Computers in Biology and Medicine 170 (2024) 108029

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Systematic comparison of 3D Deep learning and classical machine learning explanations for Alzheimer's Disease detection

Louise Bloch [a,b,c,1], Christoph M. Friedrich [a,b,2,*], for the Alzheimer's Disease Neuroimaging Initiative[3]

[a] Department of Computer Science, University of Applied Sciences and Arts Dortmund (FH Dortmund), Emil-Figge-Straße 42, Dortmund, 44227, North Rhine-Westphalia, Germany

[b] Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Hufelandstraße 55, Essen, 45122, North Rhine-Westphalia, Germany

[c] Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen, Hufelandstraße 55, Essen, 45122, North Rhine-Westphalia, Germany

ARTICLE INFO

Keywords:
Alzheimer's Disease
Interpretable Machine Learning
SHAP
LIME
GradCAM
3D CNN

ABSTRACT

Black-box deep learning (DL) models trained for the early detection of Alzheimer's Disease (AD) often lack systematic model interpretation. This work computes the activated brain regions during DL and compares those with classical Machine Learning (ML) explanations. The architectures used for DL were 3D DenseNets, EfficientNets, and Squeeze-and-Excitation (SE) networks. The classical models include Random Forests (RFs), Support Vector Machines (SVMs), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting (LightGBM), Decision Trees (DTs), and Logistic Regression (LR). For explanations, SHapley Additive exPlanations (SHAP) values, Local Interpretable Model-agnostic Explanations (LIME), Gradient-weighted Class Activation Mapping (GradCAM), GradCAM++ and permutation-based feature importance were implemented. During interpretation, correlated features were consolidated into aspects. All models were trained on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. The validation includes internal and external validation on the Australian Imaging and Lifestyle flagship study of Ageing (AIBL) and the Open Access Series of Imaging Studies (OASIS).

DL and ML models reached similar classification performances. Regarding the brain regions, both types focus on different regions. The ML models focus on the inferior and middle temporal gyri, and the hippocampus, and amygdala regions previously associated with AD. The DL models focus on a wider range of regions including the optical chiasm, the entorhinal cortices, the left and right vessels, and the 4th ventricle which were partially associated with AD. One explanation for the differences is the input features (textures vs. volumes). Both types show reasonable similarity to a ground truth Voxel-Based Morphometry (VBM) analysis. Slightly higher similarities were measured for ML models.

1. Introduction

The most frequent cause of dementia [1] and thus a globally growing health problem is Alzheimer's Disease (AD). Currently, no causal therapy can cure AD [2]. The development of pre-clinical markers can help recruit subjects for therapy studies that aim to stop the disease progression among the AD continuum. The continuum includes different stages, e.g., cognitive normals (CN), Mild Cognitive Impairment (MCI) due to AD, probable AD, and AD dementia.

In previous research [3,4], Machine Learning (ML) helped to identify patterns in high-dimensional data to improve AD detection. Black-box ML models, e.g., Random Forests (RFs) [5], eXtreme Gradient

* Corresponding author at: Department of Computer Science, University of Applied Sciences and Arts Dortmund (FH Dortmund), Emil-Figge-Straße 42, Dortmund, 44227, North Rhine-Westphalia, Germany.

E-mail addresses: louise.bloch@fh-dortmund.de (L. Bloch), christoph.friedrich@fh-dortmund.de (C.M. Friedrich).

[1] 0000-0001-7540-4980

[2] 0000-0001-7906-0038

[3] 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, Access: 2024-01-23.

https://doi.org/10.1016/j.compbiomed.2024.108029
Received 9 July 2023; Received in revised form 25 January 2024; Accepted 25 January 2024
Available online 30 January 2024

Boosting (XGBoost) [6], or Convolutional Neural Networks (CNNs) [7] often outperform more interpretable models like Decision Trees (DTs) or Logistic Regression (LR). However, black-box models give no insights to prove their biological plausibility. Therefore, interpretable ML [8] explains black-box models. The explanation of black-box ML models shows high potential in medical imaging [9,10] and has been previously used in AD detection [4,11–13]. Deep learning (DL) explanations mostly focus on visually inspecting the heatmaps of exemplary individuals [14–16]. To the best of our knowledge, there is a lack of work that systematically compares the regional feature importances of ML and DL models.

To overcome this limitation, DL-based activations of cortical and subcortical brain structures are summarized systematically. The approach facilitates the quantitative comparison of DL and classical ML explanations. Multiple interpretation methods were compared to investigate relevant brain structures across methods and models. Those methods include permutation importance, SHapley Additive exPlanations (SHAP) [17], Local Interpretable Model-agnostic Explanations (LIME) [18], Gradient-weighted Class Activation Mapping (GradCAM) [19], and GradCAM++ [20]. For DL, 3D model architectures adapted from Dense Convolutional Networks (DenseNets) [21], EfficientNets [22], and Squeeze-and-Excitation (SE) networks [23] were implemented. RF [5], XGBoost [6], Light Gradient Boosting Machines (LightGBMs) [24], and Support Vector Machines (SVMs) [25] were trained as classical black-box ML models. DTs and LR serve as interpretable comparison models. As model calibration affects the results of explainability methods [26], the implementation of Platt scaling [27] reduced the model uncertainty. Additionally, permutation-based interpretation models like SHAP and LIME assume independent input variables. This assumption is not sustainable for real-world problems such as AD detection, and thus, another problem that affects the model interpretability is feature correlation [28]. During the permutation process, correlated features lead to synthetic subjects with unrealistic combinations of features. This effect usually results in reduced importance scores for correlated features compared to the resulting feature importance of independent features. Due to this bias, a comparison in terms of biological plausibility is not possible. To overcome this problem, in this work, correlated features are consolidated using aspects [29]. All models are trained on the Alzheimer's Disease Neuroimaging Initiative (ADNI) [30] dataset and validated for an ADNI test set, the external Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL) [31], and the Open Access Series of Imaging Studies (OASIS) [32] cohorts. The activated brain regions were compared with the results of a Voxel-Based Morphometry (VBM) analysis. The contributions of this paper can be summarized as follows:

- Systematic comparison of relevant brain regions for ML and DL models in AD detection.
- Comparison of region-based feature importance scores with a ground truth VBM analysis.
- Model calibration and aspect consolidation to avoid biased model explanations.
- Internal and external validation to guarantee model robustness.

Section 2 describes related work. The datasets and explainability methods are introduced in Section 3. The implementation of the ML workflow and the details of the experiments are described in Section 4. Section 5 elaborates on the experimental results, which are discussed, including the limitations, in Section 6. Finally, Section 7 concludes the work.

2. Related work

A comparison between the performance scores of 2D- and 3D-DL models as well as classical ML models, more precisely Support Vector Machines (SVMs), is presented in [33]. All models were trained on two different subsets of the ADNI dataset. The first subset contains 509 subjects (162 CN, 76 progressive MCI (pMCI), 134 stable MCI (sMCI), 137 AD), while the second subset consists of 264 subjects (164 pMCI, 100 sMCI). The SVMs were trained using two different feature sets. Aggregated selection [34] extracted the features of the first feature set from the voxels of preprocessed Magnetic Resonance Imaging (MRI) scans. For the second feature set, kernel partial least squares [35] was used for the extraction. Four pretrained 2D CNNs, namely AlexNet [36], GoogleNet [37], ResNet [38], and Inception-v3 [39] were fine-tuned using three single-channel 2D slices of an MRI, which were combined to build a three-channel image. The 3D CNN models were trained from scratch on 3D MRI patches. Among other things, the results show improved performance for the pretrained 2D model compared to the 3D CNN. The best AUC of 93.2% for the classification between CN and AD was reached for an ensemble of two SVM models. This model outperforms an ensemble of five 2D CNNs which achieved an AUC of 90.2%. An SVM ensemble trained only on the inner cerebral structures of the brain performs best for the sMCI vs. pMCI classification (AUC: 73.3%). In comparison to this work, the paper does not introduce an explainability component, which is important in the medical domain to make a system's decisions understandable to clinicians and patients.

Some studies [11,12,40] explained models by using model-specific feature importances. RF importance scores are, e.g., used in [40] to compare the most important biomarkers during the prediction of different AD disease stages. The results achieved for 405 ADNI subjects (148 CN, 147 MCI, 110 AD) showed, that Amyloid Positron Emission Tomography (PET) uptake is more important in early AD stages (CN vs. MCI), whereas neurodegeneration biomarkers including MRI volumetric features and Fluorodeoxyglucose (FDG)-PET uptake are more relevant in late stages (MCI vs. AD, CN vs. AD).

RF feature importances also explain models trained to distinguish between 340 sMCI and 173 pMCI ADNI subjects using volumetric MRI features of two visits, cognitive test scores, and demographic data in [12]. The cognitive test scores reached higher feature importance in comparison to volumetric MRI features. External validation for AIBL was performed using 22 subjects (14 sMCI, 8 pMCI).

Overall, model-specific feature importances are not suitable for the explanation of individual predictions, which are important in clinical practice and are covered by local, model-independent methods like SHAP [17] or LIME.

The global and local predictions of RFs and XGBoost models were explained in [41] using SHAP. The dataset includes sociodemographic and lifestyle factors to predict the patient's AD risk. Transfer learning reused information from the Survey of Health, Ageing, and Retirement in Europe (SHARE) [42] (80,699 CN, 4,157 AD) to the PREVENT cohort [43] (109 high AD risk, 364 low AD risk). The results showed that age is the most relevant risk factor. Further identified risk factors are less education, physical inactivity, diabetes, and infrequent social contact. Those results support previous research [44].

The correlation between SHAP values, permutation-based feature importance, natural feature importances (XGBoost, RF), and LR log odds ratios was examined in [45] using classical ML methods. The dataset included MRI volumes describing the hemispheric asymmetry, sociodemographic features, the number of ApolipoproteinEϵ4 (ApoEϵ4) alleles, and cognitive test scores. All models were trained for an ADNI training set and validated for an ADNI test set, the external AIBL, and OASIS datasets. The datasets included 1,700 ADNI- (512 CN, 853 MCI, 335 AD), 612 AIBL- (446 CN, 95 MCI, 71 AD), and 921 OASIS-subjects (704 AD, 19 MCI, 198 CN). As permutation-based explanation methods divide feature importances between correlated features [28] and thus reduce the importance of individual features, correlated features are consolidated before model interpretation using aspects [29]. The results showed a strong correlation between the SHAP values of different models.

Deep-learning models were often explained using heatmaps. GradCAM was, e.g., used in [46] to explain Long Short-Term Memory-(LSTM-) [47] based Recurrent Neural Networks [48]. The experiments

compared techniques to augment MRIs with sociodemographic and genetic data. Those models were trained for the CN vs. MCI task using the AD subset [49] of the Heinz Nixdorf Risk Factors Evaluation of Coronary Calcification and Lifestyle (RECALL) (HNR) [50] (61 MCI and 59 CN) and ADNI-1 [30] (397 MCI, 227 CN). The heatmaps focused on biologically plausible regions.

The connection between summed relevance scores of heatmaps in the hippocampus area and hippocampal volumes was investigated in [51] using Layer-wise Relevance Propagation (LRP) [52]. The models were trained using 3D MRIs of 663 ADNI-GO/2 subjects (254 CN, 220 MCI, 189 AD) and validated using subjects from ADNI-3 (326 CN, 187 MCI, 62 AD), AIBL (448 CN, 96 MCI, 62 AD) and German Center for Neurodegenerative Diseases (DZNE) multicenter observational study on Longitudinal Cognitive Impairment and Dementia (DELCODE) [53] (215 CN, 155 MCI, 104 AD). The experiments found, that the hippocampus area is the most relevant AD brain structure. The summed relevance scores of the hippocampus correlated with the hippocampus volumes.

The results of three heatmap methods, namely LRP [52], Integrated Gradients [54], and Guided GradCAM [19] were compared to the results of a VBM meta-analysis in [55]. The DL models were trained to classify 252 CN and 250 AD ADNI subjects using 3D MRIs. For each explanation method, the mean heatmap intensity of all subjects was calculated and the accordance with the ground truth was calculated. The Dice scores showed a moderate correlation between the heatmaps and the ground truth. A comparison to SVM coefficients revealed that the LRP heatmaps Dice-Scores outperformed the SVM coefficients.

A comparison between a 3D-DL model trained on MRI scans, and a gradient boosting classifier, trained on tabular volume and cortical thickness data, automatically extracted from MRI scans was performed in [56]. All models were trained on 2619 MRI scans (782 CN, 1089 MCI, 748 AD) from 682 subjects of the ADNI dataset. External validation was performed on a subset of 2045 subjects from the National Alzheimer's Coordinating Center (NACC) study. Relevant brain regions for the DL model were extracted using the Saliency heatmap explainability method. The impurity-based feature importance scores were implemented to rank the features of the gradient boosting classifier. The results of the experiments showed improved classification performance of the DL model compared to the gradient boosting classifier during the internal and external validation. In addition, the DL model was found to focus on a higher number of brain regions, some of which have not been previously associated with AD. The gradient boosting classifier focused mainly on the hippocampus region. However, the results of the study are only calculated for one DL and ML model, which makes a more systematic comparison of multiple ML models, DL models and explanatory methods important.

Recently, no previous research has done a systematical, extensive, and quantitative comparison between 3D DL heatmaps and tabular data taking into account correlation structures and model calibration. To investigate this question, DL-based activations based on GradCAM, GradCAM++, SHAP, and LIME of pre-segmented cortical and subcortical brain structures are summarized in this research. Model calibration [26] and feature correlation [57], which affect the explainability were prevented using Platt scaling [27] and aspect consolidation [29]. Most DL models trained in previous research suffered from information leakage [3]. In this work, special attention was given to the splitting of training and test sets, which were split at the subject level.

3. Materials and methods

This section describes the datasets used to train the ML and DL models, validate the results and perform the explanations, as well as introduces explainability methods and VBM used in this work.

Table 1
Demographic data, and MRI field strength of the selected ADNI subjects, separated by diagnosis groups. For continuous features, mean and standard deviation are given.

Diagnosis	n	Age (years)	Females (%)	1.5 T (%)	3 T (%)
ADNI					
CN	512	74.20 ± 5.82	51.76	44.00	56.00
AD	335	74.95 ± 7.74	44.78	57.00	43.00
Σ	847	74.50 ± 6.66	49.00	49.00	51.00
AIBL					
CN	446	72.53 ± 6.14	56.95	19.06	80.94
AD	71	73.26 ± 7.88	59.15	16.90	83.10
Σ	517	72.63 ± 6.41	57.25	18.76	81.24
OASIS					
CN	704	68.35 ± 9.27	58.66	12.36	87.64
AD	198	75.62 ± 7.92	48.48	10.61	89.39
Σ	902	69.94 ± 9.48	56.43	11.97	88.03

3.1. Datasets

Data used in the preparation of this article were obtained from ADNI [30], AIBL [31], and OASIS [32]. ADNI[4] [30] was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal is to test whether a combination of biomarkers can measure the progression of MCI and AD. These biomarkers include MRI, PET, biological markers, as well as clinical and neuropsychological assessments. The ongoing cohort recruited subjects from more than 60 sites in the United States and Canada and consisted of four phases. The dataset was downloaded on 2020-07-27 and initially included 2250 subjects.

AIBL[5] [31], launched in 2006, is the largest AD study in Australia. AIBL aims to discover biomarkers, cognitive tests, and lifestyle factors. As AIBL focuses on early AD stages, most subjects are CN. AIBL data v3.3.0 was downloaded on 2019-09-19 and originally included 858 subjects.

The aim of the Open Access Series of Imaging Studies (OASIS) 3,[6] [32] is, to investigate the effects of healthy aging and AD. OASIS-3 subjects were recruited from several ongoing studies in the Washington University Knight AD Research Center (KnightADRC).[7] The longitudinal dataset included MRI, fMRI, Amyloid-, and FDG-PET scans, neuropsychological tests, and clinical data for 1098 subjects.

The subject selection process was previously described [45] in more detail. The demographics and MRI field strengths of the selected subjects and scans are summarized in Table 1. The selection of the MRI scans is described in Section 4.1.

3.2. Explainability methods

Explainability methods [8] were divided into model-specific and model-agnostic methods. Another distinction is made between global methods explaining the overall feature relevance, and local methods interpreting individual predictions.

3.2.1. Permutation-based feature importance

Permutation-based feature importance [5] is a global, model-agnostic method. The relevance of a feature is computed by first calculating the model's performance for the original feature set, and afterwards recomputing the accuracy for a dataset with permuted values for a single feature. The feature importance is defined as the mean difference between the baseline performance and the performance of the permuted dataset in multiple repetitions.

[4] ADNI: https://adni.loni.usc.edu, Access: 2024-01-23.
[5] AIBL: https://aibl.csiro.au/, Access: 2024-01-23.
[6] OASIS 3: https://www.oasis-brains.org/ Access: 2024-01-23.
[7] KnightADRC: https://knightadrc.wustl.edu/, Access: 2024-01-23.

