Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease

Esther E. Bron [a], Stefan Klein [a], Janne M. Papma [b], Lize C. Jiskoot [b], Vikram Venkatraghavan [a], Jara Linders [a], Pauline Aalten [c], Peter Paul De Deyn [d], Geert Jan Biessels [e], Jurgen A.H. R. Claassen [f], Huub A.M. Middelkoop [g,h], Marion Smits [a], Wiro J. Niessen [a,i], John C. van Swieten [b], Wiesje M. van der Flier [j], Inez H.G.B. Ramakers [c], Aad van der Lugt [a], for the Alzheimer's Disease Neuroimaging Initiative[1], on behalf of the Parelsnoer Neurodegenerative Diseases study group[2]

[a] *Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands*
[b] *Department of Neurology, Erasmus MC, Rotterdam, The Netherlands*
[c] *Alzheimer Center Limburg, School for Mental Health and Neuroscience (MHeNS), Maastricht University Medical Center, Maastricht, The Netherlands*
[d] *Department of Neurology and Alzheimer Center, University Medical Center Groningen, Groningen, The Netherlands*
[e] *Department of Neurology, UMC Utrecht Brain Center, University Medical Center Utrecht, Utrecht, The Netherlands*
[f] *Radboud University Medical Center, Nijmegen, The Netherlands*
[g] *Department of Neurology & Neuropsychology, Leiden University Medical Center, Leiden, The Netherlands*
[h] *Institute of Psychology, Health, Medical and Neuropsychology Unit, Leiden University, The Netherlands*
[i] *Imaging Physics, Applied Sciences, Delft University of Technology, The Netherlands*
[j] *Amsterdam University Medical Center, location VUmc, Amsterdam, The Netherlands*

ARTICLE INFO

ABSTRACT

This work validates the generalizability of MRI-based classification of Alzheimer's disease (AD) patients and controls (CN) to an external data set and to the task of prediction of conversion to AD in individuals with mild cognitive impairment (MCI).

We used a conventional support vector machine (SVM) and a deep convolutional neural network (CNN) approach based on structural MRI scans that underwent either minimal pre-processing or more extensive pre-processing into modulated gray matter (GM) maps. Classifiers were optimized and evaluated using cross-validation in the Alzheimer's Disease Neuroimaging Initiative (ADNI; 334 AD, 520 CN). Trained classifiers were subsequently applied to predict conversion to AD in ADNI MCI patients (231 converters, 628 non-converters) and in the independent Health-RI Parelsnoer Neurodegenerative Diseases Biobank data set. From this multi-center study representing a tertiary memory clinic population, we included 199 AD patients, 139 participants with subjective cognitive decline, 48 MCI patients converting to dementia, and 91 MCI patients who did not convert to dementia.

AD-CN classification based on modulated GM maps resulted in a similar area-under-the-curve (AUC) for SVM (0.940; 95%CI: 0.924–0.955) and CNN (0.933; 95%CI: 0.918–0.948). Application to conversion prediction in MCI yielded significantly higher performance for SVM (AUC = 0.756; 95%CI: 0.720 −0.788) than for CNN (AUC = 0.742; 95%CI: 0.709 −0.776) ($p < 0.01$ for McNemar's test). In external validation, performance was slightly decreased. For AD-CN, it again gave similar AUCs for SVM (0.896; 95%CI: 0.855–0.932) and CNN (0.876; 95%CI: 0.836–0.913). For prediction in MCI, performances decreased for both SVM (AUC = 0.665; 95%*CI*: 0.576 −0.760) and CNN (AUC = 0.702; 95%CI: 0.624 −0.786). Both with SVM and CNN, classification based on modulated GM maps significantly outperformed classification based on minimally processed images ($p = 0.01$).

Deep and conventional classifiers performed equally well for AD classification and their performance decreased only slightly when applied to the external cohort. We expect that this work on external validation contributes towards translation of machine learning to clinical practice.

Introduction

The diagnostic process of dementia is challenging and takes a substantial period of time after the first clinical symptoms arise: on average 2.8 years in late-onset and 4.4 years in young-onset dementia (Van Vliet et al., 2013). The window of opportunity for advancing the diagnostic process is however much larger than these few years. For Alzheimer's disease (AD), the most common form of dementia, there is increasing evidence that disease processes start 20 years or more ahead of clinical symptoms (Gordon et al., 2018). Advancing the diagnosis is essential to support the development of new disease modifying treatments, since late treatment is expected to be a major factor in the failure of clinical trials (Mehta et al., 2017). In addition, early and accurate diagnosis have great potential to reduce healthcare costs as they give patients access to supportive therapies that help to delay institutionalization (Prince et al., 2011).

Machine learning offers an approach for automatic classification by learning complex and subtle patterns from high-dimensional data. In AD research, such algorithms have been frequently developed to perform automatic diagnosis and predict the future clinical status at an individual level based on biomarkers. These algorithms aim to facilitate medical decision support by providing a potentially more objective diagnosis than that obtained by conventional clinical criteria (Klöppel et al., 2012; Rathore et al., 2017). A large body of research has been published on classification of AD and its prodromal stage, mild cognitive impairment (MCI) (Ansart et al., 2021; Wen et al., 2020; Rathore et al., 2017; Arbabshirani et al., 2017; Falahati et al., 2014; Bron et al., 2015). Overall, classification methods show high performance for classification of AD patients and control participants with an area under the receiver-operating characteristic curve (AUC) of 85–98%. Reported performances are somewhat lower for prediction of conversion to AD in patients with MCI (AUC: 62–82%). Structural T1-weighted (T1w) MRI to quantify neuronal loss is the most commonly used biomarker, whereas the support vector machine (SVM) is the most commonly used classifier. Following the trends and successes in medical image analysis and machine learning, neural network classifiers - convolutional neural networks (CNN) in particular - have increasingly been used since few years (Wen et al., 2020; Cui and Liu, 2019; Basaia et al., 2019), but have not been shown to significantly outperform conventional classifiers. Most CNN studies perform no to minimal pre-processing of the structural MRI scans as input for their classifier (Wen et al., 2020; Basaia et al., 2019; Hosseini-Asl et al., 2018; Vieira et al., 2017), while others use more extensive pre-processing strategies proven successful for conventional classifiers, such as gray matter (GM) density maps (Cui and Liu, 2019; Suk et al., 2017). Although CNNs are designed to extract high-level features from raw imaging data, it is imaginable that the learning process for complex tasks is improved by dedicated pre-processing that enhances disease-related features, which reduces model complexity and enables a more stable learning process. It is unclear yet whether CNNs would improve AD classification over conventional classifiers and whether they benefit from extensive MRI pre-processing.

