172}_{\textbf{no gado}}$	$T^{181}_{\textbf{tier 1/2}}$	T^{172}
Balanced accuracy	69.74 ± 0.55	64.61 ± 1.74	72.30 ± 0.48
Sensitivity	85.13 ± 0.79	45.53 ± 4.62	66.45 ± 1.32
Specificity	54.34 ± 1.84	83.68 ± 1.47	78.16 ± 1.92

of the classifier using a training data set without any of these two potential biases, we used the training subset called $T^{88}_{\text{no gado, tier 1/2}}$ and compared it with using the training subset T^{88} having the same training size. Results are reported in Table 6. For both tasks, there was a dramatic drop in balanced accuracy, down from about 70% to random (about 50%). Therefore, the classifier is only using the Clever Hans effect and not relevant diagnostic information. In other words, when it cannot exploit biases in image characteristics, the trained classifier is not better than a random classifier.

We aimed to assess if these observations still hold for another machine learning approach, specifically a CNN-based model. We thus conducted the same analysis using the Conv5_FC3 network. Results are reported in Table S1 in the supplementary material. We observed a 7 percent point increase for the D vs NDNL task and a 13 percent point increase for the D vs NDL task when using the biased data set. This increase is lower than that obtained for the SVM but still very large. Compared with the SVM, the CNN yielded a higher balanced accuracy on the $T^{88}_{\text{no gado, tier 1/2}}$ data subset (for D vs NDNL 58.62 ± 1.60 and for D vs NDL 55.53 ± 3.71). Nevertheless, the performance remains extremely poor when the training cannot exploit the Clever Hans effect.

Table 6: Joint influence of gadolinium injection and image quality on the classification performance. Results were obtained for the D vs NDNL and D vs NDL classification tasks with a linear SVM using as input gray matter maps and trained on two clinical data subsets ($T^{88}_{\text{no gado, tier 1/2}}$ and T^{88}).

A. D vs NDNL

Metric	$T^{88}_{\textbf{no gado, tier 1/2}}$	T^{88}
Balanced accuracy	51.51 ± 2.54	69.47 ± 2.37
Sensitivity	6.71 ± 12.44	71.97 ± 2.26
Specificity	96.32 ± 7.37	66.97 ± 2.51

B. D vs NDL

Metric	$T^{88}_{\textbf{no gado, tier 1/2}}$	T^{88}
Balanced accuracy	50.00 ± 0.00	73.03 ± 1.79
Sensitivity	40.00 ± 48.99	66.58 ± 4.51
Specificity	60.00 ± 48.99	79.47 ± 1.13

4.2.2. Classification performance obtained after gadolinium removal using image translation

In our previous work (Bottani et al., 2022b), we proposed a deep learning-based image translation approach to remove the visual effect of gadolinium from contrast-enhanced T1w MR images. In the present paper, we assess whether this approach could reduce the classification bias due to gadolinium injection. We created a training subset composed of 88 synthetic images obtained from images of good/medium quality acquired with and without gadolinium injection that all went through the gadolinium removal *Att-U-Net*, as described in section 3.2. If the gadolinium is successfully removed, training with this subset should be equivalent to training with the $T^{88}_{\text{no gado, tier 1/2}}$ subset that includes only images without gadolinium. Results of these experiments are reported in Table 7. The balanced accuracy is equivalent in both cases, meaning that the effect of gadolinium has been removed using the synthetic images. Nevertheless, it is not better than chance indicating, again, that the classifier cannot learn image characteristics which are relevant to the diagnostic classification.

However, it is possible that the low performance is due to the small size of the training set. We therefore used the image translation method to build a larger clinical data set composed only of images of good/medium quality and where the visual appearance of gadolinium has been removed. This data set was denoted Synthetic $T^{181}_{\text{tier 1/2}}$. Using this training set, we assessed both the linear SVM (using gray matter maps) and the ResNet (with minimally

pre-processed T1w MRI as input). Results appear in Table 8. We found an increased performance using this synthetic, larger, training set. The ResNet obtained a slightly higher performance than the SVM. It is thus possible that homogenizing the data set using image translation allows removing bias and increasing classification performance. Nevertheless, we cannot directly demonstrate this in the absence of a training set of the same size containing only images without gadolinium and of higher quality. It is thus possible that visually imperceptible differences still exist between the images that were initially acquired with gadolinium and those without, and that the classifiers exploit these differences.

4.2.3. Classification performance when training on a research data set and testing on the clinical data set

Another way to ensure that gadolinium or poor image quality is not exploited by the classifier is to train using the research data set (ADNI contains only images without gadolinium and of good quality). We both trained a linear SVM and a ResNet. Results appear in Table 9. No matter the task, the linear SVM trained on research data led to a slightly higher balanced accuracy than the ResNet. Note that the accuracy was also slightly higher than when training with synthetic data (Table 8). In any case, one should keep in mind that such classification performance is too low to be acceptable in clinical practice.

Table 7: Classification performance obtained after gadolinium removal using image translation, training on a set of 88 patients. Results were obtained for the D vs NDNL and D vs NDL classification tasks with a linear SVM using as input gray matter maps and trained on three clinical data subsets ($T^{88}_{\text{tier 1/2}}$, $T^{88}_{\text{tier 1/2}}$, $T^{88}_{\text{no gado, tier 1/2}}$).