3.2.2. LIME

LIME [18] is a local explanation method that explains the black-box model predictions f of individual observations x by training local surrogate models $g^*(x)$. To explain an individual model prediction, LIME generates a new dataset containing data near the observation at interest based on perturbations. For this dataset, an explainable regression model is fit as a local surrogate. The dataset observations are weighted based on their proximity (π_x) to the original observation. In the local optimization function (Eq. (1)), $L(f, g, \pi_x)$ describes the loss between the black-box model f and the local explanation model g.

$$g^*(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{1}$$

The explanation model complexity is regularized by $\Omega(g)$. The original LIME algorithm works with a linear loss function (Eq. (2)), K-Least Absolute Shrinkage and Selection Operator- (K-LASSO) [18] for feature regularization, and an exponential kernel as the proximity function. The resulting LIME values depend on the feature relevance in the explanation model.

$$L(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z)(f(z) - g(z'))^2 \tag{2}$$

3.2.3. SHAP

SHAP [17] is a model-agnostic method based on Shapley values [58] and aims to explain the individual prediction of an observation $V(D)$ by feature expressions $D = \{1, \ldots, n\}$. $V(D)$ is the prediction probability of a subject belonging to a predefined class. The summed SHAP values (Eq. (3)) are equal to the difference between the individual prediction $V(D)$ and the average model prediction Φ_0.

$$V(D) = \Phi_0 + \sum_{i=1}^{n} \Phi_i \tag{3}$$

The exact calculation of SHAP values (Eq. (4)) requires the model retraining for each subset S of features, leading to an exponential increase in the computational effort. Kernel SHAP [17] is one possibility for time-efficient estimation.

$$\Phi_i = \sum_{S \subseteq D \setminus \{i\}} \frac{V\left(S \cup \{i\}\right) - V(S)}{\binom{n-1}{|S|}} \tag{4}$$

Kernel SHAP uses LIME to fit an additive linear model (Eq. (5)) with a simplified representation (x') of the input features and the explanation model $g(x')$. The weights Φ_i of $g(x')$ estimate the SHAP values. For tabular data, the simplified features are binned, binary feature representations of the presence or absence of an expression, for image data, superpixels are used. M is the number of simplified features.

$$g(x') = \Phi_0 + \sum_{i=1}^{M} \Phi_i \cdot x'_i \tag{5}$$

The LIME parameters [17] used for SHAP estimation are described in Eqs. (6) and (7). $\Omega(g)$ is set to zero and $h_x(x') = x$ maps the simplified features to the original feature space. In this work, global SHAP relevance was systematically computed as the sum of all absolute local SHAP values.

$$\pi_x(x') = \frac{M - 1}{\binom{M}{|x'|} \cdot |x'| \cdot (M - |x'|)} \tag{6}$$

$$L(f, g, \pi_{x'}) = \sum_{x' \in X} \left(f\left(h_x(x')\right) - g(x')\right)^2 \cdot \pi_{x'}(x') \tag{7}$$

3.2.4. GradCAM/GradCAM++

GradCAM [59] is a model-specific method to locally explain black-box CNN predictions based on feature activation maps and model gradients. The result of GradCAM is a heatmap $L^c_{GradCAM}$ which highlights the most relevant regions in an image for the prediction of class c. For each input image, the gradients were computed for class c and concerning feature map activations A^k. The activations have the dimensions $d \times h$ of the last convolutional network layer k. Global average pooling was applied to the gradients of each feature map to calculate their relevance α^c_k (Eq. (8)).

$$w^c_k = \frac{1}{d \cdot h} \sum_{i=1}^{d} \sum_{j=1}^{h} \frac{\partial y^c}{\partial A^k_{ij}} \tag{8}$$

The relevance scores determine the influence of each feature map. Rectified Linear Units (ReLU) [60] were applied as the activation function (Eq. (9)) to calculate the GradCAM heatmap.

$$L^c_{GradCAM} = ReLU(\sum_k \alpha^c_k A^k) \tag{9}$$

GradCAM++ [20] is an advanced version of GradCAM. It compensates for the problem of recognizing multiple occurrences of one class in an image and improves heatmap localization. GradCAM++ is a generalization of GradCAM where α is the importance of location i, j in the activation map A^k (Eq. (10)), which depends on higher-order derivatives.

$$w^c_k = \sum_{i=1}^{d} \sum_{j=1}^{h} \alpha^{kc}_{ij} \cdot ReLU(\frac{\partial y^c}{\partial A^k_{ij}}) \tag{10}$$

3.3. Voxel-based morphometry (VBM)

Voxel-Based Morphometry (VBM) [61] is a method to do a comparison of voxel-wise gray-matter (GM), white matter (WM), or CSF concentration between groups of subjects. First, brain scans of all subjects are spatially normalized to a reference brain so that they are all in the same stereotactic space. Second, the GM, WM, and CSF concentration of each voxel is computed considering the position and the voxel intensity. The resulting 3D scans are smoothed. In the end, statistical tests compare the groups in a voxel-wise manner. To avoid bias regarding multiple statistical testing, a correction is performed.

4. Approach

The following section describes the workflow developed for early AD detection shown in Fig. 1. The MRI pre-processing was implemented using FreeSurfer v6.0 [62] and the programming language Python v3.6.9 [63] was used for hyperparameter tuning and model training. The workflow implementation and the IDs of the scans are available online.[8]

4.1. MRI selection, pre-processing, and feature extraction

To directly compare the relevance of brain regions between DL and classical ML models, contrary to previous work [45] all experiments are based exclusively on data extracted from MRIs. For each subject, one baseline T1-weighted MRI was selected. As previously explained in [45], the baseline scans which were included in the adnimerge dataset were selected. The acquisition parameters differed between studies and scanners. During the ADNI-1 study phase, scans were recorded using a field strength of 1.5 T. In the remaining ADNI study phases, MRIs with a field strength of 3.0 T were recorded (Table 1). Despite differences in the images of both field strengths, which might introduce some bias, it was decided to include scans of both field strengths to increase the size of the data set. The effect of the field strengths on the classification performance is investigated in later experiments which are described in Section 5.2. The MRIs of the AIBL dataset followed the protocol of the ADNI 3D T1-weighted sequences. All AIBL scans had a resolution of $1 \times 1 \times 1.2$ mm. The OASIS-3 dataset included T1-weighted MRIs, recorded on three scanners [32].

[8] GitHub Repository: https://github.com/LouiseBloch/AlzheimerSystematicXAIComparison.

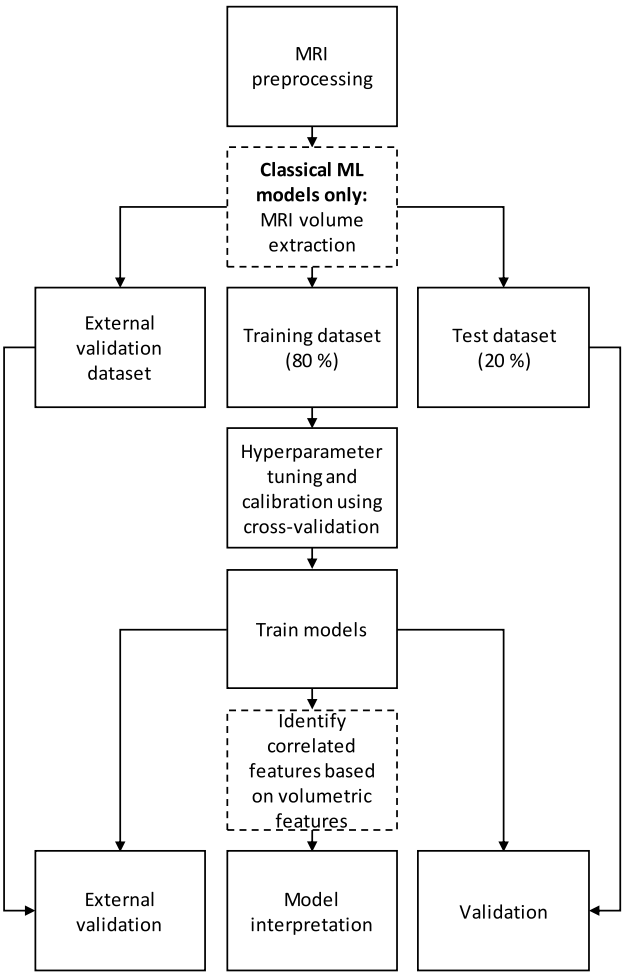


Fig. 1. ML-Workflow including MRI pre-processing and feature extraction. The dataset was split into an 80% training set, a hold-out 20% test set, and the external validation data. Hyperparameter tuning included CV, which was reused for model calibration. Model training included 3D DL and classical ML. Correlated features were consolidated into aspects. Model interpretation includes SHAP, LIME, permutation-based importance, GradCAM, and GradCAM++.

FreeSurfer v6.0 [62] was used for MRI pre-processing, segmentation of cortical and subcortical brain structures, and volume extraction. All scans were segmented into 113 regions including 30 cortical areas per hemisphere of the Desikan–Killiany–Tourville (DKT) atlas [64], 41 subcortical areas [65], five areas of the corpus callosum, the left and right cerebral cortex, left and right unknown and undetermined tissue, as well as the background (unknown). Using this segmentation, for all structures except for the left and right cerebral cortex, left and right unknown and undetermined tissue, as well as the background (unknown), volumes were extracted using FreeSurfer. All volumes were normalized for the estimated Total Intracranial Volume (eTIV) [66]. The resulting volumetric dataset includes 107 features (106 brain regions as well as the eTIV) which were used to train all classical ML models. A list of these features can be found in Suppl. Mat. A. During the training of these classical ML models, all features were centered and scaled. The pre-processing parameters were calculated for the training set and applied to the training, test, and external validation sets.

For DL, the MRIs were affinely registered to the MNI305 atlas [67], the intensities were normalized, and skull stripping was performed. The FreeSurfer-based pre-processing resulted in images of size 256 × 256 × 256 px. During model training, the intensities were scaled, and a random spatial crop of size 224 × 224 × 224 px was extracted. During model validation, the same intensity scaling, but a center-spatial cropping was performed. The FreeSurfer-based segmentations were used to group the results of the explainability methods and thus analyze the importance of specific brain structures in the DL models.

Table 2
Hyperparameters and intervals used for hyperparameter tuning.

Model	Hyperparameter	Values
DT	`criterion`	{"gini", "entropy"}
	`splitter`	{"best", "random"}
	`max_depth`	{2, 4, … , 20, None}
	`max_features`	{"sqrt", "log2", None}
XGBoost	`n_estimators`	{100, 200, … , 500}
	`eta`	{0.1, 0.2, … , 1.0}
	`gamma`	{0, 2, … , 20}
	`max_depth`	{5, 10, … 20}
	`subsample`	{0.5, 0.6, … , 0.9}
	`colsample_bytree`	{0.5, 0.6, … , 0.9}
RF	`n_estimators`	{100, 200, … , 1000}
	`criterion`	{"gini", "entropy"}
	`max_features`	{"sqrt", "log2"}
LightGBM	`n_estimators`	{100, 200, … , 500}
	`learning_rate`	{0.1, 0.2, … , 1.0}
	`colsample_bytree`	{0.5, 0.6..., 0.9}
	`max_depth`	{5, 10, … , 20}
	`subsample`	{0.5, 0.6, … , 0.9}
	`num_leaves`	{10, 20, … , 50}
SVM poly	`C`	$\{10^{-5}, 10^{-4}, \dots, 10^{5}\}$
	`degree`	{1, 2, … , 10}
	`gamma`	{"scale", "auto"}
SVM rbf	`C`	$\{10^{-5}, 10^{-4}, \dots, 10^{5}\}$
	`gamma`	{"scale", "auto"}
DenseNet	`lr`	$\{10^{-5}, 10^{-4}, \dots, 10^{-1}\}$
	`optimizer`	{"sgd", "adam", "rmsprop"}
	`scheduler`	{"step", "exp", None}
	`epochs`	{1, 2, … , 50}
All DL models	`lr`	$\{10^{-5}, 10^{-4}, \dots, 10^{-1}\}$
	`optimizer`	{"sgd", "adam", "rmsprop"}
	`scheduler`	{"step", "exp", None}
	`epochs`	{1, 2, … , 50}

4.2. Hyperparameter tuning and model calibration

All models were trained using the ADNI training set (80%) and validated using the hold-out ADNI test set (20%). The splitting was performed at the subject level and was executed within each diagnostic group to ensure similar class distributions. The AIBL and OASIS datasets were used for external validation. None of the validation subjects was used during training or model selection.

A grid search hyperparameter tuning was implemented to find the best hyperparameters. Stratified 5-fold cross-validation (CV) [68] was implemented using Scikit-learn [69] v0.23.2 and applied to the training part of the ADNI dataset to estimate the performance for an independent validation set. First, the training set was split into five distinct folds using stratification on the diagnostic level. Using these folds, five iterations were performed, each with a different fold as the hold-out validation set (20%). The training set included the remaining folds (80%). The best hyperparameters calculated during the CV were used to train the final model on the entire training set. The tuning intervals are summarized in Table 2.

The CV-predictions generated during parameter tuning were reused for model calibration. Platt scaling [27] was implemented using the Python library Scikit-learn [69] v0.23.2 to reduce uncertainty which also affects the explainability [26].

4.3. Model training

Model training was performed for two feature sets. The classical ML models (XGBoost, RF, LightGBM, radial SVM, polynomial SVM,

DT, LR) were trained using volumetric features extracted from MRIs. The XGBoost model was trained using the Python package XGBoost v1.2.0 [6] and the LightGBM was trained using the Python package lightgbm v3.3.2 [24]. The remaining classical ML models were implemented using Scikit-learn v0.23.2 [69]. End-to-end CNNs were trained on pre-processed 3D MRIs. The model architectures were adapted from DenseNets, EfficientNets, and two Squeeze-and-Excitation-Networks (SENets) which are based on the ResNet [38] and ResNeXt [70] models. All models were loaded using the Python library MONAI v0.8.0 [71] and trained using Pytorch v1.7.1+cu110 [72]. The exact model names in MONAI are `densenet121`, `EfficientNetBN ('efficientnet-b0')`, `SEResNet152`, and `SEResNext101`. A visualization of the DL architectures can be found in Suppl. Mat. B - E. During training, batch accumulation increased the mini-batch size from 2 to 64. The cross-entropy loss function was used. All experiments were performed on an NVIDIA® DGX-1,[9] supercomputer, with NVIDIA® V100[10] tensor core Graphical Processing Units (GPUs) containing 16 GB of memory. The execution environment was an NVIDIA®-optimized[11] Docker[12] [73] container, running a Deepo[13] image. All experiments were executed using a single GPU.

4.4. Transfer learning for DL models

Due to the high number of parameters in DL models, those need large datasets to generalize well. Those large datasets are not available for AD detection. One approach to overcome this problem is transfer learning [74]. The idea of transfer learning is to transfer knowledge learned from one domain typically with a larger dataset to another related field. In this paper, transfer learning was used as one possibility to increase the performance of the 3D-DL models.

The LDM-100k dataset [75] which includes 100,000 synthetically generated T1-weighted brain MRI scans was used for transfer learning due to the assumption that these images are closely connected to the AD dataset used within this paper. The dataset was split into a 60% training set, a validation dataset covering 20%, and an independent test set including the remaining 20% of the test set. Models were pretrained on the unprocessed scans of the training dataset to predict the normalized age connected to the synthetic brain scans. A hyperparameter tuning for the learning rate was performed on the validation set. For all models, Mean Square Error (MSE) was used as the loss function. Additional, mixed precision, an accumulated batch size of 120, no learning rate scheduling, as well as the Adam optimizer was used to train the models for 50 epochs. The base DL architectures, as well as the augmentation pipeline, are the same which are used for AD detection. The final results are reported on the independent test set.

The resulting models were used for fine-tuning using the ADNI dataset as is described in the remaining parts of this section.

4.5. Identify correlated features

The AD volumetric dataset contains correlated features affecting the comparison of explainability methods regarding the biological plausibility [28,57]. As permutation-based interpretation methods assume independent input variables, correlated features lead to synthetic subjects with unrealistic combinations of features. This effect can lead to reduced relevance scores for correlated features, compared to the resulting feature importance of independent features. For this reason, correlated features are consolidated before model interpretation into aspects [29]. The feature correlation was calculated using Spearman rank correlation. Hierarchical agglomerative clustering [57] with a threshold of $H = 0.5$ was used to create a dendrogram and extract the resulting aspects. The dendrogram was computed for the volumetric training set and applied to the FreeSurfer segmentations of the 3D MRIs of the deep-learning pipeline. The features in an aspect were consolidated during the LIME, SHAP, and permutation-based model interpretation with the Python library dalex v1.4.1 [76].

4.6. Evaluation

Evaluation was performed for the ADNI test set, and the external AIBL, and OASIS datasets. The model performances were measured using accuracy (ACC), balanced accuracy (BACC), F_1-Score (F_1), Matthews correlation coefficient (MCC), and Area under the Receiver Operating Curve (AUROC). Comparability to related research was increased using multiple metrics for evaluation. In comparison to accuracy, the F_1-Score focuses on incorrectly classified cases. The macro averaging F_1-Score was calculated to address both, diseased and healthy subjects. Balanced accuracy is based on sensitivity and specificity and thus suits to evaluate imbalanced class problems. The MCC and AUROC also suit imbalanced datasets.

The AUROC models the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) for different confidence thresholds and is thus independent of model calibration. For better readability, all metrics except for the MCC, which is a correlation coefficient with a range of $(-1, 1)$, are given as percentage values.

4.7. Model interpretation

For the black-box model interpretation, the local and global relevance of predefined brain structures was computed using multiple methods. The model-specific GradCAM and GradCAM++ methods were implemented using the Python library MONAI v0.8.0 [71]. The MONAI standard implementation included post-processing, which linearly scales the intensities to a range of $[1, 0]$ and thus flips the magnitudes.[14] For this reason, post-processing inverted the normalized GradCAM and GradCAM++ scores. For DL models, LIME was implemented using the Python library LIME v0.2.0.1 [18] and the Python library SHAP v0.38.1[15] was used to implement Kernel SHAP. For both methods, Simple Linear Iterative Clustering (SLIC) [77] was used to generate 100 similar-sized superpixels per scan. The compactness was set to 1. Disabled superpixels were replaced by the respective pixels of an image containing the mean intensity of all training images in each segmented structure. For the classical ML models, LIME, SHAP, and permutation-based model interpretation were implemented using the Python library dalex v1.4.1 [76].

5. Results

Internal and external classification performances of the experiments, volumetric feature correlation patterns, local and global explanations, as well as their correlations are described in this section.

5.1. Hyperparameter tuning

The CV-results achieved during the hyperparameter tuning are summarized in Table 3. The best accuracy was 89.24%±2.85 achieved for the

[9] DGX-1: https://www.nvidia.com/de-de/data-center/dgx-1/ Access: 2024-01-23.

[10] V100: https://www.nvidia.com/en-us/data-center/v100/, Access: 2024-01-23.

[11] NVIDIA®-Docker: https://github.com/NVIDIA/nvidia-docker, Access: 2024-01-23.

[12] Docker: https://www.docker.com/, Access: 2024-01-23.

[13] Deepo: https://github.com/ufoym/deepo, Access: 2024-01-23.

[14] `monai.visualize.class_activation_maps.default_normalizer`: https://docs.monai.io/en/stable/visualize.html, Access: 2024-01-23.

[15] SHAP: https://github.com/slundberg/shap, Access: 2024-01-23.

Table 3
Hyperparameters and CV-accuracies achieved during parameter tuning. The abbreviation TL marks the Deep-Learning models for which transfer learning was performed on the LDM-100k dataset. The best results are highlighted in bold. Parameters: DT: {criterion, splitter, max_depth, max_features }; XGBoost: {n_estimators, eta, gamma, max_depth, subsample, colsample_bytree}; RF: {n_estimators, criterion max_features}; LightGBM: {n_estimators, learning_rate, colsample_bytree, max_depth, subsample, num_leaves}; SVM poly: {C, degree, gamma}; SVM rbf: {C, gamma}; Deep learning models: {lr, optimizer, scheduler, epochs}.