Despite high performance of machine learning diagnosis and prediction methods for AD, it is largely unknown how these algorithms would perform in clinical practice. A next step would be to assess the generalizability of classification methods from a specific research population to another study population. There are however only very few studies assessing classification performance on an external data set (Wen et al., 2020; Bouts et al., 2019; Archetti et al., 2019; Hall et al., 2015). Results varied from only a minor reduction in performance for some experiments (Wen et al., 2020; Hall et al., 2015) to a severe drop for others (Bouts et al., 2019; Archetti et al., 2019; Wen et al., 2020). While generalizability seemed related to how well the training data represented the testing data (e.g. an external data set with similar inclusion criteria showed a smaller performance drop than a data set with very different criteria (Wen et al., 2020)), a better understanding is crucial before applying such methods in routine clinical practice.

Therefore, this work aims to assess the generalizability of MRI-based classification performance to an external data set representing a tertiary memory clinic population for both diagnosis of AD and prediction of AD in individuals with MCI. To evaluate the value of neural networks and to determine their optimal MRI pre-processing approach, we compare a CNN with a conventional SVM classifier using two pre-processing approaches: minimal pre-processing using only rough spatial alignment and more extensive pre-processing into modulated GM maps. First, we optimize the methods using a large research cohort and assess classification performance using cross-validation. Subsequently, we validate AD prediction performance in MCI patients of the same cohort as well as AD diagnosis and prediction performance in the external data set.

Methods

Study population

We used data from two cohorts. The first group of 1715 participants was included from the Alzheimer's Disease Neuroimaging Initiative (ADNI; 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 clinical and neuropsychological assessment, serial magnetic resonance imaging (MRI), positron emission tomography (PET), and other biological markers can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org. We included all participants with a T1w MRI scan at baseline from the ADNI1/GO/2 cohorts: 336 AD patients, 520 control participants (CN), 231 mild cognitive impaired (MCI) patients who converted to AD within 3 years (MCIc) and 628 MCI patients who did not convert (MCInc). The CN group consisted of 414 cognitively normal participants and 106 participants with subjective cognitive decline (SCD). Demographics are shown in Table 1. A list of included participants is made available at https://gitlab.com/radiology/neuro/bron-cross-cohort[3].

The second group of participants was included from the Health-RI Parelsnoer Neurodegenerative Diseases Biobank (PND; www.health-ri.nl/parelsnoer), a collaborative biobanking initiative of the eight university medical centers in the Netherlands (Manniën et al., 2017). The Parelsnoer Neurodegenerative Diseases Biobank focuses on the role of biomarkers on diagnosis and the course of neurodegenerative diseases,

Table 1
Demographics for the ADNI data set.

	AD	CN	MCIc	MCInc
# participants	336	520	231	628
male/ female	186/ 150	252/ 268	141/ 90	367/ 261
age (y; mean±std)	74.9 ± 7.8	74.2 ± 5.8	73.7 ± 7.0	72.9 ± 7.8

---

[3] https://doi.org/10.5281/zenodo.4896966

in particular of Alzheimer's disease (Aalten et al., 2014). It is a prospective, multi-center cohort study, focusing on tertiary memory clinic patients with cognitive problems including dementia. Patients are enrolled from March 2009 and followed annually for two to five years. In the PND biobank, a total of 1026 participants have been included. Inclusion criteria for the current research were: a high resolution T1w MRI at baseline, clinical consult at baseline, 90 days or less between MRI and clinical consult, and a baseline diagnosis of SCD, MCI, or dementia due to AD. A flow diagram of the inclusion can be found in the supplementary files (Fig. S1). A total of 557 participants met inclusion criteria. One person was excluded because image analysis failed. This led to inclusion of 199 AD patients and 138 participants with SCD. Of the MCI group, we included the 139 participants that had a follow-up period of at least 6 months. Of this group, 48 MCI patients converted towards dementia within the available follow-up time and 91 MCI patients remained stable. Demographics are shown in Table 2.

Imaging data

We used baseline T1w structural MRI acquired at 1.5T or 3T. Acquisition protocols are previously described (ADNI: Jack et al. (2008); Jack et al. (2015), PND: Aalten et al. (2014)). Variation in acquisition protocols used in PND is detailed in Table 3. For the majority of scans, a 8-channel head coil was used (N = 423; 76%); other scans used a 16-channel (N = 27), 24-channel (N = 1), 40-channel (N = 1), or unknown coil (N = 104).

Image pre-processing

We evaluated two pre-processing approaches based on T1w images: minimal pre-preprocessing and a more extensive pre-processing into modulated GM maps.

To prepare T1w images with minimal pre-processing, scans were non-uniformity corrected using the N4 algorithm (Tustison et al., 2010) and subsequently transformed to MNI-space using registration of brain masks with a similarity transformation. A similarity transformation is a rigid transformation including isotropic scaling. Registrations were performed with Elastix registration software (Klein et al., 2010; Shamonin et al., 2014). To account for variations in signal intensity, images were normalized within the brain mask to have zero mean and unit variance.

To obtain modulated GM maps encoding gray matter density, the Iris pipeline was used (Bron et al., 2014). To compute these maps a group template space was defined using a procedure that avoids bias towards any of the individual T1w images using pairwise registration (Seghers et al., 2004). The pairwise registrations were performed using a similarity, affine, and nonrigid B-spline transformation model consecutively. We selected a subset of images for the definition of the template space. This template set consisted of the images of 50 ADNI participants that were randomly selected preserving the ratio between diagnostic groups (subject list available at https://gitlab.com/radiology/neuro/bron-cross-cohort[3]). The other images of both ADNI and PND data sets were registered to the template space following the same registration procedure. For the current work, some changes to the template space construction procedure as used in Bron et al., 2014 were made: non-uniformity correction was performed, skull-stripping was performed, and the template space corresponded to MNI-space. Using similarity

registration based on brain masks, we computed the coordinate transformations of MNI space to each of the template set's images, which were subsequently concatenated with the pairwise transformations before averaging. After template space construction, probabilistic GM maps were obtained with the unified tissue segmentation method of SPM8 (Statistical Parametric Mapping) (Ashburner and Friston, 2005). To obtain the final feature maps, probabilistic GM maps were transformed to the template space and modulated, i.e. multiplied by the Jacobian determinant of the deformation field, to take compression and expansion into account (Ashburner and Friston, 2000). To correct for head size, modulated GM maps were divided by intracranial volume.

Classification approaches

Two machine learning approaches were used for classification: a support vector machine (SVM) and a convolutional neural network (CNN).

Support vector machine (SVM)

An SVM with a linear kernel was used as this approach previously showed good performance using voxel-based features for AD classification. (Klöppel et al., 2008; Cuingnet et al., 2011; Bron et al., 2014; Bron et al., 2015). The c-parameter was optimized with 5-fold cross-validation on the training set. Input features, i.e. voxel values of the pre-processed images within a brain mask, were normalized to zero mean and unit variance based on the training set. The classifier was implemented using Scikit-Learn.

To gain insight into the classifications, we calculated statistical significance maps (p-maps) that show which features contributed to the SVM decision. These maps were computed using an analytical expression that approximates permutation testing (Gaonkar et al., 2015). Clusters of significant voxels were obtained using a p-value threshold of $\alpha \leqslant 0.05$. P-maps were not corrected for multiple comparisons, as permutation testing has a low false-positive rate (Gaonkar and Davatzikos, 2013).