A. D vs NDNL

Metric	$T^{88}_{\textbf{tier 1/2}}$	**Synthetic** $T^{88}_{\textbf{tier 1/2}}$	$T^{88}_{\textbf{no gado, tier 1/2}}$
Balanced accuracy	60.26 ± 5.41	51.71 ± 1.15	51.51 ± 2.54
Sensitivity	58.68 ± 30.44	75.66 ± 34.75	6.71 ± 12.44
Specificity	61.84 ± 22.63	27.76 ± 34.98	96.32 ± 7.37

B. D vs NDL

Metric	$T^{88}_{\textbf{tier 1/2}}$	**Synthetic** $T^{88}_{\textbf{tier 1/2}}$	$T^{88}_{\textbf{no gado, tier 1/2}}$
Balanced accuracy	68.29 ± 3.55	54.08 ± 5.19	50.00 ± 0.00
Sensitivity	69.34 ± 7.71	52.50 ± 41.54	40.00 ± 48.99
Specificity	67.24 ± 14.43	55.66 ± 45.57	60.00 ± 48.99

Table 8: Classification performance obtained after gadolinium removal using image translation, training on a set of 181 patients. Results were obtained a linear SVM with probability gray matter maps or a ResNet with minimally pre-processed T1w MR images.

A. D vs NDNL

Metric	**SVM**	**ResNet**
Balanced accuracy	61.91 ± 1.34	63.22 ± 3.47
Sensitivity	81.32 ± 2.45	52.24 ± 10.65
Specificity	42.50 ± 4.59	74.21 ± 7.22

B. D vs NDL

Metric	**SVM**	**ResNet**
Balanced accuracy	64.61 ± 1.74	67.50 ± 0.98
Sensitivity	45.53 ± 4.62	64.47 ± 10.47
Specificity	83.68 ± 1.47	70.53 ± 10.05

Table 9: Classification performance when training on a research data set and testing on a clinical data set. Results were obtained for the D vs NDNL and D vs NDL classification tasks using a linear SVM with probability gray matter maps or a ResNet with minimally pre-processed T1w MR images.

A. D vs NDNL

Metric	**SVM**	**ResNet**
Balanced accuracy	64.08 ± 0.82	61.84 ± 4.07
Sensitivity	62.76 ± 0.53	60.92 ± 8.28
Specificity	65.39 ± 1.29	62.76 ± 6.55

B. D vs NDL

Metric	**SVM**	**ResNet**
Balanced accuracy	69.47 ± 0.32	61.78 ± 4.35
Sensitivity	62.76 ± 0.53	60.92 ± 8.28
Specificity	76.18 ± 0.49	62.63 ± 4.43

5. Discussion

In this paper, we studied the performance of machine learning approaches for computer-aided detection of dementia based on T1w MRI using a real-life clinical routine cohort coming from a hospital data warehouse. To the best of our knowledge, this is the first paper of this kind since previous works have used either research data sets or clinical data from specialized centers that have been carefully selected and are thus not representative of daily clinical routine. We demonstrated that the classifiers trained on clinical routine data are highly biased by image acquisition specificities such as image quality or injection of gadolinium. When such biases are removed, the performance is very poor. Models trained on research data performed poorly and their accuracy is unacceptably low for clinical use.

As a research topic, machine learning for diagnosis of Alzheimer's disease is now 15 year old (Klöppel et al., 2008; Vemuri et al., 2008; Gerardin et al., 2009; Fan et al., 2008). While high performance has been consistently reported, most of these works use research data sets for training and validation (Samper-González et al., 2018; Falahati et al., 2014; Manera et al., 2021; Bron et al., 2021). There are a few papers using clinical routine data sets but they cannot be considered representative of daily clinical routine as they come from a single or a handful of highly specialized centers and carefully select data using strict criteria regarding data quality (Morin et al., 2020; Platero et al., 2019; Sohn et al., 2015; Klöppel et al., 2015). It is thus unclear how such methods would perform on real-life clinical MRI and

ultimately translate to the clinic.

The aim of our work was to provide a proof of concept of the abilities of machine learning algorithms to work with heterogeneous data sets as the clinical routine data sets. It is a first attempt to better understand how these models behave when the acquisition setting is very different from that of research studies. Our experimental settings does not imply that the tasks the classifiers need to tackle are representative of realistic clinical diagnostic scenarii. Indeed, in the clinic, there is a lot of prior knowledge available to the radiologist, such as the reason for referal or the patient's history, that we did not consider.

The main results of our work are three-fold: i) the performance of such CAD methods is considerably lower on clinical routine data compared with research data sets; ii) on clinical routine data, classifiers were heavily biased by irrelevant characteristics and when such biases were removed, the performance was particularly low; iii) training on research data and testing on clinical data allowed reaching slightly higher accuracies but the overall performance remained low. More specifically, when both training and testing on research data, we obtained high classification performance (around 87% balanced accuracy) which is in line with the literature. When training/testing on clinical data, the performance dropped by more than 15 percent points and, more importantly, was heavily biased by irrelevant characteristics. When such confounders were removed, the performance was very poor. This was the case for both the SVM and a CNN model. The CNN achieved slightly

better results than the SVM in the absence of confounders but the accuracy was still very low. Training on the research data set and testing on the clinical routine data set allowed removing this source of bias but the performance remained poor (decrease of at least 19 percent points of balanced accuracy). Thus, classifiers that lead to high classification performance in a research framework do not necessarily generalize to clinical data sets. Part of this drop in accuracy could be explained by an increase in the difficulty of the classification task between the research and clinical setups. In the research setup, the AD and CN classes are quite homogeneous, while in the clinical setup, the D, NDL and NDNL classes are much more heterogeneous as each category corresponds to several diagnoses. However, this may not be the only factor leading to this performance difference and more analyses were performed to dissect these results.