Model	Hyperparameters	CV-Accuracy ($\bar{x} \pm \sigma$)
LR	–	84.81% ± 2.74
DT	{"gini", "best", 6.0, "sqrt"}	79.20% ± 2.16
XGBoost	{400, 0.1, 0, 5, 0.6, 0.7}	89.09% ± 2.84
RF	{1000, "gini", "log2"}	87.61% ± 3.49
LightGBM	{500, 1.0, 0.6, 5, 0.8, 30}	89.09% ± 4.60
SVM poly	{1.0, 1, "scale"}	**89.24%** ± **2.85**
SVM rbf	{1.0, "auto"}	88.65% ± 2.75
DenseNet	{10^{-4}, "none", "adam", 29}	87.02% ± 0.95
DenseNet TL	{10^{-2}, "none", "adam", 39}	87.17% ± 1.82
EfficientNet	{10^{-3}, "none", "adam", 89}	86.44% ± 3.04
EfficientNet TL	{10^{-3}, "none", "adam", 87}	88.94% ± 3.18
SEResNet	{10^{-4}, "none", "adam", 77}	88.05% ± 1.82
SEResNeXt	{10^{-3}, "none", "adam", 67}	87.46% ± 1.46

polynomial SVM. The DT reached the worst results of 79.20%±2.16. The DenseNet model trained from scratch on the original 3D MRIs reached a CV-accuracy of 87.02%±0.95. The pretrained DenseNet model achieved a slightly better performance of 87.17% ± 1.82. For the EfficientNet, SEResNet, and SEResNeXt models, the course of the CV-accuracy after 50 epochs has not converged. For this reason, it was decided, to train the best-performing models for another 50 epochs. The EfficientNet model, which was trained from scratch reached a CV-accuracy of 86.44% ± 3.04, which is slightly worse than the performance achieved for the DenseNet models. The pretrained EfficientNet outperforms this model with a CV-accuracy of 88.94% ± 3.18. The SEResNet model achieved a CV-accuracy of 88.05% ± 1.82 and the SEResNeXt model reached a CV-accuracy of 87.46% ± 1.46. Overall, ML and DL models achieved similar CV-accuracies. Except for the DT which reached a CV-accuracy of 79.20% ± 2.16 the performances of the remaining models was between 84% and 90%.

5.2. ADNI test set validation

The results achieved for the ADNI test set are summarized in Table 4. As the number of epochs of the DL models depended on the number of observations in the training set and the final model was trained on a larger dataset than the CV-models, this number was increased by 10%, leading to $29 + 29 \cdot 0.1 \approx 32$ epochs for the DenseNet which was trained from scratch, $39 + 39 \cdot 0.1 \approx 43$ epochs for the pretrained DenseNet, $89+89 \cdot 0.1 \approx 98$ epochs for the EfficientNet trained from scratch, $87 + 87 \cdot 0.1 \approx 96$ epochs for the pretrained EfficientNet, $77 + 77 \cdot 0.1 \approx 85$ epochs for the SEResNet, and $67 + 67 \cdot 0.1 \approx 74$ for the SEResNeXt models. Additionally, to obtain more robust models, Polyak averaging was implemented. Therefore, the model parameters from the last five epochs were averaged to create the final model.

As previously mentioned, the LDM-100k dataset was used to train two models, namely DenseNet and EfficientNet with transfer learning. The broad hyperparameter tuning of the learning rate for the DenseNet model leads to a learning rate of 10^{-2}. This model achieved an MSE of 0.080 on the validation dataset. For the EfficientNet model, a learning rate of 10^{-5} was selected leading to an MSE of 0.145.

Consistently with the CV, the best AUROC of 96.21% was reached for the polynomial SVM. The XGBoost, and the LightGBM which both achieved the second-best CV-accuracy, achieved the second-best accuracy (91.12%) and MCC (0.814). The LightGBM also reached the second-highest balanced accuracy of 90.34% and F_1-Score of 90.65%.

Table 4
Results achieved for the hold-out ADNI test set.

Model	ACC	BACC	AUROC	F_1	MCC
No information rate: 60.36%					
LR	90.53%	90.11%	94.85%	90.11%	0.802
DT	81.07%	77.91%	80.00%	79.00%	0.602
XGBoost	91.12%	90.09%	96.08%	90.60%	0.814
RF	88.17%	87.38%	96.06%	87.57%	0.752
LightGBM	91.12%	90.34%	94.15%	90.65%	0.814
SVM poly	88.75%	86.59%	**96.21%**	87.77%	0.768
SVM rbf	88.17%	87.38%	95.49%	87.57%	0.752
DenseNet	88.17%	87.38%	89.08%	87.57%	0.752
DenseNet TL	86.39%	85.91%	93.43%	85.82%	0.716
EfficientNet	**91.72%**	**90.58%**	92.80%	**91.20%**	**0.827**
EfficientNet TL	85.21%	83.90%	91.98%	84.33%	0.688
SEResNet	89.94%	89.36%	92.07%	89.46%	0.789
SEResNeXt	85.21%	84.42%	93.50%	84.50%	0.690

The results reached for the LR are similar to the black-box model performances (Accuracy: 90.53%, Balanced Accuracy: 90.11%, AUROC: 94.85%, F_1-Score: 90.11%, MCC: 0.802). The DT reached the worst results (Accuracy: 81.07%, Balanced Accuracy: 77.91%, AUROC: 80.00%, F_1-Score: 79.00%, MCC: 0.602).

The best accuracy (91.72%), balanced accuracy (90.58%), F_1-Score (91.20%), and MCC (0.827) were reached for the EfficientNet which was trained from scratch. The EfficientNet pretrained on the LDM-100k dataset, achieved the lowest scores for all metrics except the AUROC (Accuracy: 85.21%, Balanced Accuracy: 83.90%, AUROC: 91.98%, F_1-Score: 84.33%, MCC: 0.688) when considering only the DL models. The DenseNet trained from scratch reached an accuracy of 88.17%, a balanced accuracy of 87.38%, an AUROC of 89.08%, an F_1-Score of 87.57%, and an MCC of 0.752. The SEResNet model achieved an accuracy of 89.94%, a balanced accuracy of 89.36%, an AUROC of 92.07%, an F_1-Score of 89.46%, and a MCC of 0.789. Overall, both the ML and DL models achieved similar performances, but the EfficientNet trained from scratch attained the best results. It is noteworthy that, with the exception of the DT, all classic ML models demonstrate superior AUROC scores compared to the DL models. The remaining metrics do not exhibit such a trend.

The pretrained models did not outperform the models trained from scratch. There are different possibilities to potentially improve the performances during model fine-tuning, these include the use of a real-world dataset instead of the artificially generated LDM-100k dataset for pre-training, using a preprocessed version of the LDM-100k dataset or starting the fine-tuning for some epochs with frozen parameters. These ideas should be addressed in more detail in future work.

It was observed that eight out of the 13 models achieved better accuracies on the ADNI test set than during CV. For four of these models, the improvement is smaller than one standard deviation. The improvement for the LR is 2.09 times the standard derivation, for the DenseNet which was trained from scratch 1.21 times the standard deviation, for the EfficientNet which was trained from scratch 1.74 times the standard deviation, and for the SEResNet 1.04 times the standard deviation. Reasons for this might be the relatively small number of $n = 169$ samples in the test dataset which is caused by the overall size of the dataset as well as that the final models were trained on a slightly larger dataset (combination of training and validation dataset). It was thoroughly tested that none of the subjects in the test dataset were previously used during training or hyperparameter tuning of the model.

The summary of the dataset in Table 1 shows that it contains MRI scans with two different field strengths, namely 1.5 T and 3 T. The field strength is a covariate that has an influence on the level of detail of the MRI scans but also image artifacts. Thus it may influence the model performance. For this reason, the model performances of both groups are shown in Table 5. The ADNI test set includes 77 scans (45.56%) acquired at 1.5 T, and 92 subjects (54.44%) acquired at 3 T. The no

information rate for the 1.5 T group is 55.84% and the no information rate for the 3 T group is 64.13%.

The results show that for most of the ML models, the metrics show higher performances for the 3 T group in comparison to the 1.5 T group. Exceptions to this are the DT which achieves better results for all metrics in the 1.5 T group. It has to be mentioned, that the DT achieved the worst results across all models. Additionally, the AUROC of the LR, and the XGBoost, the F_1-Score of the LightGBM model, as well as the balanced accuracy and AUROC, of the polynomial SVM performed better on the 1.5 T test set. In summary, this means that three of the seven ML models reached better results for all metrics in the 3 T group, one model performed better in the 1.5 T group for all metrics, two models performed better in the 3 T group for four out of five metrics, and one model performed better for the 3 T group for three out of five metrics.

The DL models diverge from the trend of achieving better results on the 3 T dataset and show a more complex pattern. Both DenseNet models, as well as the EfficientNet model which was trained from scratch, performed better on the 1.5 T scans in comparison to the 3 T scans for all metrics. For the pretrained EfficientNet model, all metrics perform better on the 3T scans. The SEResNet achieved higher results for the 3T scans for all metrics except for the AUROC. The SEResNeXt model performed better on the 3T scans using the accuracy, balanced accuracy, and MCC as metrics. The AUROC and F_1-Score of the same model are better within the 1.5 T group. In summary, three of the six DL models performed better on the 1.5 T scans for all metrics, one model performed better on the 3 T scans for all metrics, one model reached better results in the 3 T scans for four out of five metrics, and the last model achieved better results in the 3 T group for three out of five metrics

The overall results show that the performance for both groups is reasonable and clearly outperforms the no information rate. Regarding the classical ML models, most models achieve slightly better results in the 3 T scan group. Exceptions from this were the DT which prefers the 1.5 T scans and the polynomial SVM which does not indicate clear preference. No clear preference was observed for the DL models, as three out of six models show better performance in the 1.5 T group. Two models perform better in the 3T group for most metrics. The SEResNeXt model does not indicate any clear preference.

5.3. External validation

The external AIBL and OASIS results are summarized in Table 6. For AIBL, the no information rate was 86.27%. It should be noted that the no information rate is higher in both external validation datasets compared to the ADNI test dataset. The imbalanced diagnostic groups in the external datasets result in higher accuracy scores during external validation. Therefore, comparing accuracy scores between imbalanced datasets is misleading and comparisons should focus on metrics that are less impacted by class imbalances, such as balanced accuracy. All models except the DT (accuracy: 85.69%) and the pretrained DenseNet model (accuracy: 84.91%) outperformed the AIBL no information rate. The best accuracy of 91.30% was reached for the radial SVM. The same model also achieved the best balanced accuracy (89.63%), F_1-Score (84.09%), and MCC (0.697). Additionally, the model reached the third-highest AUROC score of 92.37%. The best AUROC of 93.98% was reached for the XGBoost. Within the DL models, the best results on the AIBL dataset were reached for the SEResNeXt model (Accuracy: 90.91%, Balanced accuracy: 85.26%, AUROC: 90.80%, F_1-score: 82.35%, MCC: 0.652). The EfficientNet model which was pretrained on the LDM-100k dataset reached the best AUROC score of 91.74% within the DL models.

For OASIS, the no information rate was 78.05% and all models outperformed this value. The best accuracy is 86.47% achieved by the EfficientNet model which was pretrained on the LDM-100k dataset. This model also reached the best F_1-Score of 80.41% and MCC of 0.608.

Table 5
Results achieved for the hold-out ADNI test set split between field strengths 1.5 T and 3 T.

Model	ACC	BACC	AUROC	F_1	MCC
1.5 T: No information rate: 55.84%					
LR	88.31%	88.30%	96.17%	86.96%	0.764
DT	81.82%	80.95%	85.47%	78.13%	0.630
XGBoost	89.61%	89.47%	96.03%	88.24%	0.789
RF	84.42%	84.82%	95.14%	83.33%	0.692
LightGBM	89.61%	89.77%	93.43%	88.57%	0.792
SVM poly	88.31%	87.38%	96.51%	85.71%	0.766
SVM rbf	85.71%	85.98%	95.28%	84.51%	0.715
DenseNet	88.31%	88.00%	93.09%	86.57%	0.763
DenseNet TL	90.91%	90.94%	95.90%	89.86%	0.817
EfficientNet	92.21%	92.10%	94.46%	91.18%	0.842
EfficientNet TL	80.52%	80.10%	90.01%	77.61%	0.604
SEResNet	87.01%	87.14%	95.69%	85.71%	0.739
SEResNeXt	83.12%	83.34%	93.43%	81.69%	0.663
3 T: No information rate: 64.13%					
LR	92.39%	91.40%	93.58%	89.23%	0.834
DT	80.43%	74.06%	75.45%	65.38%	0.570
XGBoost	92.39%	90.06%	95.84%	88.52%	0.835
RF	91.30%	88.55%	97.33%	86.67%	0.812
LightGBM	92.39%	90.06%	95.02%	88.52%	0.835
SVM poly	89.13%	85.52%	95.35%	82.76%	0.766
SVM rbf	90.22%	87.70%	95.38%	85.25%	0.786
DenseNet	84.78%	82.13%	85.00%	77.42%	0.663
DenseNet TL	84.78%	82.79%	91.53%	78.12%	0.666
EfficientNet	91.30%	88.55%	91.37%	86.66%	0.812
EfficientNet TL	89.13%	86.85%	92.76%	83.87%	0.761
SEResNet	92.39%	90.73%	88.08%	88.89%	0.833
SEResNeXt	86.96%	84.49%	92.96%	80.65%	0.712

Table 6
External test results achieved for AIBL and OASIS.

Model	ACC	BACC	AUROC	F_1	MCC
AIBL: No information rate: 86.27%					
LR	87.62%	87.50%	91.93%	79.20%	0.617
DT	85.69%	73.94%	80.13%	72.07%	0.445
XGBoost	89.94%	87.66%	**93.98%**	81.87%	0.654
RF	90.72%	86.33%	92.56%	82.43%	0.657
LightGBM	90.33%	87.29%	92.70%	82.23%	0.658
SVM poly	91.10%	82.41%	92.77%	81.66%	0.633
SVM rbf	**91.30%**	**89.63%**	92.37%	**84.09%**	**0.697**
DenseNet	90.72%	77.45%	91.40%	79.16%	0.586
DenseNet TL	84.91%	82.37%	89.69%	74.85%	0.529
EfficientNet	89.36%	84.36%	91.45%	80.17%	0.613
EfficientNet TL	89.36%	83.77%	91.74%	79.97%	0.608
SEResNet	89.45%	82.81%	90.25%	80.10%	0.606
SEResNeXt	90.91%	85.26%	90.80%	82.35%	0.652
OASIS: No information rate: 78.05%					
LR	81.60%	80.04%	85.04%	76.18%	0.541
DT	79.27%	71.65%	76.37%	70.84%	0.418
XGBoost	82.71%	80.21%	87.15%	77.11%	0.553
RF	81.93%	80.44%	**87.24%**	76.58%	0.548
LightGBM	82.26%	**81.19%**	85.37%	77.11%	0.560
SVM poly	85.25%	77.49%	86.41%	78.04%	0.561
SVM rbf	81.71%	80.48%	86.18%	76.42%	0.547
DenseNet	86.25%	79.76%	86.63%	79.87%	0.597
DenseNet TL	79.93%	78.25%	86.04%	74.28%	0.505
EfficientNet	85.37%	79.37%	87.22%	78.95%	0.579
EfficientNet TL	**86.47%**	80.63%	86.63%	**80.41%**	**0.608**
SEResNet	80.82%	77.55%	85.29%	74.65%	0.504
SEResNeXt	81.37%	78.27%	85.89%	75.35%	0.518

The best balanced accuracy of 81.19% was reached for the LightGBM model and the best AUROC of 87.24% was achieved by the RF model. The polynomial SVM reached the best F_1-Score of 78.04% and the best MCC of 0.561 within the classical ML model. The worst results were reached for the DT (Accuracy: 79.27%, Balanced accuracy: 71.65%, AUROC: 76.37%, F_1-Score: 70.84%, MCC: 0.418). The DenseNet model which was pretrained on the LDM-100k dataset reached the worst

accuracy of 79.93% within the DL models. The SEResNet reached the smallest balanced accuracy of 77.55% within the DL models.

Most of the models reached reasonable performances during the external validation. It can therefore be inferred that both ML and DL models acquired patterns that are transferable to external datasets.

5.4. Correlated features

The aspects resulting from the feature correlation analysis are shown in Suppl. Mat. F. The 107 volumes are consolidated into 48 aspects. Of those, 14 included individual features. Nine aspects included more than two features. At least one pair of left and right hemispheric volumes is included within 30 aspects. `Aspect_27` included eight regions within the middle-temporal and inferior-temporal cortex, which were previously [78–83] associated with AD progression. `Aspect_34` included five ventricular regions corresponding with the association between ventricular enlargement and AD [79,84].

5.5. Local model interpretation

The concept behind local model explanations is to identify the regions that impact the model's prediction for a particular subject. In the case of LIME and SHAP, positive values denote that the model has recognized that a specific feature or aspect increases the subject's risk for AD. Conversely, negative values imply that the model has learnt that this feature reduces the patient's AD risk. For both GradCAM and GradCAM++, the values suggest an impact on the prediction without indicating a specific direction.

In this work, two methods were employed to illustrate local explanations for the ADNI AD subject 002_S_0816. First, Fig. 2 and Suppl. Mat. G - M visualize MRI slices and the activated brain regions for different models demonstrated as heatmaps. These heatmaps reveal the regions which are relevant for the predictions of the DL and ML models. In comparison to this, Fig. 2(a) demonstrates the inverse *p*-value map calculated during the VBM analysis as a ground truth heatmap. To further explore feature importance, a second visualization technique is employed. Fig. 3 visualizes a matrix displaying the contribution of each aspect to the model prediction for a specific subject. The rows show the aspects included in this work and the columns display the models which were trained. The color in the matrix visualizes the feature importance calculated for each explanation method and each aspect. As stated previously, for LIME and SHAP, positive values denote the model identified a particular aspect as increasing the subject's risk for AD, whereas negative values indicate the model learnt that this feature decreases the patient's risk. Both visualization techniques, the heatmaps and the feature importance plot, present identical information.