Convolutional neural network (CNN)

An all convolutional neural network was used (Springenberg et al., 2015), which is a fully convolutional network (FCN) architecture that uses standard convolutional layers with stride two instead of the pooling layers used in most CNNs. This approach was chosen as it has previously shown good classification performance for AD based on structural MRI (Cui and Liu, 2019; Basaia et al., 2019). The used architecture is shown in Fig. 1.[4] Specifically, the network was built of 7 blocks consisting of a 3D convolutional layer (filter size 3; stride 1), followed by dropout, batch normalization (BN), and a rectified linear unit (ReLU) activation function, succeeded by a second 3D convolutional layer (filter size 3; stride 2), dropout, BN, and ReLU activation (Cui and Liu, 2019; Basaia et al., 2019). The number of filters changed over blocks: 16 filters in block 1, 32 in block 2 and 3, 64 in block 4 and 5, 32 in block 6, and 16 in block 7. The final output layer of the network was a softmax activation function, providing 2 prediction values (1 per class). The total network consisted of 577,498 parameters. For artificially increasing the training data set and for removing the class imbalance, data augmentation was used. The training set was augmented to 1000 samples per class based on the 'mixup' approach (Eaton-Rosen et al., 2018; Zhang et al., 2018). Mixup is a data-agnostic augmentation approach that is not based on spatial transformations, and therefore does not degrade the spatial normalization. Augmented samples were constructed by linearly combining two randomly selected images of the same class: a fraction of 80% of the first image was added to a fraction of 20% of the second image.

The network was compiled with a binary cross-entropy loss function

Table 2
Demographics for the PND data set. FU: follow-up time.

	AD	SCD	MCIc	MCInc
# participants	199	138	48	91
male/ female	94/ 105	92/ 46	33/ 15	56/ 35
age (y; mean±std)	73.1 ± 9.6	63.2 ± 10.3	70.4 ± 7.9	68.8 ± 12.6
FU (y; mean±std)	N.A.	N.A.	2.7 ± 1.2	2.2 ± 0.8

---

[4] Figure created with https://alexlenail.me/NN-SVG

Table 3

An overview of T1-weighted imaging protocols in the PND data set from eight centers. All sequences were 3D and used gradient recalled echo (GRE). TFE: turbo field echo, FSPGR: fast spoiled GRE, TFL: turboflash, MPRAGE: magnetization prepared rapid gradient echo, MP: magnetization prepared, SS: steady state, SP: spoiled, IR: inversion recovery, Sag: sagittal, Cor: Coronal, Ax: axial, TE: echo time, TR: repetition time, TI: inversion time.

Field strength	3T						1.5T				
Vendor	Philips				GE	Siemens	Philips	GE		Siemens	
Number of scans	195	60	70	1	122	40	1	6	1	32	28
Sequence name	TFE	GRE	GRE	GRE	FSPGR	MPRAGE	GRE	FSPGR	FSPGR	TFL MPRAGE	MPRAGE
Sequence variant	MP	MP	MP	MP	SS/SP	IR/SP/ MP	MP	SS/SP	SS/SP	IR/SP/MP	IR/SP/ MP
Plane	Sag	Sag	Sag	Cor	Sag	Sag	Sag	Sag	Ax	Sag	Sag
Slice thickness (mm)	1	1	1	1.4	1	1	2	1.6	1.6	1	1
In-plane (mm*mm)	1*1	0.78*0.78	0.5*0.5	0.88*0.88	0.94*0.94	1*1	0.47*0.47	0.47*0.47	1*1	1*1	0.5*0.5
TE (ms)	4.2	4.6	3.5	4.6	3	4.7	4.6	2.1	4.2	3	3.7
TR (ms)	8.1	9.9	9	9.6	7.8	2300	15	7.1	8.3	2000	2700
TI (ms)	N. A.	N.A.	N.A.	N.A.	450	1100	N.A.	450	450	1100	950
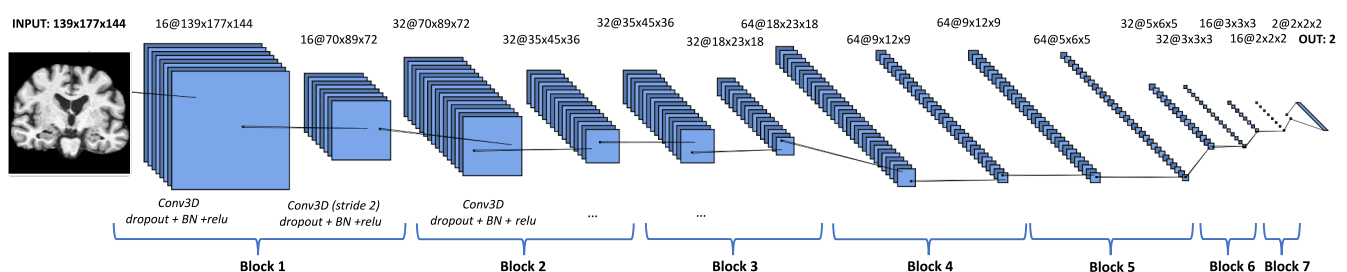
Fig. 1. CNN architecture[3].

and Adam optimizer (learning rate = 0.001, epsilon = 1e-8, decay = 0.0). To facilitate a stable convergence, learning rate followed a step decay schedule, i.e. after each ten epochs the learning rate was divided by two. The dropout rate was set to 20%. Data was propagated through the network with a batch size of 4. Input images were normalized to zero mean and unit variance based on the augmented training set. A validation set was created by randomly splitting 10% of the training data which was not used for training but only for regularization by early stopping, i.e. training was stopped when the validation AUC had not increased for 20 epochs. The model of the epoch with the highest validation AUC was selected as final model. Implementation was based on Keras and Tensorflow.

To gain insight into the classifications, we made saliency maps that show which parts of the brain contributed the most to the prediction of the CNN, i.e. which voxels lead to increase/decrease of prediction score when changed. Saliency maps were made using guided backpropagation, changing the activation function of the output layers from softmax to linear activations (Springenberg et al., 2015). Maps were averaged over correctly classified AD patients (Rieke et al., 2018).

Analysis and statistics

Classification performance was quantified by the area under the curve (AUC) and accuracy. For AD-CN classification, the data of the ADNI AD and CN groups were randomly split for 20 iterations preserving relative class sizes in each training and testing sample, using 90% for training and 10% for testing. Random splits were the same for both SVM and CNN. In each iteration, classification model parameters were optimized on the training set as explained above. The models were optimized solely on the training set; the test set was used only for evaluation of the final model. Ninety-five percent confidence intervals (95%CI) for the mean performance measures were constructed using the corrected resampled t-test based on the 20 cross-validation iterations, thereby taking into account that the samples in the cross-validation splits were

not statistically independent (Nadeau and Bengio, 2003).