In the clinical routine data set, there was a clear correlation between the diagnostic groups on the one hand and image quality and presence of gadolinium on the other hand (∼65% of images with gadolinium in NDL and NDNL and 25% in D; 37% of images of good or medium quality in NDNL, and ∼55% in D and NDL). We hypothesized that models trained on such data could exploit this bias. To assess this, we trained different models changing the characteristics of the training subsets: we used training subsets having only images without gadolinium ($T^{172}_{\text{no gado}}$) or images of good/medium quality ($T^{181}_{\text{tier 1/2}}$) or both ($T^{88}_{\text{no gado, tier 1/2}}$) and we compared their performance with a training subset of the same sample size but having the same proportions of

images with gadolinium and of low quality than the whole data set (T^{172} and T^{88}). We showed that the performance of the classifier was heavily biased by these image characteristics, a phenomenon known as the Clever Hans effect (Lapuschkin et al., 2019) or shortcut learning (Geirhos et al., 2020). Such phenomenon has been previously described in different medical image computing applications (Lapuschkin et al., 2019; Wallis and Buvat, 2022). All these sub-analyses were performed using a sub-data set not representative of the original data set but they were carried out to better understand the behavior of the models.

Of course, it does not mean that other machine learning methods could not achieve higher performance using larger, unbiased, clinical routine data sets but it was not the case in our work.

We aimed to remove the bias coming from gadolinium injection by applying an image translation *Att-U-Net* model proposed in (Bottani et al., 2022b). On the smaller set of 88 patients, its performance was close to chance and similar to that of a classifier trained on images without gadolinium and of good/medium quality. This approach still has some limitations: it may blur the images, limiting the information related to the diseases that the models could detect. When using a larger data set of synthetic images, we obtained higher accuracies. This potentially indicates that the use of image translation allows removing some of the biases while improving performance. Nevertheless, we cannot strictly assert this because there may be residual, visually imperceptible differences between images that were acquired with gadolin-

ium and those without. Overall, this stresses the importance of developing image homogenization techniques for training unbiased classifiers.

Our results contribute to a better understanding of the broader issue of generalizability of machine learning systems to clinical routine data. Such issue is mentioned as one of the major challenges faced by machine learning for medical imaging in Varoquaux and Cheplygina (2022). In particular, the authors highlight the discrepancy between results that can be achieved in research benchmarks and when the tool is applied to more realistic clinical data. Similarly to our results, it has been found that machine learning systems may in fact be exploiting confounders in different settings. Some examples of exploited confounders include site and clinical department (Zech et al., 2018), magnetic field strength (Thibeau-Sutre et al., 2022a), the presence of a chest drain that is actually implemented only after the diagnosis and not before (Oakden-Rayner et al., 2020) and skin markings associated with melanoma detection (Winkler et al., 2019). More generally, failure to generalize is often due to the fact that the data sets used for training are not representative of clinical routine. In medical imaging, this is particularly problematic because imaging devices, acquisition parameters and, more generally, the setting within which the acquisition is performed, are major sources of variability. Within a given research data set, such sources of variability are controlled, up to a certain point. However, it has been shown that machine learning can identify data sets (Wachinger et al., 2021). In clinical routine, acquisition settings are largely uncontrolled making the translation

from research data particularly difficult. Furthermore, the difficulty of translation to clinical routine goes beyond the medical imaging field (Futoma et al., 2020). It has for example been demonstrated when predicting acute kidney injury (Davis et al., 2017), heart failure (Wessler et al., 2017) or mortality from electronic health records (Singh et al., 2022).

Trustworthy medical machine learning is an increasingly important topic. However, there is still a need to raise awareness on this multifaceted issue. Methodological researchers have a tendency to often use the same research data sets. In the field of dementia, hundreds of machine learning papers have been published using ADNI (Rathore et al., 2017; Ebrahimighahnavieh et al., 2020; Ansart et al., 2021). Unfortunately, comparatively, only little research has been made using clinical routine data. Furthermore, while benchmarks and challenges are very valuable, they are often not representative enough of the clinical reality. There is thus a need to have more representative benchmarks (Varoquaux and Cheplygina, 2022). More generally, it is necessary that researchers in medical image computing are not only interested in building new models for diagnosis or prognosis but also in addressing the challenges of their performance on realistic data. Awareness is also necessary among clinicians. In particular, it is important that they are trained to understand problems such as shortcut learning (Geirhos et al., 2020), so that they can be aware of the limitations of machine learning tools. Finally, it is also our role, as scientists, to engage with the general public about such topics. Indeed, it is essential that patients are not only aware of these prob-

lems but can also contribute to shaping unbiased machine learning systems for healthcare.