The results in Fig. 3 show that the most relevant feature for all ML models is `aspect_27`, which includes regions within the middle temporal and inferior temporal cortices as well as the hippocampi and amygdalae of both hemispheres and thus corresponds with previous AD research [78–83]. The positive value means that the subject's features observed for this aspect increased the patient's AD risk. All classical ML models used additional features for the prediction of the AD status of subject 002_S_0816. Most of these features reached absolute, normalized values smaller than 0.2. Features which reached absolute, normalized values higher than 0.4 for at least one of the models and explanation methods are `aspect_30`, `aspect_12`, `aspect_3`, and `aspect_34`. `Aspect_30` reached values higher than 0.4 for the polynomial SVM explained using SHAP, as well as LIME and SHAP explanations for the LR model. For all methods, a positive association with AD was observed. The aspect includes the caudal middle frontal gyri of both hemispheres. Reduced cortical thickness [85,86] as well as reduced brain volume [86] was associated with AD in previous research. The LIME method identified `aspect_12` as a relevant feature for the DT which reduced the AD risk of subject 002_S_0816. `Aspect_12` includes the precuneus and superior parietal lobule of both hemispheres. Reduction in the cortical thickness of the left superior parietal lobule was previously associated with AD [85]. Additionally, [87] found that the volume of the precuneus is associated with impaired visuospatial functioning. By explaining the LR model, SHAP identifies a protective association with AD for `aspect_3`, which includes the left and right cerebellum cortex and was associated with AD progression in previous research [88,89]. LIME identified `aspect_34` as having a negative impact on the polynomial SVM prediction. `Aspect_34` includes multiple ventricular volumes and ventricular enlargement was previously associated with AD progression [79,84].

Compared to the classical ML models, the DL models took into consideration a higher number of regions for the prediction of the disease status of subject 002_S_0816 as can be seen in Fig. 3.

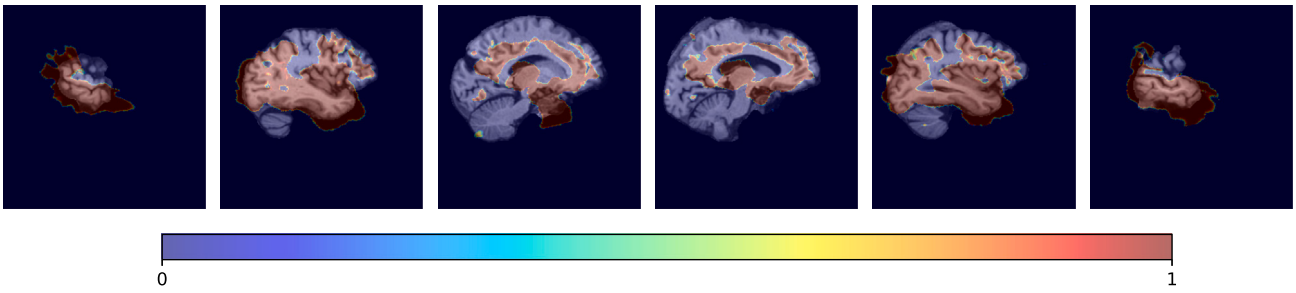
For the DenseNet model which was trained from scratch, the most relevant brain regions were the optical chiasm (Ranks: LIME and GradCAM++: 2, SHAP and GradCAM: 3), `aspect_2` (Ranks: LIME: 1, GradCAM and GradCAM++: 6), `aspect_19` (Ranks: SHAP: 1, LIME: 5), `aspect_22` (Ranks: SHAP: 2), the FreeSurfer region called CSF (Ranks: GradCAM and GradCAM++: 1, LIME and SHAP: 4), and `aspect_5` (Ranks: GradCAM: 2, GradCAM++: 3, LIME: 6).

The optical chiasm was not one of the main brain regions previously associated with AD. However, associations between the optic nerve [90], the visual pathway [91] and AD have been reported. `Aspect_2` includes the central and middle anterior parts of the corpus callosum which was previously associated with AD [92]. `Aspect_19` includes the left and right isthmus of the cingulate gyrus. For the left hemisphere of this brain structure, reduced cortical thickness and surface area were observed in [85]. The lingual gyri of both hemispheres were consolidated into `aspect_22`. For this region of the left hemisphere, reduced cortical thickness [85] was observed for subjects with AD in comparison to CN subjects. The FreeSurfer definition of the CSF region approximately corresponds to the transverse cerebral fissure, which was not in focus for AD detection. However, the dilatation of lateral parts of the transverse fissure was associated with AD in previous research [93]. In addition, the region is located near the third ventricle which was associated with AD progression [94]. Finally, `aspect_5` includes the thalamus proper of both hemispheres. This region reached the 3rd rank in the VBM analysis and was previously associated with AD [95,96].

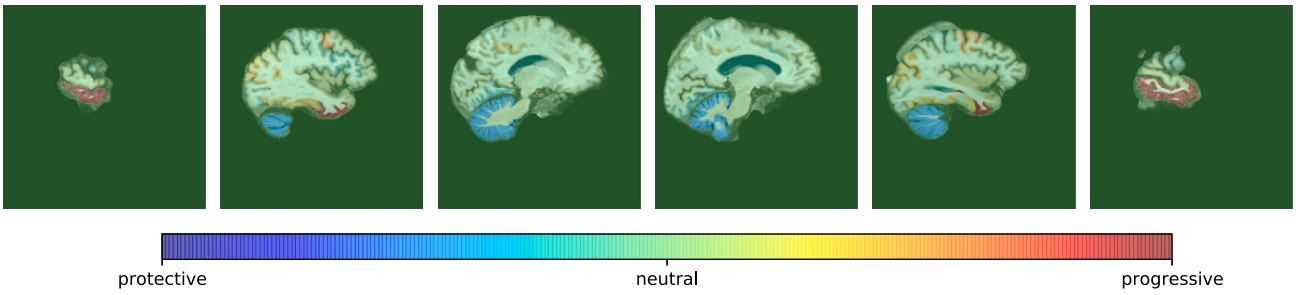
For the DenseNet model which was pretrained on the LDM-100k dataset, especially the GradCAM and GradCAM++ methods identified a large number of relevant brain regions. These regions include the left vessel (Ranks: LIME and SHAP: 1), `aspect_17` (Ranks: LIME and SHAP: 2), the rostral anterior cingulate gyrus of the right hemisphere (Ranks: GradCAM and GradCAM++: 1, LIME: 6), and `aspect_2` (Ranks: GradCAM and GradCAM++: 2).

Of these regions, `aspect_2` was also one of the most relevant regions for the DenseNet model which was trained from scratch. Although the regions do not include the hippocampus or entorhinal cortices which were the most prominent regions in AD detection, all of the selected regions have been associated with AD in previous research. These associations include modalities beyond the MRI as DL methods might detect patterns that are classically identified using different modalities such as functional MRI (fMRI), or PET. The FreeSurfer definition of the left vessel includes vessel regions in the inferior pallidum and putamen. Cholinergic neuronal loss [97] as well as atrophy [98] in these regions were previously associated with AD. `Aspect_17` includes the left and right pars opercularis regions. Previous work found altered functional connectivity [99] of subjects with AD in this region. The functional connectivity [100] as well as atrophy patterns [101] of different subregions of the anterior cingulate cortex including the caudal anterior cingulate cortex and the rostral anterior cingulate cortex have been also associated with AD.

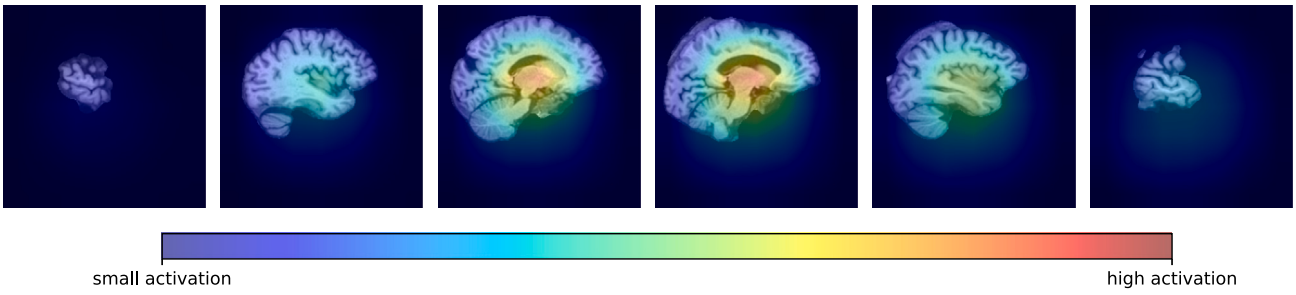
The EfficientNet model which was trained from scratch showed a similar pattern for the prediction of the disease status of this subject.



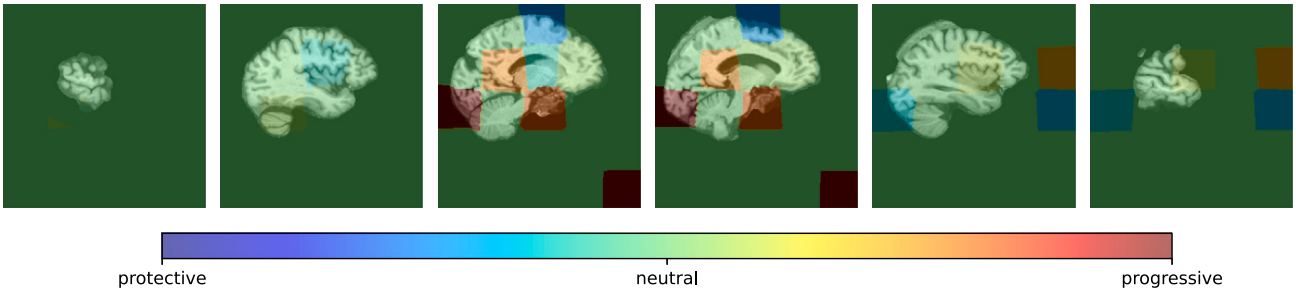
(a) VBM analysis results (inverse $p - value$)



(b) SHAP explanations for polynomial SVM



(c) GradCAM++ heatmap for DenseNet



(d) SHAP explanations for DenseNet

Fig. 2. VBM results (inverse *p*-value) and local explanations visualized as heatmaps on T1-weighted MRIs for ADNI AD subject 002_S_0816 . The visualization shows MRI intensities and heatmaps for slices 50, 75, 100, 125, 150, and 175.

The most relevant regions for this model were: the 4th ventricle (Ranks: LIME: 1), the right vessel (Ranks: LIME and SHAP: 2), `aspect_29` (Ranks: SHAP: 1), `aspect_9` (Ranks: GradCAM: 1, GradCAM++: 6), `aspect_30` (Ranks: GradCAM++: 1, GradCAM: 2), and `aspect_2` (Ranks: GradCAM++: 2, GradCAM: 5).

`Aspect_30` has been also selected as one of the most relevant regions for the prediction of this subject for the polynomial SVM and LR models. Additionally, `aspect_2` was a relevant region for both DenseNet models. For these regions, relevant associations in previous research were discussed above. In volumetric analysis, the lateral and inferior lateral ventricles were more affected by ventricular enlargement in AD in comparison to the 4th ventricle. Additionally, the 4th ventricle reached the last rank during the VBM analysis. The right vessel reached the 13th rank within the VBM analysis. The FreeSurfer definition of the vessels includes vessel regions in the inferior pallidum and putamen. For these regions, cholinergic neuronal loss [97] as well as atrophy [98] were observed in association with AD. The fusiform gyri of both hemispheres are included in `aspect_29` and have been

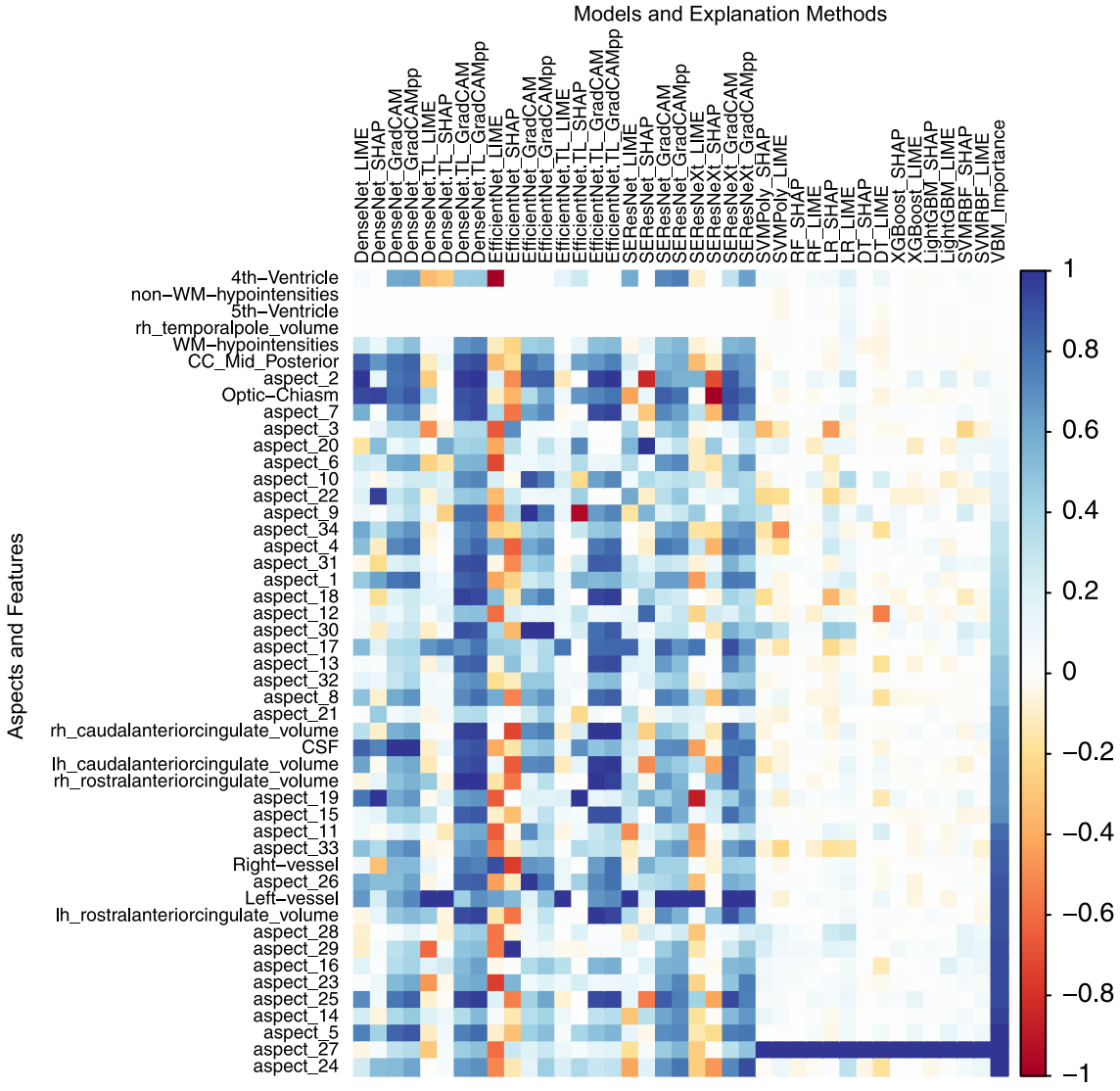


Fig. 3. Matrix displaying the contribution of each aspect to the model prediction for ADNI subject 002_S_0816 with diagnosis AD. The rows show the aspects included in this work and the columns display the models which were trained. The color in the matrix visualizes the feature importance calculated for each model and each aspect. For LIME and SHAP, positive values denote the model identified a particular aspect as increasing the subject's risk for AD. Negative values indicate the model learned that this feature decreases the patient's AD risk.

associated with atrophy in the preclinical stages of AD [102]. `Aspect_9` includes the paracentral lobule of both hemispheres. Reduced cortical thickness was previously observed for subjects with AD in this region, in [85]. Additionally, differences in the structural cortical network of the paracentral lobule were found in [103].

For the pretrained EfficientNet model, the most relevant features were the left vessel (Ranks: LIME: 1), `aspect_17` (Ranks: LIME: 2), `aspect_19` (Ranks: SHAP: 1, LIME: 6), `aspect_9` (Ranks: SHAP: 2, LIME: 9), the right caudal anterior cingulate gyrus (Ranks: GradCAM and GradCAM++: 1), the right rostral anterior cingulate gyrus (Ranks: GradCAM: 2, GradCAM++: 6), and `aspect_2`. (Ranks: GradCAM++: 2, GradCAM: 5, LIME: 8).

Many of these regions were also classified as relevant regions in the DL models described above. These are the left vessel (pretrained DenseNet), `aspect_17` (pretrained DenseNet), `aspect_19` (DenseNet trained from scratch), `aspect_9` (EfficientNet trained from scratch), the right rostral anterior cingulate gyrus (pretrained DenseNet), and `aspect_2` (included in all DL models described above). Most of these regions were not in focus of AD detection in previous research. However, for most regions or at least for adjacent regions associations were found which were previously discussed. In addition to this, changes within the functional connectivity [100] and atrophy patterns [101] of different subregions of the anterior cingulate cortex including the caudal anterior cingulate cortex and the rostral anterior cingulate cortex have been found in subjects with AD.

The regions identified as most relevant for the local prediction of subject 002_S_0816 using the SEResNet model were the left vessel (Ranks: LIME, GradCAM, and GradCAM++: 1), `aspect_17` (Ranks: LIME: 2, GradCAM: 4), `aspect_20` (Ranks: SHAP: 1, LIME: 5), `aspect_2` (Ranks: SHAP: 2), `aspect_25` (Ranks: GradCAM: 2, GradCAM++: 5, SHAP: 6), and the optical chiasm (GradCAM++: 2, GradCAM: 3 LIME: 7).

Similar to the results of the models which were described before, this model does not concentrate on the regions which were mostly associated with AD in previous research. However, for most of the regions, there have been some articles that associated these or adjacent regions with AD progression. Some of the regions which were identified as being relevant for the SEResNet model were also relevant for the DL models introduced before. These regions are the left vessel (pretrained DenseNet and pretrained EfficientNet), `aspect_17` (pretrained DenseNet and pretrained EfficientNet), `aspect_2` (included in the local explanations of all DL models described above) and the

optical chiasm (DenseNet trained from scratch). The integration of these regions in previous AD detection has been discussed above. The regions of the cuneus and pericalcarine cortices of both hemispheres are consolidated in aspect_20. Although both regions were not the focus of early AD prediction, for both brain structures, reduced cortical thickness was observed in [85] for the left hemisphere. Aspect_25 included the left and right accumbens area, which also has been previously associated with AD [104]. Additionally, this region achieved the 5th rank in the VBM analysis projected to subject 002_S_0816.

For the SEResNeXt model, similar patterns were found. The most relevant regions were: the left vessel (Ranks: LIME, GradCAM, and GradCAM++: 1), aspect_19 (Ranks: LIME: 2), the optical chiasm (Ranks: SHAP: 1, GradCAM and GradCAM++: 3), aspect_2 (Ranks: SHAP: 2, LIME: 4, GradCAM: 6), aspect_25 (Ranks: GradCAM: 2, SHAP and GradCAM++: 8, LIME: 9), and aspect_24 (Ranks: GradCAM++: 2, SHAP: 5).