Subsequently, we retrained classifiers using all AD and CN participants from the ADNI as training set. These retrained classifiers were used for visualization and their performance was evaluated on three independent test sets: ADNI MCIc-MCInc, PND AD-SCD, and PND MCIc-MCInc. 95%CIs were obtained based on 500 bootstrap samples of the test set. Significant differences between classifiers were assessed using the non-parametric McNemar Chi-square test (Dietterich, 1998) ($\alpha < 0.013$ after Bonferroni correction for 4 comparisons in each test set).

Trained models, lists of included subjects and all code used in preparation of this article are available from https://gitlab.com/radiology/neuro/bron-cross-cohort[3].

Results

Cross-validation performance for the ADNI AD-CN classification is shown in Fig. 2. For SVM, the AUC using modulated GM maps (0.940, 95%CI: 0.924 −0.955) was higher than the AUC using T1w images (0.801, 95%CI: 0.765 −0.837). For CNN, the same effect was observed, with modulated GM maps yielding a higher AUC (0.933, 95%CI: 0.918 −0.948) than T1w images (0.898, 95%CI: 0.875 −0.920), albeit here with overlapping confidence intervals. For classification based on modulated GM maps, the AUC for SVM (0.940, 95%CI: 0.924 −0.955) was similar to that of CNN (0.933, 95%CI: 0.918 −0.948). Accuracy measures showed the same patterns.

The performance of the classifiers trained on all ADNI AD and CN data to predict MCI conversion is shown in Fig. 3. While AUCs with both SVM and CNN were slightly higher for modulated GM maps than for T1w images, the accuracy measures showed similar performance for both inputs. Using modulated GM maps, performance for SVM (AUC = 0.756, 95%CI: 0.720 −0.788; accuracy = 0.695, 95%CI: 0.665 −0.723) was higher than for CNN (AUC = 0.742, 95%CI: 0.709 −0.776); accuracy = 0.658, 95%CI: 0.628 −0.690). This difference was significant
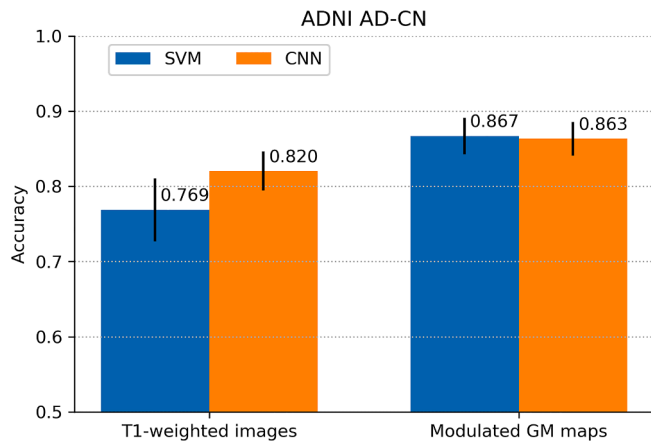
Fig. 2. Cross-validation performance for classification of the Alzheimer's disease patients (AD) and controls (CN) of the ADNI data set expressed by (a) area under-the-ROC–curve (AUC) and (b) accuracy. Performance is shown for SVM and CNN classifiers using two inputs: minimally processed T1w scans and modulated GM images. Error bars indicate 95%CIs.

according to McNemar's test ($p < 0.01$).

The performance of external validation, i.e. the application of the classifiers in the PND data set, is shown in Fig. 4. For AD-SCD diagnosis, the AUC for SVM was 0.896 (95%CI: 0.855 −0.932) and that for CNN was 0.876 (95%CI: 0.836 −0.913). Both AUC and accuracy followed the same patterns as in ADNI: SVM and CNN showed similar performance and modulated GM maps yielded higher classification performance than minimally processed T1w images (McNemar's test; $p < 0.01$ for SVM, $p = 0.01$ for CNN). Performances were however slightly lower; PND confidence intervals for AUC (but not for accuracy) overlapped with those of ADNI.

For prediction of MCI conversion in PND, classification performance was also lower than that in ADNI. For the GM modulated maps, the AUC for CNN was 0.702 (95%CI: 0.624 −0.786) and that for SVM was 0.665 (95%CI: 0.576 −0.760). Confidence intervals were relatively large and overlapped with those in the ADNI data. No significant differences between classifiers and between pre-processing approaches were seen.

Brains regions that contributed to the classifications are visualized using SVM p-maps in Fig. 5 and using CNN saliency maps in Fig. 6. The SVM p-map for the minimally processed T1w images showed small clusters of significant voxels, mainly located in the medial temporal lobe (hippocampus), around the ventricles and at larger sulci at the outside of the brain. For modulated GM maps, clusters of significant voxels in the

p-map were larger and predominantly visible in the hippocampus. In addition, smaller clusters were located in the rest of the temporal lobe and the cerebellum. CNN saliency maps showed a very limited contribution of the temporal lobe. Instead, the saliency map for the T1w images mainly showed contribution of voxels at the edge of the brain, in white matter regions around the ventricles and in the cerebellum. For modulated GM maps, clusters of contributing voxels were located in the subcortical structures, the white matter around the ventricles and the cerebellum.

Discussion

We performed a comparative study focusing on the generalizability of diagnostic and predictive performance of machine learning based on MRI data of the ADNI research cohort, to the PND multi-center data set representing a tertiary memory clinic population. Both cross-validation and external validation results for AD-CN diagnosis showed similar performance using the used deep learning classifier and conventional classifier. Both approaches significantly benefited from the use of modulated GM maps instead of raw T1w images. Application to MCI conversion prediction yielded higher performance for SVM than for CNN in ADNI, but this was not seen in PND. Performances were in line with the state-of-the-art (Rathore et al., 2017; Wen et al., 2020; Ansart et al., 2021). For MCI conversion prediction, Ansart et al., 2021 showed that performance of current methods converges to an AUC of about 75% as the number of subjects increases, which aligns with our results.