Our study has the following limitations. Unlike in research studies, the diagnosis may not be trustworthy as it is assigned using ICD-10 codes, which could be a source of bias. Indeed, in the French healthcare system, they are assigned during hospitalization by the clinical department for the billing of the expenses. In addition, ICD-10 codes do not undergo quality control and it is likely that mistakes occur when entering the codes. These limitations of the diagnostic labels may hamper the performance of the classifiers. In order to have more reliable diagnostic labels, it would be necessary to use information from medical reports. This could be done by medical experts but this is time consuming and may not scale up to large populations. Another option is to use natural language processing but it may also lead to errors. Other limitations concern the composition of the clinical data set: there is a part of arbitrariness in the way we defined the three groups. For example, in the dementia class we included just the two most common causes of dementia but other causes exist and they may also lead to brain atrophy. In addition, we limited the size of the dementia category by deciding to only consider patients with no multiple codes and whose code did not change over time. This decision was mainly motivated by the unreliability of the ICD-10 codes which may vary depending on the hospital service or the person in charge of attributing them. We assumed that a stable and unique code would be more reliable. In addition, we decided to select patients belonging to NDL and

NDNL matching by age/sex the class of dementia: this has further reduced the size of our data set, even if the main limitation was the size of the dementia class, but it was the easier way to obtain a data set not biased by these two cofactors.

Due to all the choices we have made, we have reduced the sample size of the data set. Further evaluations should be done to assess whether the performance of the classifiers could improve over the present work by adding more subjects in the training. Finally, we have limited our experimental settings to the use of standard machine learning approaches (linear SVM or standard CNN classifiers). More sophisticated approaches (e.g. (Lian et al., 2018; Li et al., 2018; Hett et al., 2018; Coupé et al., 2012; Liu et al., 2012; Tong et al., 2014; Suk et al., 2017; Basaia et al., 2019; Wee et al., 2019; Hett et al., 2021)) have been proposed in the literature which, on research data, resulted in higher classification accuracies than those obtained with the standard techniques used in the present work. In the context of our work, one of the most important limiting factors was the amount of time needed for training a CNN model in the closed environment of the clinical data warehouse (about 24 hours for one CNN model for 50 epochs). This was one of the motivations for relying on standard models. Another motivation was that we believed it was important to first study the behavior of standard, widely-used, techniques. Future work could assess whether more sophisticated techniques, which have achieved better results on research data, could also lead to improvements in the context of heterogeneous clinical data sets. To im-

prove the reproducibility, we extensively reported implementation details, software/OS versions, repositories and commit numbers as recommended by (Kennedy et al., 2019). However, the reproducibility remains limited by the fact that, as previously mentioned, the data must remain within the hospital network. Still, we believe that, given the information we have provided, other researchers who would be granted access to the hospital data warehouse for that purpose by the AP-HP Scientific and Ethics Board would be able to reproduce the present work.

Overall, our results highlight the challenges for translation of CAD systems from research to clinical routine. A major result of this study is uncovering the strong influence of biases coming from image heterogeneity. We specifically studied the case of gadolinium injection and image quality but other sources of biases such as image resolution, sequence parameters or scanner type could exist. They could in turn induce Clever Hans effects on the CAD systems if they are correlated with the diagnosis of interest. This highlights the need for automatic quality control tools in order to identify the various sources of biases as well as for homogenization tools that could remove these biases.

Acknowledgments

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 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-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6).

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.; 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.

Disclosure statement

Competing financial interests related to the present article: none to disclose for all authors.

Competing financial interests unrelated to the present article: OC reports having received consulting fees from AskBio and Therapanacea and having received fees for writing a lay audience short paper from Expression Santé. Members from his laboratory have co-supervised a PhD thesis with myBrainTechnologies and with Qynapse. OC's spouse is an employee and holds stock-options of myBrainTechnologies. O.C. holds a patent registered at the International Bureau of the World Intellectual Property Organization (PCT/IB2016/0526993, Schiratti J-B, Allassonniere S, Colliot O, Durrleman S, A method for determining the temporal progression of a biological phenomenon and associated methods and devices).

APPRIMAGE Study Group

Olivier Colliot, Ninon Burgos, Simona Bottani, Sophie Loizillon [1]
Didier Dormont [1,2], Samia Si Smail Belkacem, Sebastian Ströer [2]
Nathalie Boddaert [3]
Farida Benoudiba, Ghaida Nasser, Claire Ancelet, Laurent Spelle [4]
Hubert Ducou-Le-Pointe[5]
Catherine Adamsbaum[6]
Marianne Alison[7]
Emmanuel Houdart[8]
Robert Carlier [9,17]
Myriam Edjlali[9]
Betty Marro[10,11]
Lionel Arrive[10]
Alain Luciani[12]
Antoine Khalil[13]
Elisabeth Dion[14]
Laurence Rocher[15]
Pierre-Yves Brillet[16]
Paul Legmann, Jean-Luc Drape [18]
Aurélien Maire, Stéphane Bréant, Christel Daniel, Martin Hilka, Yannick

Jacob, Julien Dubiel, Cyrina Saussol, Rafael Gozlan [19]
Florence Tubach, Jacques Ropers, Antoine Rozès, Camille Nevoret [20]