Many of these regions overlap with regions identified as relevant in the different DL models trained in this work. These regions include the left vessel (SEResNet, pretrained EfficientNet, and pretrained DenseNet), aspect_19 (pretrained EfficientNet and DenseNet trained from scratch), the optical chiasm (DenseNet trained from scratch and SEResNet), aspect_2 (included in the local explanations of all previously described models), and aspect_25 (SEResNet). Explanations about the relevance of these regions within previous research have been given above. Additionally, aspect_24 consolidated the entorhinal gyri and is one of the regions which are quite prominent in previous AD research [79,105]. For the VBM analysis results which were projected to the MRI scan of subject 002_S_0816, this region achieved the highest score.

To compare the aspect rankings of the methods for the individual explanations of subject 002_S_0816 with a focus on highly ranked aspects, the Normalized Discounted Cumulative Gain (NDCG) was used. The similarity plot is shown in Fig. 4. NDCG values lie in the range between 0 and 1. It should be noted, that due to the dependence on feature importances, the NDCG matrix is not symmetric. Each value in the matrix describes the normalized feature importances of a reference method ranked in the order of the feature importances of the comparison method after applying a logarithmic discount. Changing the reference method changes the ground truth feature importance scores, resulting in asymmetries.

Both axes of Fig. 4 can be split into three different parts. The first part consists of the explanations of the DL models, the second part includes the explanations of the ML models and the last part is the VBM ground truth ranking.

The first experiment focuses on the rankings of the DL models. These are visualized in the upper left corner of the matrix. The plot shows that most of the regional rankings of the DL models show high similarity across each other. The 10% quantile of all comparisons in this group was 0.626. This observation means, that more than 90% of the model comparisons show moderate to high similarities. The smallest NDCG score was 0.464 which was reached when the SHAP explanations of the pretrained DenseNet models were compared to the GradCAM explanations of the same model. Overall it was observed, that most of the models reached smaller similarities if the SHAP explanations of the pretrained DenseNet models were used as reference method.

The second experiment investigates the similarities in the feature rankings across ML models. The data for this experiment can be found in the lower right corner of the matrix. The results showed higher similarities in this group in comparison to the DL models. The smallest NDCG score in this comparison was 0.791 which was reached by comparing the SHAP explanations of the LR to the LIME explanations of the LightGBM model. This leads to the conclusion, that all ML models show moderate to high similarities in the feature rankings.

Due to the fact that the NDCG scores are not symmetric, two different experiments are possible for the comparison of ML and DL feature importance. The first comparison uses the DL explanations as a reference and can be found in the upper right corner of the matrix. These comparisons showed smaller similarities than the DL models. The NDCG scores were between 0.496 and 0.917. The smallest score was reached for the comparison between the SHAP explanations of the pretrained DenseNet and the LIME explanations of the LR model. The highest score was reached when the GradCAM++ explanations were compared to the LIME explanations of the XGBoost model. The 10% quantile of the data was 0.589.

The second experiment uses the ML models as a reference method. It can be seen that those comparisons achieved worse NDCG scores than the previously described comparison. All values are in the range of 0.234 and 0.726. The 10% quantile is 0.314. The comparison which achieved the smallest NDCG score was the comparison between the SHAP explanations of the DT and the LIME explanations of the DenseNet trained from scratch. The highest similarity was reached for the comparison of the LIME explanations of the LR model and the GradCAM++ explanations of the EfficientNet model trained from scratch.

The results show that there are structural differences in the local feature explanations between DL and classical ML models for subject 002_S_0816. The classical ML models focus on a small number of features and all prefer regions which were previously associated with AD such as the hippocampi and middle temporal gyri. The DL models instead focused on a larger number of regions and combined regions which were in the focus of AD in previous research, like aspect_24 and regions which were not prominent in AD detection before. However, for most of the regions, previous research was found that shows some associations with these regions in different modalities.

The comparison to the VBM analysis (last row of Fig. 4) shows high similarities for both, ML and DL models. The DL models reached NDCG scores between 0.753 which was reached for the LIME explanations of the DenseNet model and 0.890 which was achieved for the GradCAM++ explanations of the SEResNeXt model. The 10% quantile was 0.772. The results reached for the classical ML models were higher than those achieved for the DL models. The NDCG scores were in the range of 0.834 which was reached for the local SHAP explanations of the LR model and 0.922 achieved for the SHAP explanations of the RF. The 10% quantile was 0.840. An analysis showing similar results, was performed for the ADNI CN subject 941_S_4292. These results are shown in Suppl. Mat. N - W.

5.6. Global model interpretation

In this section, the global rankings of the activated brain regions for the ML and DL models were presented and compared to each other. The most relevant brain regions to differentiate subjects with AD from CN subjects were investigated using feature importance in classical ML methods as well as mean absolute heatmap activations per aspect in DL models.

As a ground truth comparison, VBM is used. The registration and GM segmentation are implemented using the CAT12 software [106], which is based on the SPM [107]. Smoothing was performed using a Gaussian kernel with a full-width at half-maximum (FWHM) of 8 mm. An ANCOVA statistical test was implemented using the Python package pingouin v0.5.3 [108]. The model was corrected for the covariates age, gender, MRI-field strength, and eTIV. The Bonferroni correction [109] was used to correct for bias due to multiple testing. It was implemented using the Python package statsmodels v0.13.2 [110]. The significance level was set to $\alpha = 0.01$. As ground truth, aspect-wise region importance was calculated using the mean inverted p-value. The ten most important aspects and their mean inverted p-values were summarized in Fig. 5. The most important aspect was aspect_27. Aspect_27 includes regions within the middle temporal and inferior temporal cortex, also associated with AD in previous research [78–83]. The second most important region was aspect_24, which includes the entorhinal gyri and was previously associated with AD [79,105]. The

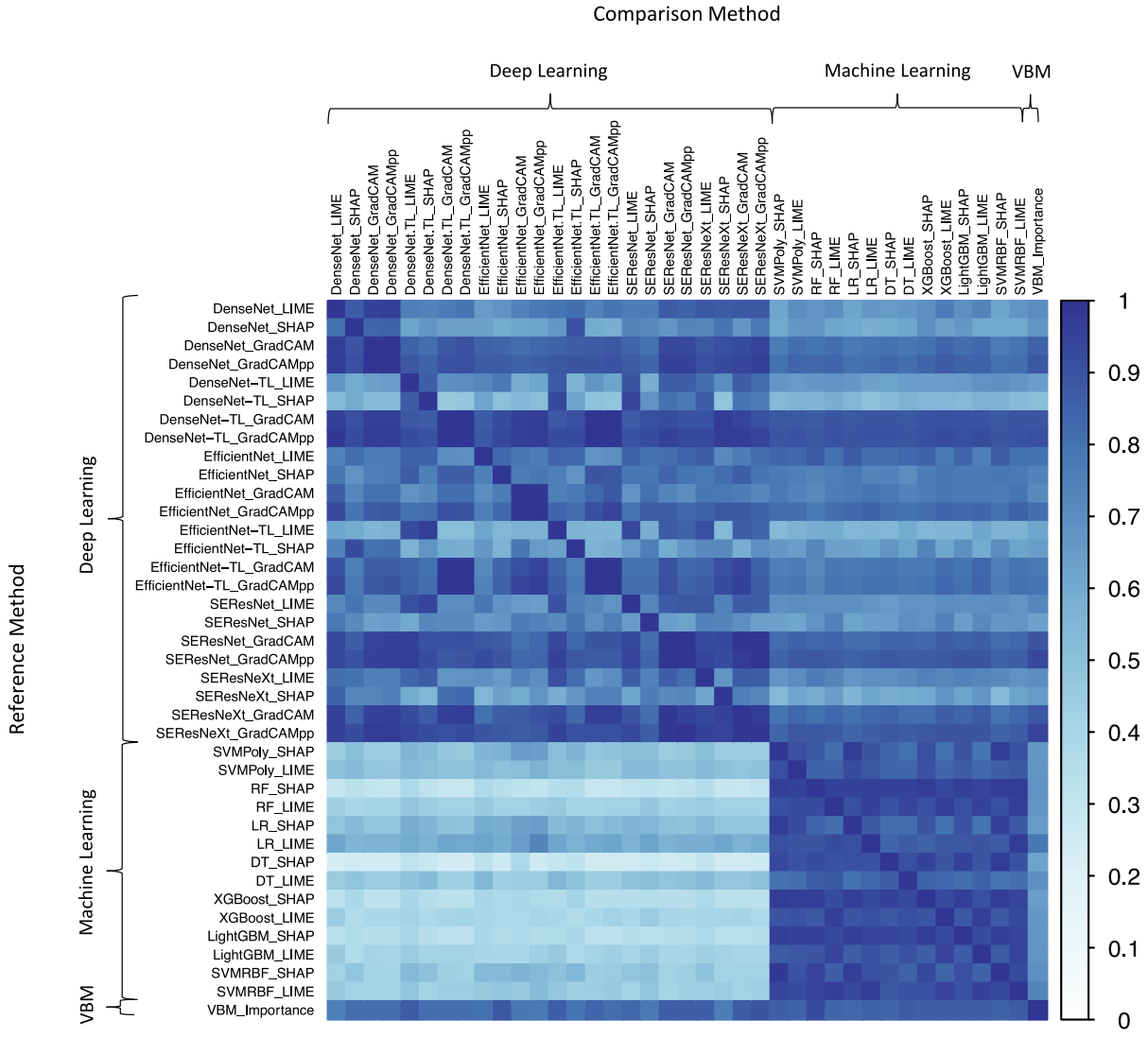


Fig. 4. Correlation plot showing the Normalized Discounted Cumulative Gain (NDCG) between local feature importances of the explainability methods for AD subject with ID 002_S_0816. The correlation matrix is not symmetric because the NDCG is based on feature importance scores of the features visualized as rows (reference method) and rankings visualized as columns (comparison method). Changing the reference method changes the ground truth feature importance scores, resulting in asymmetries.

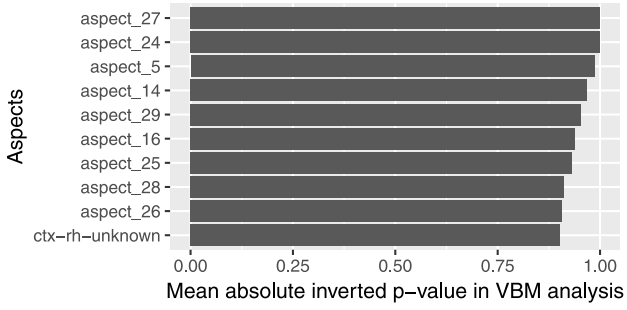


Fig. 5. Mean inverted *p*-value of aspects calculated during VBM.

third most relevant region was `aspect_5`, including the Thalamus-Proper of both hemispheres. The thalamus was connected to AD in the previously published literature [95,96].

The global LIME and SHAP feature rankings of the polynomial SVM are summarized in Fig. 6 using SHAP summary plots. The SHAP and LIME values were calculated by consolidating correlated features into aspects. In these plots, the interactions of the importance scores and the feature values are visualized for individual volumes to improve interpretability. This leads to identical SHAP and LIME value distributions for all features in an aspect. The aspects are stated behind the feature names. The plots show that `aspect_27`, including the volumes of the left and right hippocampi, the amygdalae, the middle temporal, and the inferior temporal gyri, was most important. Decreased brain volumes (blue) are associated with AD. All regions consolidated in `aspect_27` were previously connected with AD [78–83] and the relationship, the model learned corresponds to the atrophy pattern. The same applies to `aspect_24` [79,105] which ranked second using SHAP and fourth using LIME. `Aspect_24` includes the volumes of the left and right entorhinal gyri. The second most important LIME aspect was `aspect_34`, including multiple ventricular volumes. Ventricular enlargement was previously associated with AD [79,84] and matches

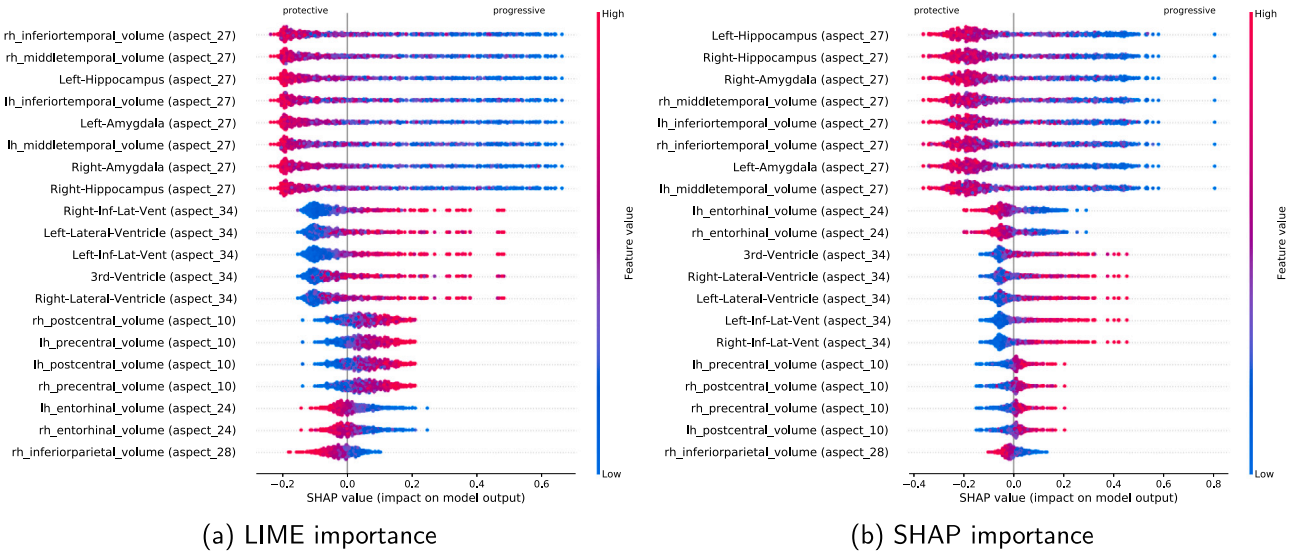


Fig. 6. Global feature importances for polynomial SVM, ADNI training set.

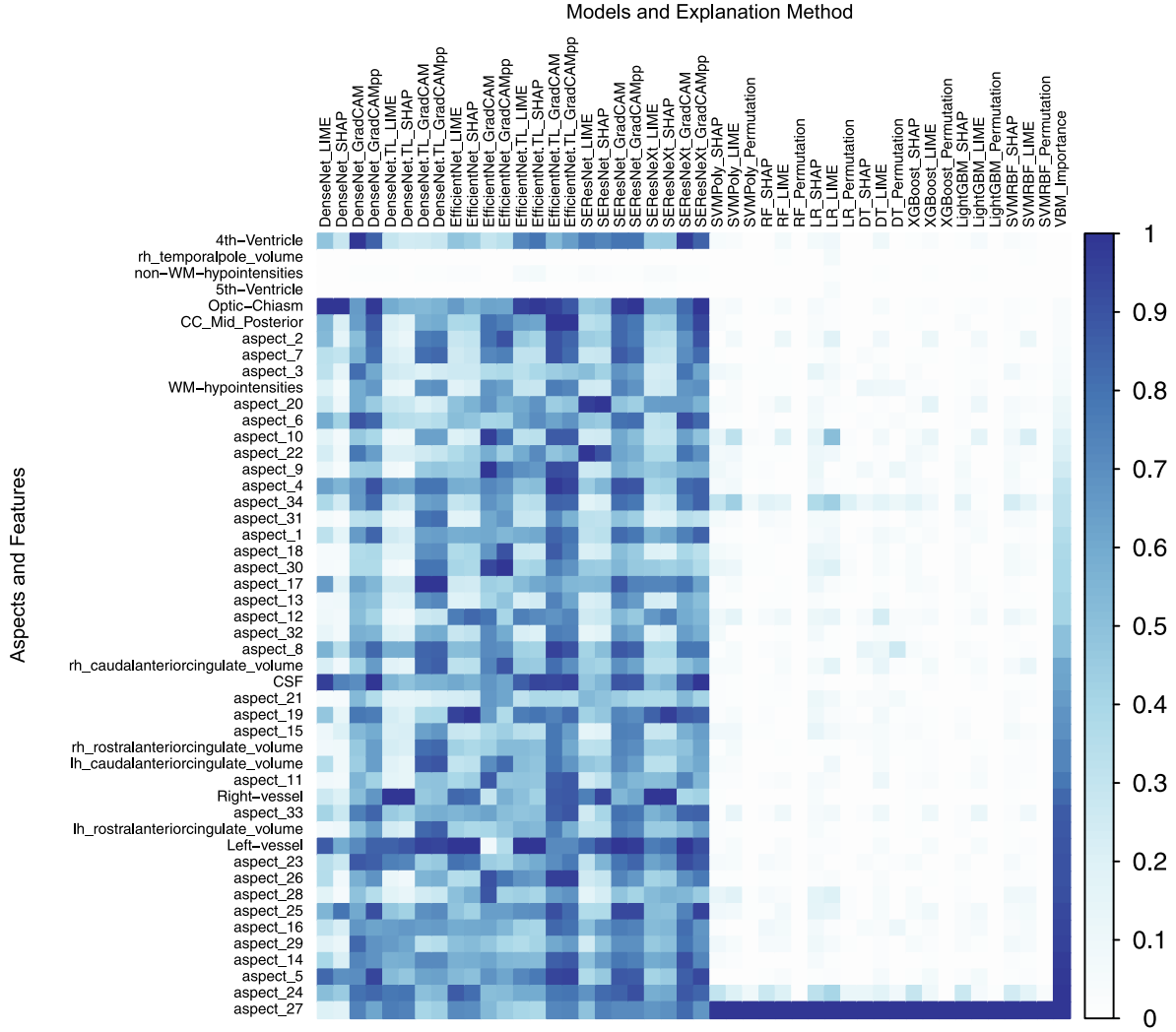


Fig. 7. Heatmap showing the global, normalized feature importances for the DL and classical ML models using different explanation methods.

the pattern, the model learned. aspect_34 was ranked third for SHAP.

Fig. 7 shows a heatmap summarizing the normalized feature importances calculated for the DL and ML models using the described explanation methods. The DL feature importances include mean local GradCAM and GradCAM++ scores on the training set for the FreeSurfer segmented brain structures which are consolidated into aspects. For SHAP and LIME, the absolute sum of all local scores was calculated to take into account both protective and progressive regions.