While in many medical imaging applications CNNs convincingly outperformed conventional classifiers (Litjens et al., 2017), our results showed similar performance for CNN and SVM, which confirms the findings by Wen et al., 2020. Other CNN designs could possibly improve on this, but we made an effort to follow the state-of-the-art for CNN design. Promising developments to further improve performance could come from changes in network architecture (e.g., successful standard architectures like InceptionNet or ResNet, adversarial training, discriminative auto-encoders) and improvements in data collection and handling (e.g., larger datasets to learn more complex models, or pre-training on other collections of brain imaging data). In addition, data augmentation could play a role in further improvement. While a strength of the mix-up approach is that it is data-agnostic, an augmentation approach using for example prior knowledge may have added value. This work shows that the need for dedicated pre-processing is lower for CNN than for SVM, but nevertheless has an added value for the performance. While we evaluated only one implementation of the pre-processing procedure (Bron et al., 2014), we expect that alternative implementations (e.g. SPM12, FSL-VBM) could have slightly changed
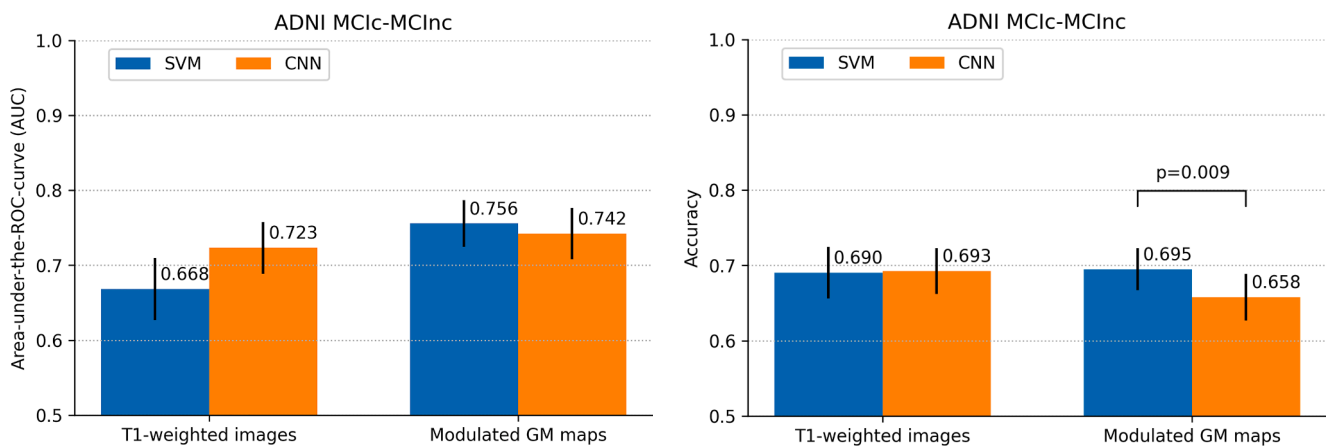

Fig. 3. Classification performance of patients with mild cognitive impairment (MCI) that do or do not convert to Alzheimer's disease (MCIc vs MCInc) in the ADNI data set expressed by (a) area under-the-ROC–curve (AUC) and (b) accuracy. Performance is shown for SVM and CNN classifiers using two inputs: minimally processed T1w scans and modulated GM images. Error bars indicate 95%CIs. P-values for significant differences are shown in (b).
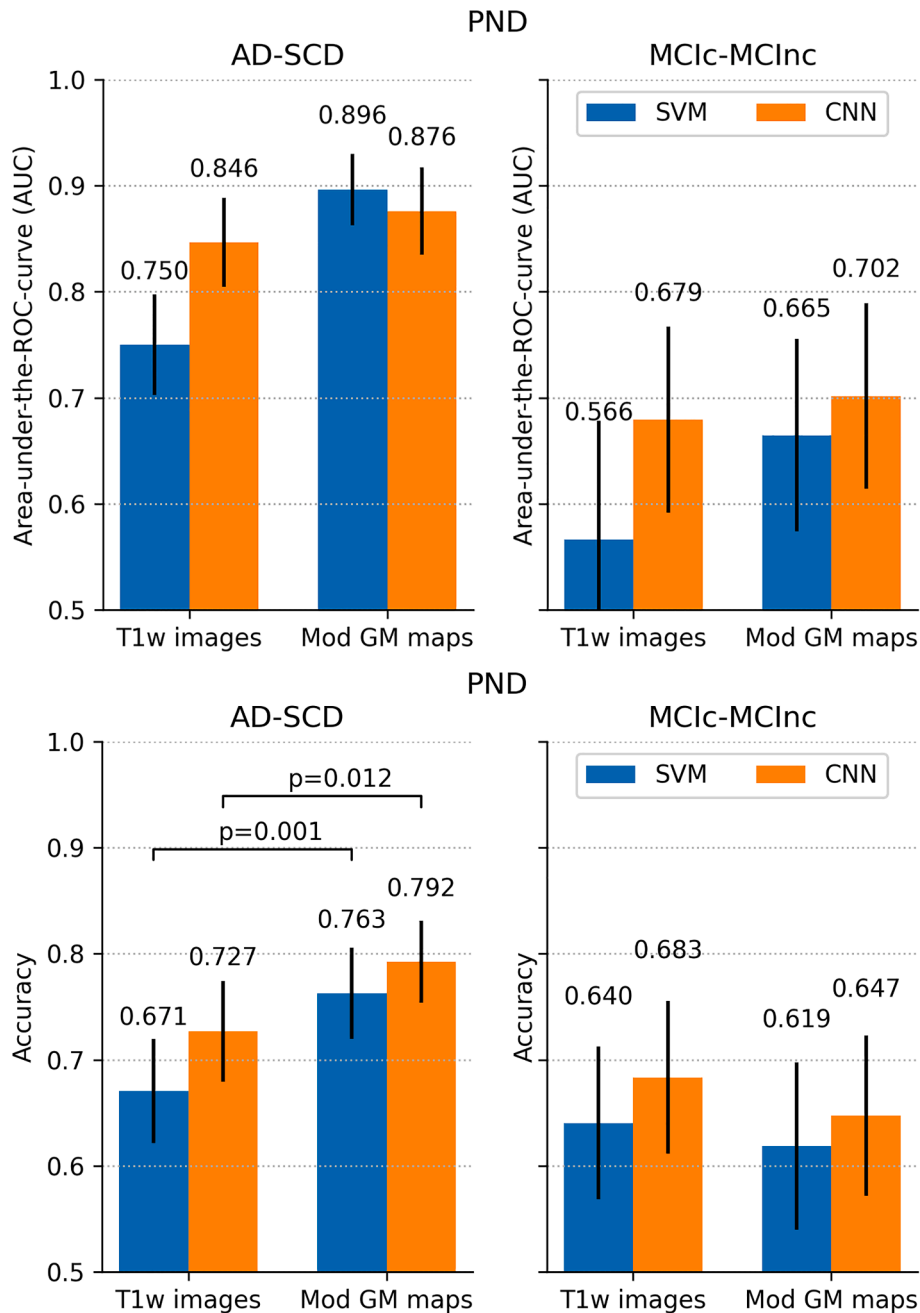
Fig. 4. Classification performance on the PND data set: (a) area under-the-ROC–curve (AUC) and (b) accuracy. Classifiers were trained on ADNI AD-CN and applied to PND AD-SCD (left figures) and PND MCIc-MCInc (right figures). Performance is shown for SVM and CNN classifiers using two inputs: minimally processed T1w scans and modulated GM images. Error bars indicate 95%CIs. P-values for significant differences are shown in (b).

results but would have led to the same conclusions. With sufficiently large datasets the need for dedicated pre-processing including spatial normalization may reduce.