[1] Sorbonne Université, Institut du Cerveau - Paris Brain Institute, Inserm, CNRS, AP-HP, Hôpital de la Pitié Salpêtrière, Inria, Aramis project-team, F-75013, Paris, France
[2] AP-HP, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, F-75013, Paris, France
[3] AP-HP, Hôpital Necker, Department of Radiology, F-75015, Paris, France
[4] AP-HP, Hôpital Bicêtre, Department of Radiology, F-94270, Le Kremlin-Bicêtre, France
[5] AP-HP, Hôpital Armand-Trousseau, Department of Radiology, F-75012, Paris, France
[6] AP-HP, Hôpital Bicêtre, Department of Pediatric Radiology, F-94270, Le Kremlin-Bicêtre, France
[7] AP-HP, Hôpital Robert-Debré, Department of Radiology, F-75019, Paris, France
[8] AP-HP, Hôpital Lariboisière , Department of Neuroradiology, F-75010, Paris, France
[9] AP-HP, Hôpital Raymond-Poincaré, Department of Radiology, F-92380, Garches, France
[10] AP-HP, Hôpital Saint-Antoine, Department of Radiology, F-75012, Paris, France
[11] AP-HP, Hôpital Tenon, Department of Radiology, F-75020, Paris, France
[12] AP-HP, Hôpital Henri-Mondor, Department of Radiology, F-94000, Créteil, France
[13] AP-HP, Hôpital Bichat, Department of Radiology, F-75018, Paris, France
[14] AP-HP, Hôpital Hôtel-Dieu, Department of Radiology, F-75004, Paris, France
[15] AP-HP, Hôpital Antoine-Béclère, Department of Radiology, F-92140, Clamart, France
[16] AP-HP, Hôpital Avicenne, Department of Radiology, F-93000, Bobigny, France
[17] AP-HP, Hôpital Ambroise Paré, Department of Radiology, F-92100 104, Boulogne-Billancourt, France
[18] AP-HP, Hôpital Cochin, Department of Radiology, F-75014, Paris, France
[19] AP-HP, WIND department, F-75012, Paris, France
[20] AP-HP, Unité de Recherche Clinique, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, F-75013, Paris, France

References

Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T., Couronné, R., Faouzi, J., Koval, I., Louis, M., et al., 2021. Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review. Medical Image Analysis 67, 101848.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. NeuroImage 26, 839–851.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12, 26–41.

Avants, B.B., Tustison, N.J., Stauffer, M., Song, G., Wu, B., Gee, J.C., 2014. The Insight ToolKit image registration framework. Frontiers in Neuroinformatics 8, 44.

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., Filippi, M., Initiative, A.D.N., et al., 2019. Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks. NeuroImage: Clinical 21, 101645.

Bidani, A., Gouider, M.S., Travieso-González, C.M., 2019. Dementia detection and classification from MRI images using deep neural networks and transfer learning, in: International Work-Conference on Artificial Neural Networks, Springer. pp. 925–933.

Böhle, M., Eitel, F., Weygandt, M., Ritter, K., 2019. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Frontiers in aging neuroscience 11, 194.

Bottani, S., 2022. Machine learning for neuroimaging using a very large scale clinical datawarehouse. Ph.D. thesis. Sorbonne Université-EDITE.

Bottani, S., Burgos, N., Maire, A., Wild, A., Ströer, S., Dormont, D., Colliot, O., 2022a. Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse. Medical Image Analysis 75, 102219.

Bottani, S., Thibeau-Sutre, E., Maire, A., Ströer, S., Dormont, D., Colliot, O., Burgos, N., 2022b. Homogenization of brain MRI from a clinical data warehouse using contrast-enhanced to non-contrast-enhanced image translation with U-Net derived models, in: SPIE Medical Imaging 2022.

Bron, E.E., Klein, S., Papma, J.M., Jiskoot, L.C., Venkatraghavan, V., Linders, J., Aalten, P., De Deyn, P.P., Biessels, G.J., Claassen, J.A., et al., 2021. Cross-cohort generalizability of deep and conventional machine learning for mri-based diagnosis and prediction of alzheimer's disease. NeuroImage: Clinical 31, 102712.

Burgos, N., Bottani, S., Faouzi, J., Thibeau-Sutre, E., Colliot, O., 2021. Deep learning for brain disorders: from data processing to disease treatment. Briefings in Bioinformatics 22, 1560–1576.

Chagué, P., Marro, B., Fadili, S., Houot, M., Morin, A., Samper-González, J., Beunon, P., Arrivé, L., Dormont, D., Dubois, B., et al., 2021. Radiological classification of dementia from anatomical MRI assisted by machine learning-derived maps. Journal of Neuroradiology 48, 412–418.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O., 2009. Fully automatic hippocampus segmentation and classification in alzheimer's disease and mild cognitive impairment applied on data from adni. Hippocampus 19, 579–587.

Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Pruessner, J.C., Allard, M., Collins, D.L., Initiative, A.D.N., et al., 2012. Scoring by nonlocal image patch estimator for early detection of alzheimer's disease. NeuroImage: clinical 1, 141–152.