The results show that similar to the polynomial SVM, all ML models strongly focus on aspect_27. As was previously mentioned, aspect_27 includes regions within the middle temporal and inferior temporal cortex, which were strongly associated with AD in previous research [78–83]. Aspect_27 is also the region which reached the highest scores during the VBM analysis. Beyond aspect_27, the classical ML models ranked different regions with smaller importance scores. These regions include aspect_24 consisting of the left and right entorhinal gyri, aspect_34 consisting of ventricular volumes, as well as aspect_10 consisting of the postcentral and precentral gyri. For the entorhinal gyri, atrophy patterns were previously associated with AD [79,105]. For aspect_34, ventricular enlargement has been identified as a relevant sign for AD [79,84]. Previous research shows reduced cortical thickness for subjects with AD in the postcentral and precentral gyri [85].

In comparison to this, all DL models show high feature importance for a wider range of regions. For the DenseNet model, which was trained from scratch, the optical chiasm was identified as one of the most important regions (Ranks: LIME and SHAP: 1, GradCAM++: 2). Additionally, aspect_25 (Ranks: SHAP: 2, GradCAM++: 4) which includes the left and right accumbens area, the FreeSurfer CSF region (Ranks: GradCAM++: 1, LIME: 2, SHAP: 3), aspect_6 (Ranks: GradCAM: 2, LIME: 7, SHAP: 8) which includes the Cerebellum White Matter and the Diencephalon of both hemispheres, as well as the brain stem, and the 4th ventricle (Ranks: GradCAM: 1, GradCAM++: 8) were identified as relevant regions for this model.

The optical chiasm was not one of the main MRI regions previously associated with AD. However, associations between the optic nerve [90], the visual pathway [91] and AD were reported. The left and right accumbens area was previously associated with AD in [104] and achieved the 7th rank in the VBM analysis. The CSF region approximately corresponds to the transverse cerebral fissure, which was not in focus for AD detection in previous literature. However, [93] found some connection between the dilatation of lateral parts of the transverse fissure and AD. The left and right Cerebellum White Matter, the brain stem, as well as the left and right ventral Diencephalon are not in focus in most AD analyses. However, there is some previous research that identified changes in the morpho-functional mesoscopic traits [111] as well as amyloid plaques [112] in these regions. In volumetric analysis, the lateral and inferior lateral ventricles were more affected by ventricular enlargement in AD in comparison to the 4th ventricle.

The most relevant regions chosen by the pretrained DenseNet model are the right (Ranks: LIME and SHAP: 1) and left vessels (Ranks: LIME, SHAP, GradCAM and GradCAM++: 2), and aspect_17 (Ranks: GradCAM and GradCAM++: 1) which includes the left and right pars opercularis regions.

The FreeSurfer definition of the vessels includes vessel regions in the inferior pallidum and putamen. For these regions, cholinergic neuronal loss [97] as well as atrophy [98] were observed in association with AD. Additionally, the left vessel reached the 11th rank within the VBM analysis and the right vessel was ranked 14th. Aspect_17 includes the left and right pars opercularis regions. This region is not in the focus of AD research in previous articles. However, altered functional connectivity [99] of subjects with AD was identified previously in this region.

The EfficientNet which was trained from scratch achieved high feature importance for the left vessel (Ranks: LIME: 1, SHAP: 2), aspect_19 (Ranks: SHAP: 1, LIME: 2) which includes the left and right isthmus of the cingulate gyri, aspect_9 (Ranks: GradCAM: 1, GradCAM++: 8), consolidating the paracentral lobule of both hemispheres, aspect_10 (Ranks: GradCAM: 2, GradCAM++: 6) consisting of the postcentral and precentral gyri, aspect_30 (GradCAM++: 1, GradCAM: 3) which includes the caudal middle frontal gyri of both hemispheres, and aspect_2 (GradCAM++: 2, GradCAM: 7) which consolidates the caudal middle frontal gyri of both hemispheres. The left vessel was also selected as a relevant region for the pretrained DenseNet model. For the left isthmus of the cingulate gyrus, reduced cortical thickness and surface area were observed in [85]. For the paracentral lobule of both hemispheres, reduced cortical thickness was previously observed for subjects with AD in [85]. Additionally, differences in the structural cortical network of the paracentral lobule were found in [103]. Aspect_10 consists of the postcentral and precentral gyri. Previous research shows reduced cortical thickness for subjects with AD in these regions [85]. For the caudal middle frontal gyri of both hemispheres, reduced cortical thickness [85,86] and brain volume [86] were associated with AD. The central and middle anterior parts of the corpus callosum were consolidated in aspect_2 and have been previously associated with AD progression [92].

The relevant regions identified for the EfficientNet which has been pretrained on the LDM-100k dataset were the left vessel (Ranks: LIME and SHAP: 1), the optical chiasm (Ranks: LIME and SHAP: 2, GradCAM: 4, GradCAM++: 9), middle posterior parts of the corpus callosum (Ranks: GradCAM and GradCAM++: 1), aspect_4 (Ranks: GradCAM: 2, GradCAM++: 5), which includes the pallidum of both hemispheres and aspect_26 (Ranks: GradCAM++: 2, GradCAM: 3). As was previously described, the optical chiasm has been selected as one of the most relevant regions in the DenseNet model which was trained from scratch. The left vessel was selected for the pretrained DenseNet, as well as both EfficientNet models. Different parts of the corpus callosum were also associated with AD in previous research [92]. Cholinergic neuronal loss within the pallidum was associated with AD in [97]. Aspect_26 consolidated the posterior cingulate cortex of both hemispheres and reached the 9th-highest relevance score during the VBM analysis. Atrophy [101] within the posterior cingulate cortex was previously found as an early sign of AD.

For the SEResNet model, the most relevant brain regions were: aspect_22 (Ranks: LIME: 1, SHAP: 4), aspect_20 (Ranks: SHAP: 1, LIME: 2), consolidating the regions of the cuneus and pericalcarine cortices of both hemispheres, the right (Ranks: SHAP: 2, LIME: 5), and left vessels (Ranks: GradCAM: 1, GradCAM++: 2), and the optical chiasm (Ranks: GradCAM++: 1, GradCAM: 2).

Some of these regions overlap with the regions selected in different DL models and have been thus already discussed above. These regions are the right vessel (pretrained DenseNet), the left vessel (pretrained DenseNet, both EfficienNets), and the optical chiasm (DenseNet trained from scratch and pretrained EfficientNet). Aspect_22 includes the lingual gyrus of both hemispheres. In the left hemisphere of this region, reduced cortical thickness [85] was observed for subjects with AD. Although the cuneus and pericalcarine regions were not the focus of early AD prediction, for both brain structures, reduced cortical thickness was observed in [85] for the left hemisphere.

The most relevant regions identified for the SEResNeXt model are the right vessel (Ranks: LIME and SHAP: 1), aspect_19 (Ranks: LIME and SHAP: 2, GradCAM: 8), which includes the left and right isthmus of the cingulate gyri, the left vessel (Ranks: GradCAM: 1), the 4th ventricle (Ranks: GradCAM: 2, LIME and SHAP: 3, GradCAM++: 6), the FreeSurfer CSF region (Ranks: GradCAM++: 1, GradCAM: 7) and the optical chiasm (Ranks: GradCAM++: 2). All of these regions were also identified as highly relevant regions in the DL models described before. The left vessel was included in the explanations of the pretrained DenseNet, both EfficientNet models as well as the SEResNet

model. The right vessel overlaps with the explanations of the pretrained DenseNet and the SEResNet model. Aspect_19 was already included in the EfficientNet model trained from scratch. The 4th ventricle was also selected in the explanations of the DenseNet model trained from scratch. The FreeSurfer CSF region was included in the explanations of the DenseNet model trained from scratch. Finally, the optical chiasm was selected as a relevant region in the explanations of the DenseNet model trained from scratch, the pretrained EfficientNet model, and the SEResNet model.

It has been previously mentioned that aspect_27 which includes regions within the middle temporal and inferior temporal cortices, e.g. the hippocampi of both hemispheres, was the most relevant region for the classical ML models and was also the focus of previous AD research [78–83]. This region was not one of the most relevant regions identified for the DL models. However, all explanation methods of the DL models weighted this region with a relevance score larger than 0. The smallest normalized feature relevance score of 0.185 was calculated for the SHAP explanations of the DenseNet model trained from scratch. The highest normalized feature relevance score was 0.822 which was calculated for the GradCAM explanation of the SEResNeXt model. The 10% quantile of the feature rankings for aspect_27 in the DL models was 0.390. This means, that 10% of the DL model explanations calculated a normalized relevance score of less than 0.390 for aspect_27.

Additionally, the average normalized feature importances across all DL models and all explanation methods were calculated. The most relevant regions were the left vessel (average, normalized feature importance: 0.822), the optical chiasm (average, normalized feature importance: 0.750), the CSF region (average, normalized feature importance: 0.728), and aspect_24 (average, normalized feature importance: 0.706). The FreeSurfer definition of the left vessel includes vessel regions in the inferior pallidum and putamen which were associated with cholinergic neuronal loss [97] as well as atrophy [98] in previous research. Although regions near the optical chiasm and the CSF regions were associated with AD in previous research, these are not in focus. Aspect_24, however, consolidated the entorhinal gyri of both hemispheres and is one of the regions which are quite prominent in previous AD research [79,105]. This region reached the second rank within the VBM analysis. These findings lead to the assumption that DL models combine information extracted from regions which were prominent in AD detection with information from other regions. The focus on the less prominent regions can lead to the assumption that DL models are able to extract textural information from the MRI scans that exceed the known AD biomarkers.

The results show, that the most important regions correspond with previous research. Regions that are mainly associated with early AD atrophy, such as the hippocampi, or the entorhinal gyri seem to be more relevant in classical ML. The classical ML models focus on a handful of regions for the prediction of AD. In comparison to this, the DL explanations focused on a larger number of regions and combined regions which were previously associated with AD, like the left and right vessels which FreeSurfer definitions are located near the inferior pallidum and putamen, or aspect_24 which includes the entorhinal gyri of both hemispheres and regions which were less prominent in previous AD detection, like the optical chiasm, the 4th ventricle and the CSF region. One reason for the different regions is, that DL focuses on textural whereas classical ML concentrates on volumetric features.

Similar to the comparison of the local rankings, the NDCG score is used to compare the similarity of the rankings with a focus on highly ranked aspects. A summary of the NDCG is visualized in Fig. 8. NDCG values are in the range of 0 and 1, where 1 indicates perfect similarity and 0 indicates no similarity. It should be noted, that the matrix is not symmetric. Each value in the matrix describes the normalized feature importance of a reference method (visualized as rows) ranked in the order of the comparison method (visualized as columns) after applying a logarithmic discount. If "Method A" detects only a few relevant features and calculates a score close to 0.0 for the remaining features, like classical ML models found in Fig. 7, and "Method B" selects a broader spectrum of features, this results in a higher NDCG score if "Method B" was used as the reference method than if "Method A" was used.

Both axes of the similarity matrix can be split into three different parts which are labeled using curly braces in the plot. These parts are methods that explain DL models, methods that explain classical ML models, and the VBM ground truth. The results show high similarity if methods that explain DL models are compared to each other. The smallest similarity of 0.592 was observed if the SHAP explanations of the DenseNet which was trained from scratch were compared to the GradCAM explanations of the EfficientNet trained from scratch. The second smallest similarity was 0.615 and was observed if the SHAP explanations of the DenseNet which was trained from scratch were compared to the GradCAM++ explanations of the EfficientNet trained from scratch. The 10% quantile of this group is 0.820, meaning that, only 10% of the similarity scores fall below a value of 0.820. Thus, for more than 90% of the comparisons a high similarity was observed.

If the explanations of the classical ML models are compared against each other, a similar pattern was observed with even higher NDCG scores. The smallest NDCG value in this group is 0.846 observed for the comparison of the LIME explanations used for the LR method which was compared to the permutation importance scores of the DT. This means that all comparisons of feature rankings that include ML models show a high similarity.

Due to the fact that the NDCG scores are not symmetric, two different experiments are possible to compare the feature importance computed for the classical ML models with those of the DL models. The first comparison uses the DL explanations as a reference and can be found in the upper right corner of the matrix in Fig. 8. The highest NDCG score of 0.948 was observed if the GradCAM explanations of the SEResNeXt model were compared to the SHAP explanations of the RF model. The smallest score was 0.637 which was observed for the SHAP explanation of the DenseNet model trained from scratch compared to the LIME explanations of the LR model. Overall, relatively small scores between 0.637 and 0.688 were observed if the SHAP explanations of the DenseNet model trained from scratch were used as the reference method. For the entire comparison of explanations generated for DL models used as reference methods and explanations of ML models as comparison methods, 10% of the NDCG scores undercut a value of 0.802.

A different pattern was observed if the classical ML models were used as reference methods and the DL models were used as comparison methods. A visualization of this comparison can be found in the lower-left corner of the matrix in Fig. 8. It can be observed, that smaller NDCG scores were calculated in this comparison. The reason for this observation is that the classical ML methods are based on a small number of features, which means that they have feature importance scores close to 0.0 for many of the remaining features. However, some of these features are highly rated by the DL models. The smallest NDCG score in this group is 0.192 which was observed if the RF permutation importance was compared to the SHAP explanations generated for the fine-tuned EfficientNet. The highest NDCG score is 0.638 observed if the SHAP explanations of the LR were compared to the SHAP explanations of the fine-tuned DenseNet. 10% of the NDCG scores have a value below 0.250.

In conclusion, the analysis revealed noticeable differences between the explanations of the DL and ML models. The first observation was that DL models select a larger number of aspects for the prediction of AD in comparison to the ML models. Additionally, the selected regions differed between both approaches. The classical ML models focused on the hippocampus region and the entorhinal cortex which were previously associated with AD. The DL models used a combination of regions which were previously associated with AD, e.g. the left and right vessels which FreeSurfer definitions are located near the inferior

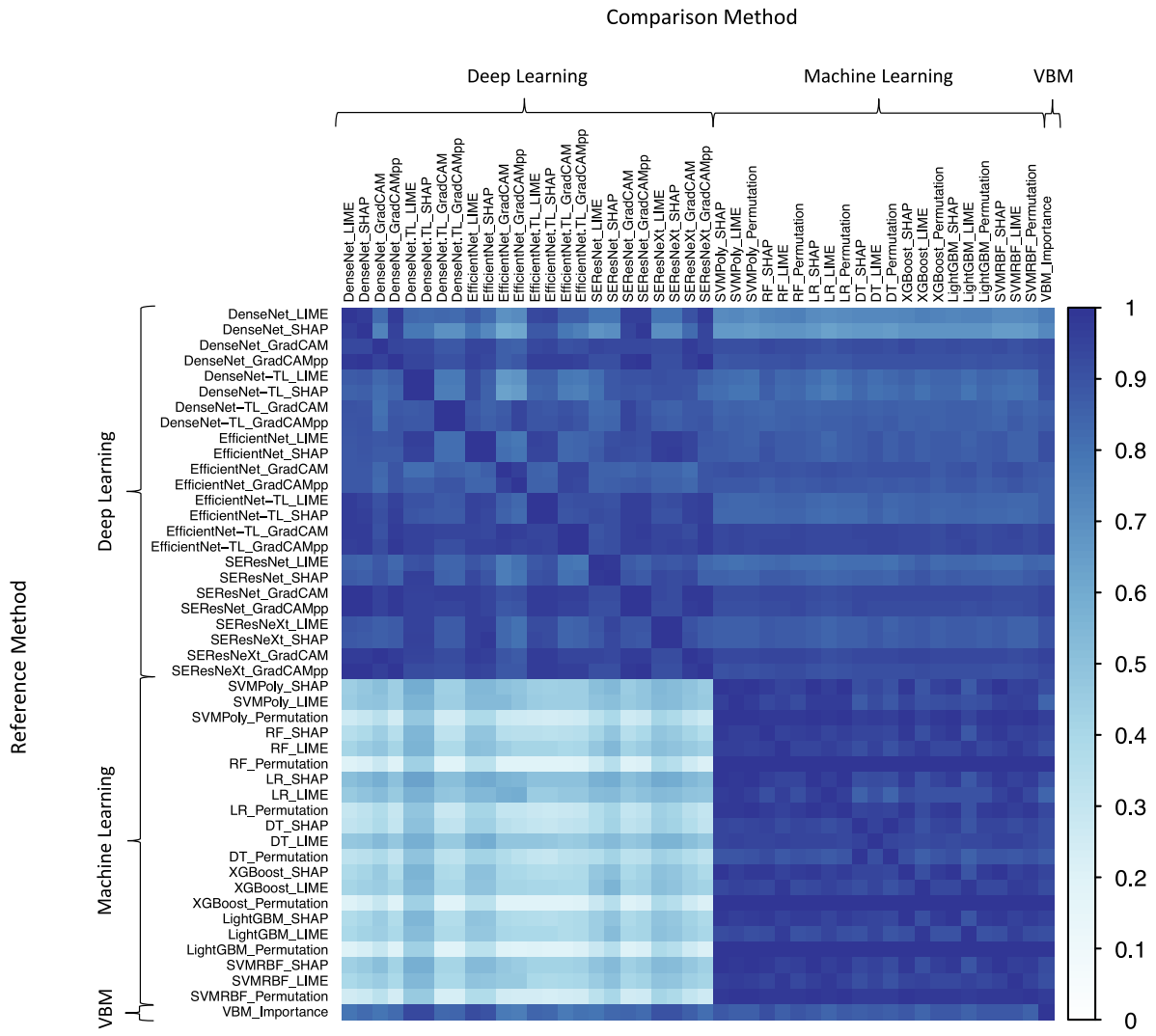


Fig. 8. Similarity plot showing the Normalized Discounted Cumulative Gain (NDCG) between global feature importances of the explainability methods. The correlation matrix is not symmetric because the NDCG is based on feature importance scores of the features visualized as rows (reference method) and rankings visualized as columns (comparison method). Changing the reference method changes the ground truth feature importance scores, resulting in asymmetries.

pallidum and putamen, and `aspect_24` which includes the entorhinal gyri of both hemispheres, and regions which were less prominent in previous AD detection, e.g., the optical chiasm, the 4th ventricle or the CSF region. The NDCG matrix confirmed structural differences between the ML and DL explanations and showed relatively high similarities within the different groups.

Fig. 8 also enables the comparison between the rankings within the ground truth VBM analysis and the ML and DL performances. The case when the VBM analysis is used as a reference method is particularly interesting (last row of Fig. 8). The DL models show NDCG scores between 0.769 for the GradCAM++ explanations of the EfficientNet which was trained from scratch and 0.935 for the SHAP explanations of the fine-tuned DenseNet model. The ML models reached NDCG scores between 0.848 for the LIME explanations of the LR model and 0.917 for the SHAP explanations of the RF. This comparison shows, that all of the models and explanation methods show strong similarity to the VBM ground truth with slight preferences of the ML models. This means that all of the methods rank features that are relevant in the VBM ground truth analysis high. The DL models show a larger variance in the NDCG scores in comparison to the classical ML models. Additionally, Fig. 7 shows that DL models ranked some of the regions that reached small relevance during the VBM analysis relatively high, e.g., the 4th ventricle or the optical chiasm. Those regions were combined with regions which achieved higher scores during the VBM analysis. One reason for this is, that DL models focus more on textures than on gray matter volumes.