Although SVM and CNN classifiers yielded similar performance, their visualizations showed different brain regions to be involved in the classification. SVM significance maps showed a clear contribution of the hippocampus and medial temporal lobe as previously shown and expected based on prior knowledge (Bron et al., 2017). CNN saliency maps showed involvement of subcortical structures, regions prone to white matter hyperintensities and the cerebellum. For both classifiers, classification based on minimally processed T1w images showed voxels at the edge of the brain to be involved, which is expected as only similarity transformation to template space had been performed. In addition to the brain edges, the CNN classifier, which outperformed the SVM for these

minimally processed input images, also highlights regions similar to those shown by the saliency map for the modulated GM images. This may implicate that the CNNs non-linear operations, in contrast to the linear kernel of the SVM, could extract feature maps that partly resemble GM modulated maps. The regions highlighted by the CNN saliency maps could possibly be related to AD using prior knowledge, but we will refrain from over-interpretation here. It is however unexpected that the medial temporal lobe is not covered as previously shown with CNN saliency maps on ADNI data (Dyrba et al., 2020; Rieke et al., 2018). Differences between the SVM and CNN classifiers in involved brain regions could be contributed to both the differences in the classification approaches as well as to the differences in the used visualization techniques. If the first reason dominates, hence if the classifiers actually use different brain regions, combining classifiers into a hybrid approach
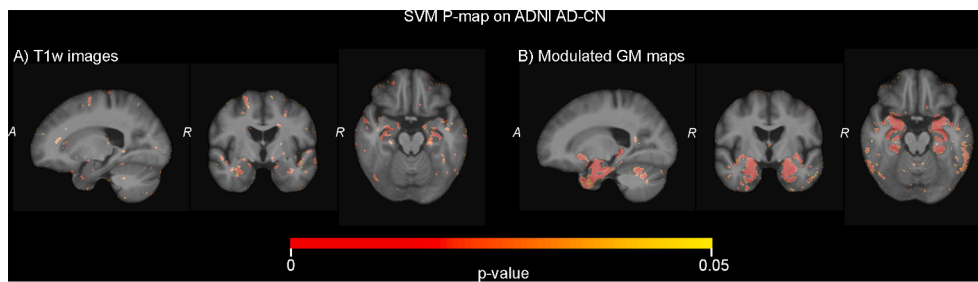
Fig. 5. Visualization of the SVM classifiers using analytic significance maps (p-maps) based on two inputs: (a) minimally processed T1w images and (b) modulated GM maps.
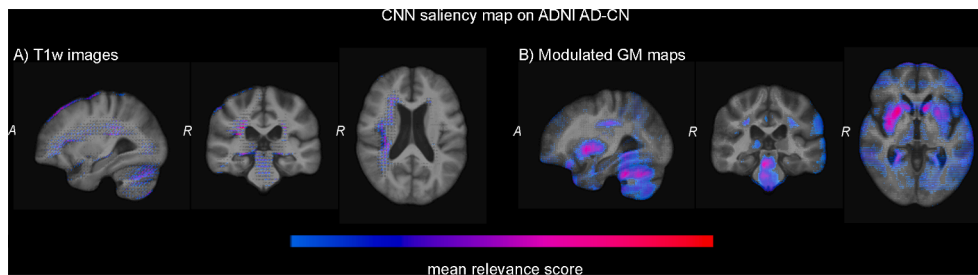

Fig. 6. Visualization of the CNN classifiers using guided back-propagation saliency maps based on two inputs: (a) minimally processed T1w images and (b) modulated GM maps. Relevance maps were averaged over all correctly classified AD participants and thresholded at $\frac{1}{3}$ of the maximum intensity.

would be an interesting future direction. However, for full understanding of brain regions involved in CNN-based classification of AD, further research is required.

This work is one of the few to address how AD classification performance of MRI-based machine learning generalizes to an independent cohort (Wen et al., 2020; Hall et al., 2015; Bouts et al., 2019; Archetti et al., 2019). On the PND data, the resulting AUC values (0.896 for SVM, 0.876 for CNN) were competitive with values reported for AD-CN in the literature, but still they were 0.04–0.07 lower than those in the ADNI cross-validation experiment. The main patterns in the results corresponded between ADNI and PND data, i.e. similar performance for SVM and CNN and added value of dedicated MRI processing. For prediction in MCI, AUC values in the PND data set were 0.04–0.10 lower than those in ADNI. Overall, similar to experiments by Wen et al. (2020) and Hall et al. (2015), we observed only a minor performance drop. This largely preserved performance could be related to the similarities between the ADNI and PND studies that include a multi-center set-up, within-study standardization of cognitive protocols, and diagnostic criteria for AD (McKhann et al., 1984; McKhann et al., 2011) and MCI (Petersen, 2004). The performance reduction could be contributed to differences between the studies, such as the MRI protocols (all high resolution T1w, but more homogeneous within ADNI than within PND), country of origin (United States vs. the Netherlands), control population (a combination of cognitively normal and SCD vs. SCD only), MCI population (amnestic MCI only vs. a broad MCI group) and patient inclusion criteria (ADNI used hard cut-offs on cognitive scores and clinical dementia rating whereas PND did not) (Petersen et al., 2010; Aalten et al., 2014). Studies that found much worse generalizability in their experiments described larger differences in inclusion and diagnostic criteria between training data and validation data than we did (Bouts et al., 2019; Wen et al., 2020).

A limitation of this study is that the diagnosis was based on clinical criteria rather than post-mortem histopathological examination. Although diagnosis was typically confirmed by follow-up, it is possible that some of the patients were misdiagnosed. An alternative could be to use amyloid data from PET imaging or cerebrospinal fluid to classify AD pathology instead of relying on the clinical diagnosis (e.g., Son et al.,

2020). In addition, because of the limited availability of diagnostic information at follow-up in the PND data set, its MCI data is relatively small. This is reflected by the large confidence intervals for the performance metrics in the prediction task. To maximize the number of PND MCI participants, we chose to use the last available time point for final diagnosis. As a result the time-to-prediction ranged between 1–5 years, whereas for ADNI a fixed time interval of three years is chosen. As time-to-prediction is related to predictive performance (Ansart et al., 2021), a fixed time interval would be preferred for inter-cohort performance comparison.

While the external validation performance was quite high, as expected some performance drop was observed. Therefore, research focusing on approaches to mitigate such performance drops, such as transfer learning, is highly relevant (Wachinger and Reuter, 2016). In addition, whereas this work only exploited structural MRI, other works have shown that performance can be increased with the use of multimodal inputs, i.e. cognitive test scores, fluid-based biomarker measurements, genetic information and other imaging modalities such as PET, diffusion MRI or perfusion MRI (Bron et al., 2017; Ansart et al., 2021; Venkatraghavan et al., 2019). While multi-modal classification would therefore be a logical and important extension, this may also lead to a decrease of generalizability as chances of differences between studies increase with multiple modalities.

In conclusion, classification performance of ADNI data generalized well to the multi-center PND biobank cohort representing tertiary memory clinic patients, with only a minor drop in performance. Conventional SVM classifiers and deep learning approaches using CNN showed comparable results, and both methods benefited from dedicated MRI processing using GM modulated maps. We hope that external validation results like those presented here will contribute to setting next steps towards the implementation of machine learning in clinical practice for aiding diagnosis and prediction.