Couvy-Duchesne, B., Faouzi, J., Martin, B., Thibeau-Sutre, E., Wild, A., Ansart, M., Durrleman, S., Dormont, D., Burgos, N., Colliot, O., 2020. 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. Frontiers in Psychiatry 11.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification

of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage 56, 766–781.

Daniel, C., Salamanca, E., 2020. Hospital Databases, in: Healthcare and Artificial Intelligence. Springer, pp. 57–67.

Davis, S.E., Lasko, T.A., Chen, G., Siew, E.D., Matheny, M.E., 2017. Calibration drift in regression and machine learning models for acute kidney injury. Journal of the American Medical Informatics Association 24, 1052–1061.

Ebrahimighahnavieh, M.A., Luo, S., Chiong, R., 2020. Deep learning to detect alzheimer's disease from neuroimaging: A systematic literature review. Computer methods and programs in biomedicine 187, 105242.

Falahati, F., Westman, E., Simmons, A., 2014. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. Journal of Alzheimer's disease 41, 685–708.

Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., Initiative, A.D.N., et al., 2008. Spatial patterns of brain atrophy in mci patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. Neuroimage 39, 1731–1743.

Farooq, A., Anwar, S., Awais, M., Rehman, S., 2017. A deep CNN based multi-class classification of Alzheimer's disease using MRI, in: 2017 IEEE International Conference on Imaging systems and techniques (IST), IEEE. pp. 1–6.

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., Celi, L.A., 2020. The myth of generalisability in clinical research and machine learning in health care. The Lancet Digital Health 2, e489–e492.

Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A., 2020. Shortcut learning in deep neural networks. Nature Machine Intelligence 2, 665–673.

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., et al., 2009.

Multidimensional classification of hippocampal shape features discriminates alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage 47, 1476–1486.

Gorgolewski, K.J., Auer, T., Calhoun, V.D., Craddock, R.C., Das, S., Duff, E.P., Flandin, G., Ghosh, S.S., Glatard, T., Halchenko, Y.O., Handwerker, D.A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B.N., Nichols, T.E., Pellman, J., Poline, J.B., Rokem, A., Schaefer, G., Sochat, V., Triplett, W., Turner, J.A., Varoquaux, G., Poldrack, R.A., 2016. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific data 3, 1–9.

Hett, K., Ta, V.T., Manjón, J.V., Coupé, P., Initiative, A.D.N., et al., 2018. Adaptive fusion of texture-based grading for alzheimer's disease classification. Computerized Medical Imaging and Graphics 70, 8–16.

Hett, K., Ta, V.T., Oguz, I., Manjón, J.V., Coupé, P., Initiative, A.D.N., et al., 2021. Multi-scale graph-based grading for alzheimer's disease prediction. Medical image analysis 67, 101850.

Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M.K., Johnson, S.C., Initiative, A.D.N., et al., 2009. Spatially augmented lpboosting for ad classification with evaluations on the adni dataset. Neuroimage 48, 138–149.

Jónsson, B.A., Bjornsdottir, G., Thorgeirsson, T., Ellingsen, L.M., Walters, G.B., Gudbjartsson, D., Stefansson, H., Stefansson, K., Ulfarsson, M., 2019. Brain age prediction using deep learning uncovers associated sequence variants. Nature Communications 10, 1–10.

Kennedy, D.N., Abraham, S.A., Bates, J.F., Crowley, A., Ghosh, S., Gillespie, T., Goncalves, M., Grethe, J.S., Halchenko, Y.O., Hanke, M., et al., 2019. Everything matters: the repronim perspective on reproducible neuroimaging. Frontiers in neuroinformatics , 1.

Klöppel, S., Peter, J., Ludl, A., Pilatus, A., Maier, S., Mader, I., Heimbach, B., Frings, L., Egger, K., Dukart, J., et al., 2015. Applying automated mr-based diagnostic methods to the memory clinic: a prospective study. Journal of Alzheimer's disease 47, 939–954.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack Jr, C.R., Ashburner, J., Frackowiak, R.S., 2008. Automatic classification of mr scans in alzheimer's disease. Brain 131, 681–689.

Koikkalainen, J., Rhodius-Meester, H., Tolonen, A., Barkhof, F., Tijms, B., Lemstra, A.W., Tong, T., Guerrero, R., Schuh, A., Ledig, C., Rueckert, D., Soininen, H., Remes, A.M., Waldemar, G., Hasselbalch, S., Mecocci, P., van der Flier, W., Lötjönen, J., 2016. Differential diagnosis of neurodegenerative diseases using structural MRI data. NeuroImage: Clinical 11, 435–449.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R., 2019. Unmasking clever hans predictors and assessing what machines really learn. Nature communications 10, 1–8.

Li, F., Liu, M., Initiative, A.D.N., et al., 2018. Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks. Computerized Medical Imaging and Graphics 70, 101–110.

Li, X., Morgan, P.S., Ashburner, J., Smith, J., Rorden, C., 2016. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. Journal of Neuroscience Methods 264, 47–56.

Lian, C., Liu, M., Zhang, J., Shen, D., 2018. Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri. IEEE transactions on pattern analysis and machine intelligence 42, 880–893.

Liu, M., Zhang, D., Shen, D., Initiative, A.D.N., et al., 2012. Ensemble sparse classification of alzheimer's disease. NeuroImage 60, 1106–1116.