Overall, it can be stated, that the most relevant regions of the DenseNet differed from those of the classical ML models. Reasons for this are that ML models focus on a smaller number of regions, as well as the DL methods focus on textural differences whereas the classical ML models focus on volumes. Both model types showed reasonable similarity to the VBM ground truth with the classical ML model explanations showing slightly higher similarity to the VBM results than the DL models.

5.7. Correlation of feature importance and model performance

In this section, the influence of the feature importance on the model performance is investigated. For this reason, the feature importances which were calculated in Section 5.6 were first normalized to a range

between 0 and 1 for all models. For each aspect, the Spearman rank correlation was calculated between the normalized feature importance and the five metrics which were calculated on the ADNI test set. The resulting correlation plot is visualized in Fig. 9.

The highest absolute correlation coefficient for the accuracy is −0.321 which was reached for the white matter hypointensities. This correlation coefficient can be interpreted as a low negative correlation. For the balanced accuracy, the highest absolute correlation of −0.300 was reached for the rostral anterior cingulate gyrus, which can be also interpreted as a low correlation. A negative correlation means, that an increase in feature importance leads to a decrease in model performance. For the AUROC, a moderate correlation was observed for a few regions, the highest absolute correlation which was −0.663 was reached for `aspect_6` which includes the left and right Cerebellum White Matter, the brain stem, as well as the left and right ventral Diencephalon. Similar to the aspects mentioned before, high feature relevances were associated with decreased performance. These regions are not in focus in most AD analyses. However, there is some previous research that identified changes in the morpho-functional mesoscopic traits [111] as well as amyloid plaques [112] in these regions. The higher correlation coefficients for the AUROC metric are caused by the fact that the ML models outperform the DL models for this metric. The F_1-Score shows low correlations for all features. As for the accuracy, the white matter hypointensities reached the highest absolute correlation coefficient of −0.325. The MCC shows a similar pattern and the white matter hypointensities achieved the highest absolute correlation of −0.325. There are only two aspects showing a positive correlation between the feature importance and the model performances, namely `aspect_27` and the temporal pole of the right hemisphere. `Aspect_27` includes regions within the middle temporal and inferior temporal cortices and thus corresponds with previous AD research [78–83] and is the most relevant region within the classical ML models.

Overall, no clear pattern was detected that shows a correlation between the feature importance and the model performances. Most of the correlations were low, moderate correlations were observed only for the AUROC metric. Combining those observations with the differences in feature importances of different models which were observed in Section 5.6, allows the conclusion that ML and DL models can detect AD based on different regions without one region outperforming the remaining ones. However, this conclusion needs to be handled with caution due to the relatively small number of data points in comparison to the high number of features.

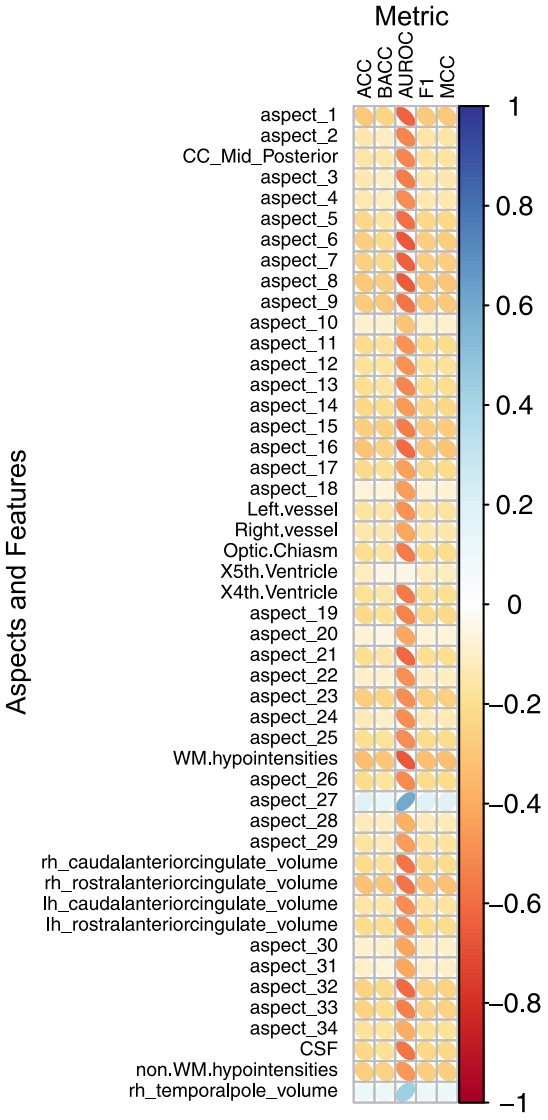


Fig. 9. Correlation plot showing the Spearman rank correlation between feature importances and model performances.

6. Discussion

This work aims to compare classical ML and DL models regarding the relevance of brain regions identified using interpretable ML. Seven classical ML models and six 3D CNNs were trained for the classification of CN vs. AD. An extensive hyperparameter tuning was applied using a 5-fold CV and the final models were calibrated using Platt scaling. For the classical ML models, three explanation methods were implemented, namely permutation-based importance, LIME, and SHAP. The DL models were analyzed using four explanation methods (GradCAM, GradCAM++, LIME, SHAP). Correlated volumes were consolidated using aspects to improve interpretability.

The results showed, that the metrics (accuracy, balanced accuracy, AUROC, F_1-Score, MCC) are similar for the classical ML and DL models. With an exception for the AUROC metric for which all ML models except the DT outperformed the DL models. Additionally, it was found that the classical ML models performed slightly better for 3 T MRI scans in comparison to 1.5 T scans. For the DL models the different models preferred different field strengths with slight preferences on 1.5 T scans. Both, the investigation of local explanations for individual subjects and the global summary of the entire training set showed the following results: The comparison between four explainability methods using activations summarized in brain structures for 3D DL models shows strong correlations. For classical ML models, the most important brain structures mainly correspond to previous research. For the DL models, high relevance scores were computed for a wider range of brain structures containing both, previously associated brain regions and brain structures which were not previously related to AD. Both approaches show a reasonable correlation to the VBM ground truth analysis with the ML models achieving slightly higher similarity. One reason for this might be, that the VBM analysis is based on GM concentration whereas the DL models were trained on voxel intensities to save time for intensive MRI pre-processing. A comparison to a deep learning model trained on GM concentration will be interesting for future work. The comparison of DL and ML feature rankings showed less similarity. This observation indicates that there are AD-related patterns in brain MRIs which cannot be represented by volumetric features.

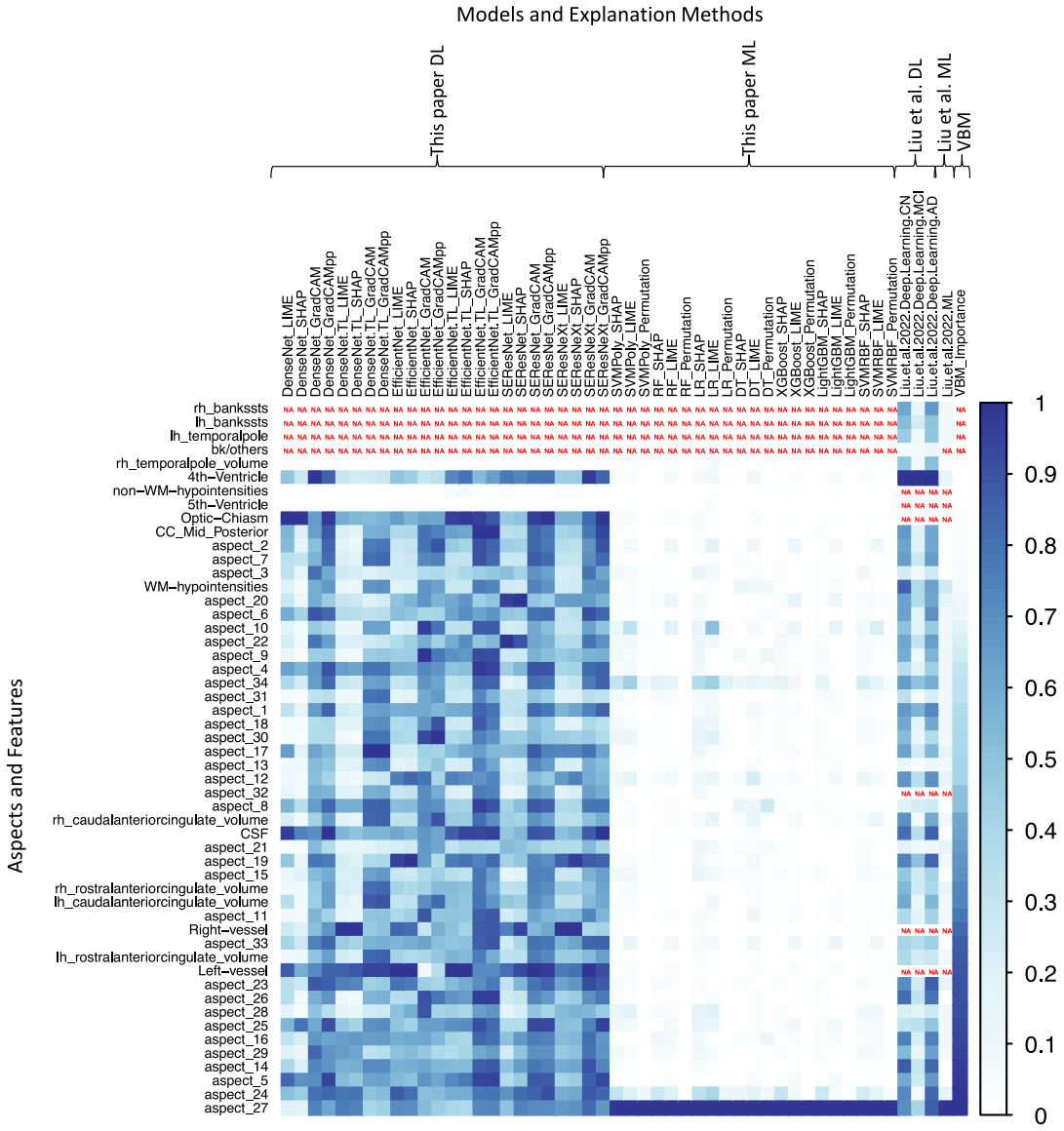


Fig. 10. Heatmap plot that compares the global, normalized feature importances in [56] with the feature importances of the DL and classical ML models of this paper using different explanation methods. As this paper consolidates correlated features in comparison to [56], only the highest ranking in [56] was used for each aspect. Regions that are not included in one of the papers are marked as NA.

As mentioned in Section 2 "Related Work", [56] compared the regional relevance of a DL model and a gradient boosting model to each other. Similar to this paper, the authors found that the DL model focused on a wider range of regions than the gradient boosting model, with some of the DL regions not previously associated with AD. A comparison of the feature importance scores computed in [56] and the explanation methods used in this work is visualized in Fig. 10. In comparison to [56], this work uses aspects that summarize correlated features. For this reason, during the comparison, only the highest ranking in [56] was used for each aspect. Additionally, regions that are not included in one of the papers are marked as NA.

The most relevant regions in the DL models implemented in [56] are the 4th ventricle (Ranks: CN: 1, MCI: 1, AD: 1), `aspect_14` (Ranks: CN: 2, AD: 7), `aspect_1` (Ranks: MCI: 2), and `aspect_27` (Ranks: AD: 2, MCI: 5). The 4th ventricle was also one of the most relevant regions of the DenseNet model which was trained from scratch and the SEResNeXt model in this work. `Aspect_14` includes the transverse temporal gyri of both hemispheres. Previous research [85] found reduced cortical thickness in the left hemisphere of this region. Additionally, `aspect_14` reached the fourth rank during the VBM analysis. In this research, `aspect_14` was not one of the main regions found for one of the DL models. However, this region reached the 5th rank within the GradCAM explanations of the SEResNeXt model. `Aspect_1` consolidates the posterior and anterior parts of the corpus callosum. Different parts of the corpus callosum have been also identified as being relevant for the classification of AD in some of the models presented

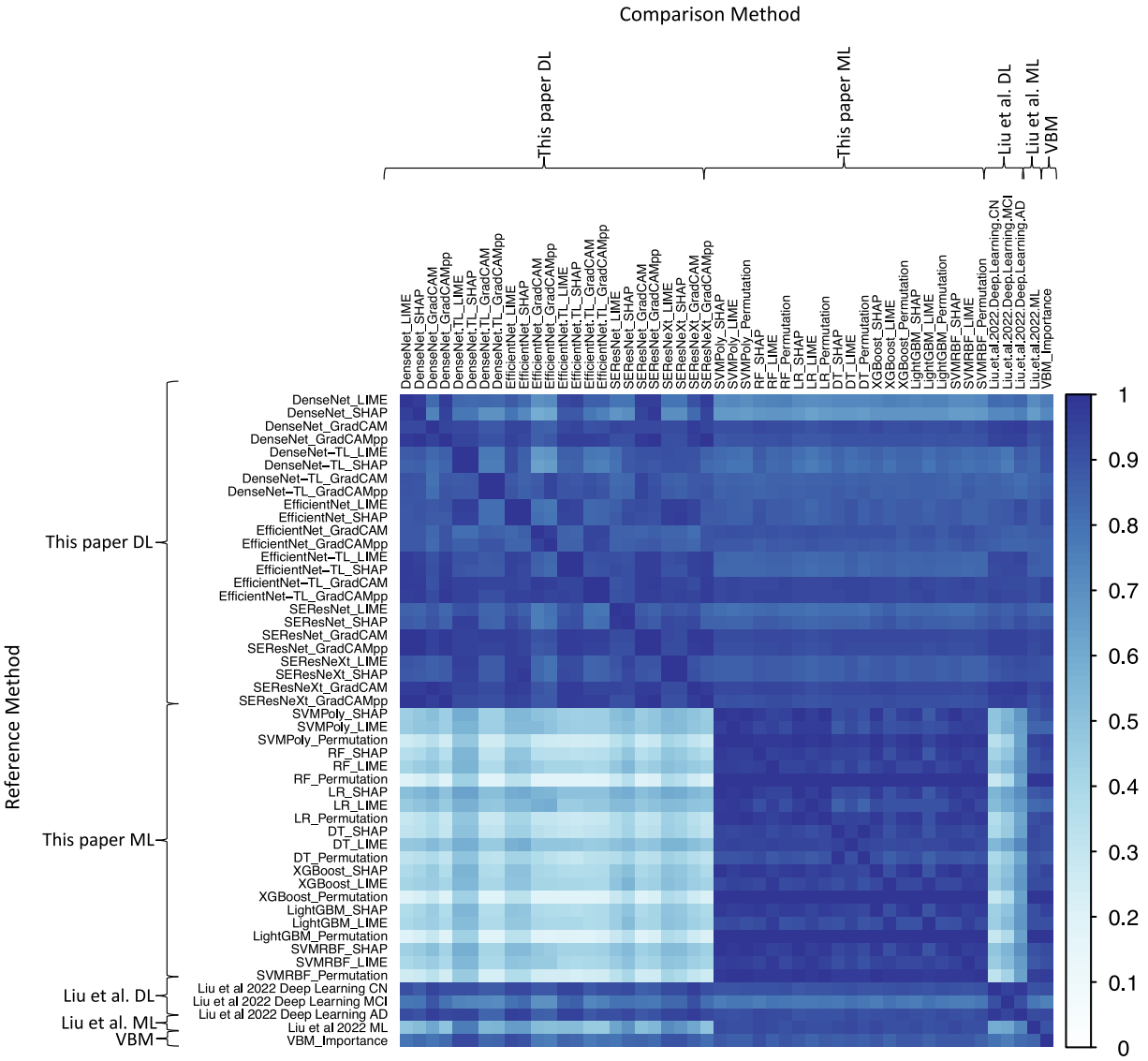


Fig. 11. Correlation plot showing the Normalized Discounted Cumulative Gain (NDCG) between global feature importances of the explainability methods and the feature importances documented in [56]. The correlation matrix is not symmetric because the NDCG is based on feature importance scores of the features visualized as rows (reference method) and rankings visualized as columns (comparison method). Changing the reference method changes the ground truth feature importance scores, resulting in asymmetries.

in this paper. Finally, `aspect_27` includes the inferior temporal, and the middle temporal gyri as well as the hippocampi and amygdala of both hemispheres. These regions are the most prominent ones in AD detection [78–83]. The DL models developed in this work were not mainly focused on `aspect_27`. However, as previously investigated in Section 5.6 these regions were nevertheless identified as relevant regions for many of the models.

Overall, the regional relevances of the models trained in this paper differed from those calculated in [56]. Nevertheless, both models also showed regions which were relevant in both approaches, e.g., the 4th ventricle, or the CSF region. Additionally, both approaches showed that DL models used a larger number of regions in comparison to the classical ML models. Some of the regions that were highly ranked in this work, such as the optical chiasm, the inferior lateral ventricle or the vessels, are excluded in [56] due to small numbers of voxels. Fig. 11 shows the similarity of the rankings computed in this paper and compares them to the rankings of [56] using the NDCG score. The matrix is built in the same way as Fig. 8, but the results presented in [56] are added. In comparison to [56], this work uses aspects that summarize correlated features. For this reason, during the comparison, only the highest ranking in [56] was used for each aspect. Additionally, regions that are not graded in one of the rankings are deleted pairwise. Fig. 11 shows a reasonable similarity between the DL methods of this paper and the rankings provided for the DL models in [56]. Using the DL models trained in this work as reference methods, the rankings reported for the DL model trained in [56] reached NDCG scores between 0.722 and 0.985. The smallest score was reached in the comparison of the SHAP explanations of the DenseNet which was trained from

scratch and the MCI ranking presented in [56]. The highest score was reached during the comparison of the GradCAM explanations of the SEResNeXt model and the AD ranking of the comparison paper. If the results from the comparison paper were used as the reference method, similar results were observed. The comparison of the MCI ranking of the comparison paper and the GradCAM++ explanations of the EfficientNet model which was pretrained from scratch reached the smallest NDCG score of 0.693. The highest NDCG score of 0.977 was achieved by comparing the AD ranking of the comparison paper with the GradCAM ranking computed for the SEResNeXt model. The comparison between the DL model of [56] and the ML models of this paper shows a similar correlation to the DL models trained in this work. In particular, when the feature importances of the classical ML models were used as the reference method, lower NDCG scores between 0.289 and 0.723 were achieved. The explanations calculated for the AD classification of the DL model showed higher similarity to the classical ML models of this paper than the explanations which were computed for the CN and MCI classes. In terms of classification performances, the DL model trained in [56] outperformed the gradient boosting classifier. This relationship was not observed in this paper. However, a fair comparison of the classification performances of the two approaches is not possible due to the different classification goals (three classes in [56] vs. two classes in this work) as well as different test sets.