CRediT authorship contribution statement

Esther E. Bron: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.nicl.2021.102712.

References

Aalten, P., Ramakers, I.H., Biessels, G.J., de Deyn, P.P., Koek, H.L., Olde Rikkert, M.G., Oleksik, A.M., Richard, E., Smits, L.L., van Swieten, J.C., Teune, L.K., van der Lugt, A., Barkhof, F., Teunissen, C.E., Rozendaal, N., Verhey, F.R., van der Flier, W. M., 2014. The Dutch Parelsnoer Institute – neurodegenerative diseases; methods, design and baseline results. BMC Neurol. 14, 1–8. https://doi.org/10.1186/s12883-014-0254-4.

Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T., Couronne, R., Faouzi, J., Koval, I., Louis, M., Couronné, R., Thibeau-Sutre, E., Wen, J., Wild, A., Burgos, N., Dormont, D., Colliot, O., Durrleman, S., 2021. Predicting the progression of mild cognitive impairment using machine learning: a systematic and quantitative review. Med. Image Anal. 67, 101848 https://doi.org/10.1016/j.media.2020.101848.

Arbabshirani, M.R., Plis, S.M., Sui, J., Calhoun, V.D., 2017. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage 145, 137–165. https://doi.org/10.1016/j.neuroimage.2016.02.079.

Archetti, D., Ingala, S., Venkatraghavan, V., Wottschel, V., Young, A.L., Bellio, M., Bron, E.E., Klein, S., Barkhof, F., Alexander, D.C., Oxtoby, N.P., Frisoni, G.B., Redolfi, A., 2019. Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease. Neuroimage Clin. 24, 101954 https://doi.org/10.1016/j.nicl.2019.101954.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry – the methods. Neuroimage 11, 805–821. https://doi.org/10.1006/nimg.2000.0582.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851. https://doi.org/10.1016/j.neuroimage.2005.02.018.

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., Filippi, M., 2019. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. NeuroImage: Clin. 21, 101645 https://doi.org/10.1016/j.nicl.2018.101645.

Bouts, M.J.R.J., Grond, J.V.D., Vernooij, M.W., Koini, M., Cremers, L.G.M., Schouten, T. M., Vos, F.D., Feis, R.A., Ikram, M.A., Rombouts, S.A.R.B., Lechner, A., Schmidt, R., Rooij, M.D., Niessen, W.J., 2019. Detection of mild cognitive impairment in a community-dwelling population using quantitative, multiparametric MRI-based classification. Hum Brain Mapp. 1–12 https://doi.org/10.1002/hbm.24554.

Bron, E.E., Smits, M., van der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M.E., Méndez Orellana, C., Meijboom, R., Pinto, M., Meireles, J.R., Garrett, C., Bastos-Leite, A.J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cárdenas-Peña, D., Álvarez-Meza, A.M., Dolph, C.V., Iftekharuddin, K.M., Eskildsen, S.F., Coupé, P., Fonov, V.S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K.R., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Di Fatta, G., Sensi, F., Chincarini, A., Smith, G.M., Stoyanov, Z.V., Sørensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J.C., Niessen, W.J., Klein, S., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. Neuroimage 111, 562–579. https://doi.org/10.1016/j.neuroimage.2015.01.048.

Bron, E.E., Smits, M., Papma, J.M., Steketee, R.M.E., Meijboom, R., de Groot, M., van Swieten, J., Niessen, W.J., Klein, S., 2017. Multiparametric computer-aided differential diagnosis of Alzheimer's disease and frontotemporal dementia using structural and advanced MRI. Eur. Radiol. 27, 3372–3382. https://doi.org/10.1007/s00330-016-4691-x.

Bron, E.E., Steketee, R.M.E., Houston, G.C., Oliver, R.A., Achterberg, H.C., Loog, M., van Swieten, J.C., Hammers, A., Niessen, W.J., Smits, M., Klein, S., 2014. Diagnostic classification of arterial spin labeling and structural MRI in presenile early stage dementia. Hum. Brain Mapp. 35, 4916–4931. https://doi.org/10.1002/hbm.22522.

Cui, R., Liu, M., 2019. RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. Comput. Med. Imaging Graph. 73, 1–10. https://doi.org/10.1016/j.compmedimag.2019.01.005.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O.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. https://doi.org/10.1016/j.neuroimage.2010.06.013.

Dietterich, T., 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput. 10, 1895–1924. https://doi.org/10.1162/089976698300017197.

Dyrba, M., Pallath, A.H., Marzban, E.N., 2020. Comparison of CNN Visualization Methods to Aid Model Interpretability for Detecting Alzheimer's Disease.

Bildverarbeitung für die Medizin 29267. https://doi.org/10.1007/978-3-658-29267-6.

Eaton-Rosen, Z., Bragman, F., Ourselin, S., Cardoso, M.J., 2018. Improving Data Augmentation for Medical Image Segmentation, in: Medical Imaging with Deep Learning, p. 1.

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. J. Alzheimer Disease 41, 685–708. https://doi.org/10.3233/JAD-131928.

Gaonkar, B., Davatzikos, C., 2013. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. Neuroimage 78, 270–283. https://doi.org/10.1016/j.neuroimage.2013.03.066.

Gaonkar, B., Shinohara, R.T., Davatzikos, C., 2015. Interpreting support vector machine models for multivariate group wise analysis in neuroimaging. Med. Image Anal. 24, 190–204. https://doi.org/10.1016/j.media.2015.06.008.

Gordon, B.A., Blazey, T.M., Su, Y., Hari-Raj, A., Dincer, A., Flores, S., Christensen, J., McDade, E., Wang, G., Xiong, C., Cairns, N.J., Hassenstab, J., Marcus, D.S., Fagan, A. M., Jack, C.R., Hornbeck, R.C., Paumier, K.L., Ances, B.M., Berman, S.B., Brickman, A.M., Cash, D.M., Chhatwal, J.P., Correia, S., Förster, S., Fox, N.C., Graff-Radford, N.R., la Fougère, C., Levin, J., Masters, C.L., Rossor, M.N., Salloway, S., Saykin, A.J., Schofield, P.R., Thompson, P.M., Weiner, M.M., Holtzman, D.M., Raichle, M.E., Morris, J.C., Bateman, R.J., Benzinger, T.L., 2018. Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant Alzheimer's disease: a longitudinal study. Lancet Neurol. 17, 241–250. https://doi.org/10.1016/S1474-4422(18)30028-0.

Hall, A., Muñoz-Ruiz, M., Mattila, J., Koikkalainen, J., Tsolaki, M., Mecocci, P., Kloszewska, I., Vellas, B., Lovestone, S., Visser, P.J., Lötjonen, J., Soininen, H., 2015. Generalizability of the disease state index prediction model for identifying patients progressing from mild cognitive impairment to Alzheimer's disease. J. Alzheimer Disease 44, 79–92. https://doi.org/10.3233/JAD-140942.