Ma, D., Lu, D., Popuri, K., Wang, L., Beg, M.F., Initiative, A.D.N., et al., 2020. Differential Diagnosis of Frontotemporal Dementia, Alzheimer's Disease, and Normal Aging Using a Multi-Scale Multi-Type Feature Generative Adversarial Deep Neural Network on Structural Magnetic Resonance Images. Frontiers in Neuroscience 14, 853.

Manera, A.L., Dadar, M., Van Swieten, J.C., Borroni, B., Sanchez-Valle, R., Moreno, F., Laforce Jr, R., Graff, C., Synofzik, M., Galimberti, D.,

et al., 2021. MRI data-driven algorithm for the diagnosis of behavioural variant frontotemporal dementia. Journal of Neurology, Neurosurgery & Psychiatry 92, 608–616.

Misra, C., Fan, Y., Davatzikos, C., 2009. Baseline and longitudinal patterns of brain atrophy in mci patients, and their use in prediction of short-term conversion to ad: results from adni. Neuroimage 44, 1415–1422.

Morin, A., Samper-Gonzalez, J., Bertrand, A., Ströer, S., Dormont, D., Mendes, A., Coupé, P., Ahdidan, J., Lévy, M., Samri, D., Hampel, H., Dubois, B., Teichmann, M., Epelbaum, S., Colliot, O., 2020. Accuracy of MRI Classification Algorithms in a Tertiary Memory Center Clinical Routine Cohort. Journal of Alzheimer's Disease 74, 1157–1166.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C., 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging, in: Proceedings of the ACM conference on health, inference, and learning, pp. 151–159.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. the Journal of machine Learning research 12, 2825–2830.

Platero, C., López, M.E., Carmen Tobar, M.d., Yus, M., Maestu, F., 2019. Discriminating alzheimer's disease progression using a new hippocampal marker from t1-weighted mri: The local surface roughness. Human brain mapping 40, 1666–1676.

Punjabi, A., Martersteck, A., Wang, Y., Parrish, T.B., Katsaggelos, A.K., 2019. Neuroimaging modality fusion in Alzheimer's classification using convolutional neural networks. PloS one 14, e0225759.

Rathore, S., Habes, M., Iftikhar, M.A., Shacklett, A., Davatzikos, C., 2017. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. NeuroImage 155, 530–548.

Routier, A., Burgos, N., Díaz, M., Bacci, M., Bottani, S., El-Rifai, O., Fontanella, S., Gori, P., Guillon, J., Guyot, A., Hassanaly, R., Jacque-

mont, T., Lu, P., Marcoux, A., Moreau, T., Samper-González, J., Teichmann, M., Thibeau–Sutre, E., Vaillant, G., Wen, J., Wild, A., Habert, M.O., Durrleman, S., Colliot, O., 2021. Clinica: An Open Source Software Platform for Reproducible Clinical Neuroscience Studies. hal-02308126 URL: https://hal.inria.fr/hal-02308126.

Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., Bertrand, A., Bertin, H., Habert, M.O., Durrleman, S., Evgeniou, T., Colliot, O., 2018. Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. NeuroImage 183, 504–521.

Singh, H., Mhasawade, V., Chunara, R., 2022. Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. PLOS Digital Health 1, e0000023.

Sohn, B.K., Yi, D., Seo, E.H., Choe, Y.M., Kim, J.W., Kim, S.G., Choi, H.J., Byun, M.S., Jhoo, J.H., Woo, J.I., et al., 2015. Comparison of regional gray matter atrophy, white matter alteration, and glucose metabolism as a predictor of the conversion to alzheimer's disease in mild cognitive impairment. Journal of Korean medical science 30, 779–787.

Spasov, S., Passamonti, L., Duggento, A., Lio, P., Toschi, N., Initiative, A.D.N., et al., 2019. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease. Neuroimage 189, 276–287.

Suk, H.I., Lee, S.W., Shen, D., Initiative, A.D.N., et al., 2017. Deep ensemble learning of sparse regression models for brain disease diagnosis. Medical image analysis 37, 101–113.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.

Thibeau-Sutre, E., Couvy-Duchesne, B., Dormont, D., Colliot, O., Burgos, N., 2022a. MRI field strength predicts Alzheimer's disease: a case example of bias in the ADNI data set, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–4.

Thibeau-Sutre, E., Diaz, M., Hassanaly, R., Routier, A., Dormont, D., Colliot, O., Burgos, N., 2022b. ClinicaDL: An open-source deep learning software for reproducible neuroimaging processing. Computer Methods and Programs in Biomedicine 220, 106818.

Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J.V., Rueckert, D., Initiative, A.D.N., et al., 2014. Multiple instance learning for classification of dementia in brain mri. Medical image analysis 18, 808–818.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Transactions on Medical Imaging 29, 1310–1320.

Varoquaux, G., Cheplygina, V., 2022. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ digital medicine 5, 48.

Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack Jr, C.R., 2008. Alzheimer's disease diagnosis in individual subjects using structural mr images: validation studies. Neuroimage 39, 1186–1197.

Wachinger, C., Rieckmann, A., Pölsterl, S., Initiative, A.D.N., et al., 2021. Detect and correct bias in multi-site neuroimaging datasets. Medical Image Analysis 67, 101879.