Hosseini-Asl, E., Ghazal, M., Mahmoud, A., Aslantas, A., Shalaby, A., Casanova, M., Barnes, G., Gimel'farb, G., Keynton, R., Baz, A.E., 2018. Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network. Front. Biosci. 23, 584–896. https://doi.org/10.2741/4606.

Jack, C.R., Barnes, J., Bernstein, M.A., Borowski, B.J., Brewer, J., Clegg, S., Dale, A.M., Carmichael, O., Ching, C., DeCarli, C., Desikan, R.S., Fennema-Notestine, C., Fjell, A. M., Fletcher, E., Fox, N.C., Gunter, J., Gutman, B.A., Holland, D., Hua, X., Insel, P., Kantarci, K., Killiany, R.J., Krueger, G., Leung, K.K., Mackin, S., Maillard, P., Malone, I.B., Mattsson, N., McEvoy, L., Modat, M., Mueller, S., Nosheny, R., Ourselin, S., Schuff, N., Senjem, M.L., Simonson, A., Thompson, P.M., Rettmann, D., Vemuri, P., Walhovd, K., Zhao, Y., Zuk, S., Weiner, M., 2015. Magnetic resonance imaging in Alzheimer's Disease Neuroimaging Initiative 2. Alzheimers Dement 11, 740–756. https://doi.org/10.1016/j.jalz.2015.05.002.

Jack, C.R., Bernstein, M., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging 27, 685–691. doi: 10.1002/jmri.21049.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W., 2010. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging 29, 196–205. https://doi.org/10.1109/TMI.2009.2035616.

Klöppel, S., Abdulkadir, A., Jack, C.R., Koutsouleris, N., Mourão-Miranda, J., Vemuri, P., 2012. Diagnostic neuroimaging across diseases. Neuroimage 61, 457–463. https://doi.org/10.1016/j.neuroimage.2011.11.002.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, I., Rohrer, J.D., Fox, N.C., Jack Jr, C.R., Ashburner, J., Frackowiak, R.S.J., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131, 681–689. https://doi.org/10.1093/brain/awm319.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005.

Manniën, J., Ledderhof, T., Verspaget, H.W., Snijder, R.R., Flikkenschild, E.F., Scherrenburg, N.P.C.V., Stolk, R.P., Zielhuis, G.A., 2017. The Parelsnoer Institute: A National Network of Standardized Clinical Biobanks in the Netherlands. Open J. Bioresour. 4, 3. https://doi.org/10.5334/ojb.23.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E.M., 1984. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 34, 939–944. https://doi.org/10.1212/wnl.34.7.939.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack Jr., C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., Mohs, R.C., Morris, J.C., Rossor, M.N., Scheltens, P., Carillo, M.C., Thies, B., Weintraub, S., Phelps, C.H., 2011. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 7, 263–269. https://doi.org/10.1016/j.jalz.2011.03.005.

Mehta, D., Jackson, R., Paul, G., Shi, J., Sabbagh, M., 2017. Why do trials for Alzheimer's disease drugs keep failing? Expert Opin. Investig. Drugs 26, 735–739. https://doi.org/10.1080/13543784.2017.1323868.

Nadeau, C., Bengio, Y., 2003. Inference for the generalization error. Mach. Learn. 52, 239–281. https://doi.org/10.1023/A:1024068626366.

Petersen, R.C., 2004. Mild cognitive impairment as a diagnostic entity. J. Intern Med. 256, 183–194. https://doi.org/10.1111/j.1365-2796.2004.01388.x.

Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Jack, C.R., Jagust, W.J., Shaw, L.M., Toga, A.W., Trojanowski, J.Q., Weiner, M.W., 2010. Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. Neurology 74, 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25.

Prince, M., Bryce, R., Ferri, C., 2011. World Alzheimer Report 2011, The benefits of early diagnosis and intervention. Alzheimer's Disease International.

Rathore, S., Habes, M., Aksam Iftikhar, M., 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. https://doi.org/10.1016/j.neuroimage.2017.03.057.

Rieke, J., Eitel, F., Weygandt, M., Haynes, J.D., Ritter, K., 2018. Visualizing convolutional networks for MRI-based diagnosis of alzheimer's disease. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11038 LNCS, 24–31. https://doi.org/10.1007/978-3-030-02628-8_3.

Seghers, D., D'Agostino, E., Maes, F., Vandermeulen, D., Suetens, P., 2004. Construction of a brain template from MR images using state-of-the-art registration and segmentation techniques, in: Proc Intl Conf Med Image Comput Comp Ass Intervent, Springer. pp. 696–703. doi: 10.1007/978-3-540-30135-6_85.

Shamonin, D.P., Bron, E.E., Lelieveldt, B.P., Smits, M., Klein, S., 2014. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. Front. Neuroinform. 7, 1–15. https://doi.org/10.3389/fninf.2013.00050.

Son, H.J., Oh, J.S., Oh, M., Kim, S.J., Lee, J.H., Roh, J.H., Kim, J.S., 2020. The clinical feasibility of deep learning-based classification of amyloid PET images in visually equivocal cases. Eur. J. Nucl. Med. Mol. Imaging 47, 332–341. https://doi.org/10.1007/s00259-019-04595-y.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for simplicity: The all convolutional net. In: 3rd International Conference on Learning Representations, p. 1 http://arxiv.org/abs/1412.6806 arXiv:1412.6806.

Suk, H.I., Lee, S.W., Shen, D., 2017. Deep ensemble learning of sparse regression models for brain disease diagnosis. Med. Image Anal. 37, 101–113. https://doi.org/10.1016/j.media.2017.01.008.

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 Trans. Med. Imaging 29, 1310–1320. https://doi.org/10.1109/TMI.2010.2046908.

Van Vliet, D., De Vugt, M.E., Bakker, C., Pijnenburg, Y.A., Vernooij-Dassen, M.J., Koopmans, R.T., Verhey, F.R., 2013. Time to diagnosis in young-onset dementia as compared with late-onset dementia. Psychol. Med. 43, 423–432. https://doi.org/10.1017/S0033291712001122.

Venkatraghavan, V., Bron, E., Niessen, W., Klein, S., 2019. Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling. Neuroimage 186. https://doi.org/10.1016/j.neuroimage.2018.11.024.

Vieira, S., Pinaya, W.H., Mechelli, A., 2017. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. Neurosci. Biobehav. Rev. 74, 58–75. https://doi.org/10.1016/j.neubiorev.2017.01.002.

Wachinger, C., Reuter, M., 2016. Domain Adaptation for Alzheimer's Disease Diagnostics. Neuroimage 139, 470–479. https://doi.org/10.1016/j.neuroimage.2016.05.053.

Wen, J., Thibeau-Sutre, E., Samper-Gonzalez, J., Routier, A., Bottani, S., Durrleman, S., Burgos, N., Colliot, O., 2020. Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation. Med. Image Anal. 63 https://doi.org/10.1016/j.media.2020.101694.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018. Mixup: Beyond empirical risk minimization. Int Conf Learn Repres http://arxiv.org/abs/1710.09412 arXiv:1710.09412.