Wallis, D., Buvat, I., 2022. Clever hans effect found in a widely used brain tumour mri dataset. Medical Image Analysis , 102368.

Wee, C.Y., Liu, C., Lee, A., Poh, J.S., Ji, H., Qiu, A., Initiative, A.D.N., et al., 2019. Cortical graph neural network for ad and mci diagnosis and transfer learning across populations. NeuroImage: Clinical 23, 101929.

Wegmayr, V., Aitharaju, S., Buhmann, J., 2018. Classification of brain MRI with big data and deep 3D convolutional neural networks, in: Medical Imaging 2018: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 105751S.

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., Colliot, O., 2020.

Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation. Medical Image Analysis , 101694.

Wessler, B.S., Ruthazer, R., Udelson, J.E., Gheorghiade, M., Zannad, F., Maggioni, A., Konstam, M.A., Kent, D.M., 2017. Regional validation and recalibration of clinical predictive models for patients with acute heart failure. Journal of the American Heart Association 6, e006121.

Winkler, J.K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., et al., 2019. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA dermatology 155, 1135–1141.

Wood, D.A., Kafiabadi, S., Al Busaidi, A., Guilhem, E., Montvila, A., Lynch, J., Townend, M., Agarwal, S., Mazumder, A., Barker, G.J., et al., 2022. Accurate brain-age models for routine clinical mri examinations. NeuroImage , 118871.

World Health Organization, et al., 2007. International classification of diseases and related health problems, 10^{th} revision. http://www. who. int/classifications/apps/icd/icd10online .

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine 15, e1002683.

Evaluation of MRI-based machine learning approaches for computer-aided diagnosis of dementia in a clinical data warehouse

-

Supplementary Material

1. Ethics approval, procedure and regulations allowing the access and the use of patient data

The AP-HP obtained the authorization of the CNIL (Commission Nationale de l'informatique et des Libertés, the French regulatory body for data collection and management) in 2017 to share data for research purposes in compliance with the MR004 reference methodology (Daniel and Salamanca, 2020). The MR004 reference controls data processing for the purpose of studying, evaluating and/or researching that does not involve human persons (in the sense of not involving an intervention or a prospective collection of research data in patients that would not be necessary for clinical evaluation, but which allows retrospective use of data previously acquired in patients). The goals of the clinical data warehouse are the development of decision support algorithms, the support of clinical trials and the promotion of multi-centre studies. According to French regulation, and as authorised by the CNIL, patients' consent to use their data in the projects of the CDW can be waived as these data were acquired as part of the clinical routine care of the patients. At the same time, AP-HP committed to keep patients updated about the different research projects of the clinical data warehouse through a portal on the internet[6] and individual information is systematically provided to all the patients admitted to the AP-HP. In addition, a retrospective information campaign was conducted by the AP-HP in 2017: it involved around 500,000 patients who were contacted by e-mail and by postal mail to be informed of the development of the CDW. Accessing the data is possible with the following procedure. A detailed project must be submitted to the Scientific and Ethics Board of the AP-HP. If the project

[6] `https://eds.aphp.fr/recherches-en-cours`

participants are external to AP-HP, they have to sign a contract with the Clinical Research and Innovation Board (Direction de la Recherche Clinique et de l'Innovation). The project must include the goals of the research, the different steps that will be pursued, a detailed description of the data needed, of the software tools necessary for the processing, and a clear statement of the public health benefits. Once the project is approved, the research team is granted access to the Big Data Platform (BDP), which was created by a sub-department of the IT of the AP-HP. The BDP is a platform internal to the AP-HP where data are collected and that external users can access to perform all their analyses, in accordance with the CNIL regulation. It is strictly forbidden to export any kind of data and each user can access only a workspace that is specific to their project. Each person of the research team can access the BDP with an AP-HP account after two-factor authentication. If the research team includes people that are not employed by the AP-HP, a temporary account associated to the project is activated. The project on which the proposed work is based is called APPRIMAGE, it is led by the ARAMIS team (current AP-HP PI: Didier Dormont; initial AP-HP PI: Anne Bertrand, deceased March 2nd 2018) at the Paris Brain Institute and it was approved by the Scientific and Ethics Board of the AP-HP in 2018 (Bottani, 2022).

2. Supplementary tables

Table S1: Joint influence of gadolinium injection and image quality on the classification performance. Results were obtained for the D vs NDNL and D vs NDL classification tasks using the Conv5_FC3 network with the minimally pre-processed T1w MR images as inputs and trained on two clinical data subsets ($T^{88}_{\text{no gado, tier 1/2}}$ and T^{88}).

A. D vs NDNL

Metric	$T^{88}_{\textbf{no gado, tier 1/2}}$	T^{88}
Balanced accuracy	58.62 ± 1.60	65.13 ± 5.91
Sensitivity	52.50 ± 9.67	57.24 ± 21.42
Specificity	64.74 ± 8.40	73.03 ± 10.89

B. D vs NDL

Metric	$T^{88}_{\textbf{no gado, tier 1/2}}$	T^{88}
Balanced accuracy	55.53 ± 3.71	68.42 ± 4.90
Sensitivity	59.47 ± 22.74	65.26 ± 20.72
Specificity	51.58 ± 29.79	71.58 ± 13.68