6.1. Limitations

The approach proposed in this article features some limitations. First, the method requires the time-consuming segmentation of cortical and subcortical brain structures. This process can be accelerated, for example, by using the DL-based FastSurfer [113] pipeline instead of FreeSurfer. Another time-consuming factor is the calculation of LIME and SHAP values for the 3D DL models using superpixels. An alternative solution would be to use an implementation of DeepSHAP [17]. For this work, it was not used as some case studies showed pixel activations that faded out by summarizing large brain structures.

During the calculation of SHAP and LIME values for the DenseNet, the pixel intensities are substituted with the mean intensity in the subsequent brain structure. A more robust method is to replace the intensities with MRI intensities of different training subjects. However, this method requires high GPU memory consumption.

In addition, the use of superpixels during the calculation of SHAP and LIME could be improved by considering the brain region segmentation mask. In this work, it was decided to use superpixels in order to avoid biases, which can be introduced by varying sizes. At this point, a balance must be taken between similarly sized and shaped, as well as, biologically plausible regions.

Another limitation of this work regarding the summarization of the explanations in DL models is that correlated features were calculated for the volumes and applied to the 3D MRIs. This problem could be addressed by extracting correlations directly from the MRI intensities.

There are some limitations regarding the DL models and their training. First, the class imbalance was not considered during the training of the DL models. One possibility to solve this problem is, to use focal [114] instead of cross entropy loss. Another point is that using more recently developed DL architectures like Transformer-based models can be a potential factor to improve the performance of DL models, which can also influence the relevance of the brain structures. However, at the moment, VisionTransformer [115] is the only Transformer-based architecture implemented in the MONAI framework, which was used in this work. These models require large datasets to achieve reasonable results and were thus not implemented in this study.

In this work, transfer learning was implemented for the DenseNet and EfficientNet models based on the LDM-100k datatset. The results showed that the pretraining did not work well in the experiments as it did not lead to improved performances. There are several potential options for enhancing model performance during fine-tuning. These options comprise employing a real-world dataset instead of the artificially generated LDM-100k dataset for pre-training, utilizing a preprocessed version of the LDM-100k dataset, or initiating fine-tuning with frozen parameters for several epochs.

Additionally, this work only focuses on the clinically less relevant question of CN vs. AD classification. The reasons for this are the influence of the classification performances on the explanations, which can lead to less accurate explanations and the small number of MCI subjects in the external AIBL and OASIS datasets.

6.2. Future work

The results showed that partially different brain structures were activated for DL and classical ML models. It would be interesting to investigate the performance and explanations of an ensemble to exploit the patterns learned as well as the relevance of both model predictions. An additional comparison which is interesting in this context is the comparison to models trained on textural MRI features. As DL models internally compute textual representations of the MRI scans, it is assumed that textual features like Radiomics [116] potentially show a higher similarity in the brain regions compared to DL models.

The experiments can be expanded by increasing the number of interpretation methods to investigate their similarities and differences. For this reason, it is interesting to investigate interpretable ML methods like LRP or integrated gradients. Additionally, the comparison of regional feature relevance can be expanded by computing the correlations between heatmaps on a voxel level. This approach can potentially reveal a new perspective on the regional differences between models.

Validation on multiple datasets, e.g., the AD subset [49] of the HNR [50] or a subset of the National Alzheimer's Coordinating Center [117], could improve the external validity. This required a precise analysis of the inclusion and exclusion criteria of the datasets. Instead of diagnoses, it can be more promising to predict biomarkers which could also further improve the clinical relevance of the experiments.

Another relevant aspect is that this work compares the results of multiple explanation methods with VBM results and prior knowledge. An additional important step for future work is, to conduct a survey with physicians based on these studies to verify their usefulness in everyday clinical practice.

There are some new aspects of interest, which should be additionally considered in future work. The first aspect is the investigation of the impact of multimodal features. Here, approaches similar to [46, 118] could be examined using systematic explainability experiments. For example, it is interesting to combine MRI scans with genetic data, or sociodemographic risk factors to investigate the interactions of the modalities. Additionally, the use of multiple scans per subject taken at longitudinal visits can lead to further improvements by making the ML and DL models more robust. Another topic is that this work only investigates 3D DL models and compares them with classical ML models. However, most DL models are developed and pretrained for 2D images, which could potentially improve model performance for AD detection and thus also affects the interpretation models. For this reason, future work should compare the activated brain regions between 3D models and different implementations of slice-based 2D models.

7. Conclusion

In this research, a workflow was introduced to systematically compare the explanations of DL and classical ML in early AD detection based on 3D MRIs. For evaluation, seven classical ML models, namely RF, XGBoost, LightGBM, DT, LR, polynomial SVM, and radial SVM, were trained on an ADNI training set, validated for a hold-out ADNI test set, and externally validated for AIBL and OASIS. For the classical ML models, three explanation methods were applied (permutation

importance, LIME, SHAP). The resulting explanations were compared to four DL explanations (GradCAM, GradCAM++, LIME, SHAP) of six DL models.

The results show similar performances for DL in comparison to classical ML models. On an exemplary base, the activated regions of individual observations, which are clinically relevant, were examined and compared between the explanation methods and ML models. These explanations were systematically summarized for the entire training set for a systematic comparison of global regional relevance. The results showed that different explanation methods showed similar rankings for the same model. Overall, ML models focused on a small number of regions which were previously associated with AD. The explanations of the ML models showed high similarities across each other. DL models instead have high relevance scores for a larger number of brain regions. These include regions which were associated with AD in previous research, like the entorhinal gyri as well as regions not associated with AD before, like the optical chiasm. These observations are similar to those presented in [56]. The comparison to the VBM ground truth showed that ML and DL models both show reasonable similarities with ML models achieving slightly higher similarity scores. Nevertheless, the most relevant brain structures clearly differed between models. The most prominent regions like the hippocampus were more important for the classical ML models although these are also included within the DL models with varying relevance.

The developed workflow demonstrates one possibility to systematically explain the results of 3D DL models in medical contexts and compare the activated regions to the most important brain regions of classical ML models. The experiments showed different regions activated in both types of models. Nevertheless, both model types show reasonable similarity with the ground truth regions calculated via VBM.

CRediT authorship contribution statement

Louise Bloch: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Christoph M. Friedrich:** Writing – review & editing, Supervision, Resources, Conceptualization.

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.

Acknowledgments

The work of Louise Bloch was partially funded by a PhD grant from University of Applied Sciences and Arts Dortmund, Dortmund, Germany.

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. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.[16] As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found online.[17]

Consent for publication has been provided by ADNI administrators.

Human and animal rights

The ADNI study was approved by the institutional review boards of the participating institutions. All participants gave informed written consent. More details can be found online (https://adni.loni.usc.edu, Access: 2024-01-23). The AIBL study was approved by the institutional ethics committees of Austin Health, StVincent's Health, Hollywood Private Hospital and Edith Cowan University. All participants gave written informed consent before participating in the study. The OASIS study was approved by the institutional review boards of the participating institutions. All participants gave informed written consent.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108029.

References

[1] Alzheimer's Association, 2022 Alzheimer's disease facts and figures, Alzheimer's Dementia 18 (4) (2022) 700–789.

[2] J. Cao, J. Hou, J. Ping, D. Cai, Advances in developing novel therapeutic strategies for Alzheimer's disease, Molecul. Neurodegener. 13 (1) (2018).

[3] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation, Med. Image Anal. 63 (2020) 101694.

[4] L. Bloch, C.M. Friedrich, Developing a machine learning workflow to explain black-box models for Alzheimer's disease classification, in: C. Pesquita, A. Fred, H. Gamboa (Eds.), Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: HEALTHINF, HEALTHINF 2021, SciTePress, INSTICC, 2021, pp. 87–99, http://dx.doi.org/10.5220/0010211300870099.

[5] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[6] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD 2016, ACM, New York, NY, USA, 2016, pp. 785–794, http://dx.doi.org/10.1145/2939672.2939785.

[7] Y. LeCun, Y. Bengio, G.E. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[8] C. Molnar, Interpretable Machine Learning, 2019, URL https://christophm.github.io/interpretable-ml-book/.

[9] K. Borys, Y.A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C.M. Friedrich, F. Nensa, Explainable AI in medical imaging: An overview for clinical practitioners – beyond saliency-based XAI approaches, Eur. J. Radiol. 162 (2023) 110786.

[10] K. Borys, Y.A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C.M. Friedrich, F. Nensa, Explainable AI in medical imaging: An overview for clinical practitioners – saliency-based XAI approaches, Eur. J. Radiol. 162 (2023) 110787.

[16] ADNI: https://adni.loni.usc.edu, Access: 2024-01-23.

[17] ADNI acknowledgement list: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf, Access: 2024-01-23.

[11] L. Akter, Ferdib-Al-Islam, Dementia identification for diagnosing Alzheimer's disease using XGBoost algorithm, in: Proceedings of the International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021, 2021, pp. 205–209, http://dx.doi.org/10.1109/ICICT4SD50815.2021.9396777.

[12] L. Bloch, C.M. Friedrich, Classification of Alzheimer's disease using volumetric features of multiple MRI scans, in: Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2019, IEEE, 2019, pp. 2396–2401, http://dx.doi.org/10.1109/EMBC.2019.8857188.

[13] L. Bloch, C.M. Friedrich, Data analysis with Shapley values for automatic subject selection in Alzheimer's disease data sets using interpretable machine learning, Alzheimer's Res. Therapy 13 (1) (2021) 155.

[14] M.L. Leavitt, A.S. Morcos, Towards falsifiable interpretability research, 2020, http://dx.doi.org/10.48550/arXiv.2010.12016, arXiv:2010.12016v1.

[15] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI, ACM Comput. Surv. (2023).

[16] H. Zhang, J. Chen, H. Xue, Q. Zhang, Towards a unified evaluation of explanation methods without ground truth, 2019, http://dx.doi.org/10.48550/arXiv.1911.09017, arXiv:1911.09017v1.

[17] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), NIPS 2017, in: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017, pp. 4765–4774, URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[18] M.T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, 2016, pp. 1135–1144, http://dx.doi.org/10.1145/2939672.2939778.

[19] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, IEEE, 2017, pp. 618–626, http://dx.doi.org/10.1109/ICCV.2017.74.

[20] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV 2018, IEEE, 2018, pp. 839–847, http://dx.doi.org/10.1109/WACV.2018.00097.

[21] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, IEEE, 2017, pp. 2261–2269, http://dx.doi.org/10.1109/CVPR.2017.243.

[22] M. Tan, Q.V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: Proceedings of the International Conference on Machine Learning, ICML 2019, 97, PMLR, 2019, pp. 6105–6114, URL https://proceedings.mlr.press/v97/tan19a/tan19a.pdf.

[23] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, IEEE, 2017, pp. 7132–7141, http://dx.doi.org/10.1109/cvpr.2018.00745.

[24] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, in: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), in: Advances in Neural Information Processing Systems (NIPS 2017), vol. 30, Curran Associates, Inc., 2017, pp. 3146–3154, URL https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[25] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[26] G. Scafarto, N. Posocco, A. Bonnefoy, Calibrate to interpret, in: M.-R. Amini, S. Canu, A. Fischer, T. Guns, P. Kralj Novak, G. Tsoumakas (Eds.), Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases – Machine Learning and Knowledge Discovery in Databases, ECML–PKDD 2022, Springer International Publishing, Cham, 2023, pp. 340–355, http://dx.doi.org/10.1007/978-3-031-26387-3_21.

[27] J.C. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, in: Advances in Large Margin Classifiers, MIT Press, 1999, pp. 61–74.

[28] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C.A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, B. Bischl, General pitfalls of model-agnostic interpretation methods for machine learning models, in: A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, W. Samek (Eds.), Proceedings of the xxAI - beyond Explainable AI: International Workshop, Held in Conjunction with International Conference on Machine Learning 2020, Revised and Extended Papers, xxAI 2020, Springer International Publishing, Cham, 2022, pp. 39–68, http://dx.doi.org/10.1007/978-3-031-04083-2_4.

[29] K. Pekala, K. Woznica, P. Biecek, Triplot: Model agnostic measures and visualisations for variable importance in predictive models that take into account the hierarchical correlation structure, 2021, http://dx.doi.org/10.48550/arXiv.2104.03403, arXiv:2104.03403v1.

[30] R.C. Petersen, P.S. Aisen, L.A. Beckett, M.C. Donohue, A.C. Gamst, D.J. Harvey, C.R. Jack, W.J. Jagust, L.M. Shaw, A.W. Toga, J.Q. Trojanowski, M.W. Weiner, Alzheimer's disease neuroimaging initiative (ADNI), Neurology 74 (3) (2010) 201–209.

[31] K.A. Ellis, A.I. Bush, D. Darby, D. De Fazio, J. Foster, P. Hudson, N.T. Lautenschlager, N. Lenzo, R.N. Martins, P. Maruff, C. Masters, A. Milner, K. Pike, C. Rowe, G. Savage, C. Szoeke, K. Taddei, V. Villemagne, M. Woodward, D. Ames, AIBL Research Group, The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease, Int. Psychogeriatr. 21 (4) (2009) 672–687.

[32] P.J. LaMontagne, T.L.S. Benzinger, J.C. Morris, S. Keefe, R.C. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K.L. Moulder, A.G. Vlassenko, M.E. Raichle, C. Cruchaga, D.S. Marcus, OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease, medRxiv, 2019.

[33] L. Nanni, M. Interlenghi, S. Brahnam, C. Salvatore, S. Papa, R. Nemni, I. Castiglioni, The Alzheimer's Disease Neuroimaging Initiative, Comparison of transfer learning and conventional machine learning applied to structural brain MRI for the early diagnosis and prognosis of Alzheimer's disease, Front. Neurol. 11 (2020).

[34] T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi, A novel aggregate gene selection method for microarray data classification, Pattern Recognit. Lett. 60–61 (2015) 16–23.

[35] F. Lindgren, P. Geladi, S. Wold, The kernel algorithm for PLS, J. Chemometr. 7 (1) (1993) 45–59.

[36] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), in: Advances in Neural Information Processing Systems, vol. 25, Curran Associates, Inc., 2012, URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S.E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, IEEE, 2015, pp. 1–9, http://dx.doi.org/10.1109/cvpr.2015.7298594.

[38] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, IEEE, 2016, pp. 770–778, http://dx.doi.org/10.1109/cvpr.2016.90.

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, IEEE, 2016, pp. 2818–2826, http://dx.doi.org/10.1109/cvpr.2016.308.

[40] T.C. Hammond, X. Xing, C.K. Wang, D. Ma, K. Nho, P.K. Crane, F.M. Elahi, D.A. Ziegler, G. Liang, Q. Cheng, L.M. Yanckello, N. Jacobs, A.-L. Lin, β-amyloid and tau drive early Alzheimer's disease decline while glucose hypometabolism drives late decline, Commun. Biol. 3 (1) (2020).

[41] S.O. Danso, Z. Zeng, G. Muniz-Terrera, C.W. Ritchie, Developing an explainable machine learning-based personalised dementia risk prediction model: A transfer learning approach with ensemble learning algorithms, Front. Big Data 4 (2021) 613047.

[42] A. Börsch-Supan, M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, S. Zuber, Data resource profile: The survey of health, ageing and retirement in europe (SHARE), Int. J. Epidemiol. 42 (4) (2013) 992–1001.

[43] C.W. Ritchie, K. Ritchie, The PREVENT study: A prospective cohort study to identify mid-life biomarkers of late-onset Alzheimer's disease, BMJ Open 2 (6) (2012) e001893.

[44] G. Livingston, J.M. Huntley, A. Sommerlad, D. Ames, C. Ballard, S. Banerjee, C. Brayne, A. Burns, J. Cohen-Mansfield, C. Cooper, S.G. Costafreda, A. Dias, N.C. Fox, L.N. Gitlin, R. Howard, H.C. Kales, M. Kivimäki, E.B. Larson, A. Ogunniyi, V. Orgeta, K. Ritchie, K. Rockwood, E.L. Sampson, Q.M. Samus, L.S. Schneider, G. Selbæk, L. Teri, N. Mukadam, Dementia prevention, intervention, and care: 2020 report of the lancet commission, Lancet 396 (10248) (2020) 413–446.

[45] L. Bloch, C.M. Friedrich, for the Alzheimer's Disease Neuroimaging Initiative, Machine learning workflow to explain black-box models for early Alzheimer's disease classification evaluated for multiple datasets, SN Comput. Sci. 3 (6) (2022) 509.

[46] O. Pelka, C.M. Friedrich, F. Nensa, C. Mönninghoff, L. Bloch, K.-H. Jöckel, S. Schramm, S. Sanchez Hoffmann, A. Winkler, C. Weimar, M. Jokisch, for the Alzheimer's Disease Neuroimaging Initiative, Sociodemographic data and APOE-ϵ4 augmentation for MRI-based detection of amnestic mild cognitive impairment using deep learning systems, PLoS One 15 (9) (2020) e0236868.

[47] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[48] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.

[49] M. Dlugaj, C. Weimar, N. Wege, P. Verde, M. Gerwig, N. Dragano, S. Moebus, K.-H. Jöckel, R. Erbel, J. Siegrist, Prevalence of mild cognitive impairment and its subtypes in the Heinz Nixdorf RECALL study cohort, Dementia Geriatric Cogn. Disord. 30 (4) (2010) 362–373.

[50] A. Schmermund, S. Möhlenkamp, Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: Rationale and design of the Heinz Nixdorf RECALL study, Am. Heart J. 144 (2) (2002) 212–218.

[51] M. Dyrba, M. Hanzig, S. Altenstein, S. Bader, T. Ballarini, F. Brosseron, K. Buerger, D. Cantré, P. Dechent, L. Dobisch, E. Düzel, M. Ewers, K. Fliessbach, W. Glanz, J.-D. Haynes, M.T. Heneka, D. Janowitz, D.B. Keles, I. Kilimann, C. Laske, F. Maier, C.D. Metzger, M.H. Munk, R. Perneczky, O. Peters, L. Preis, J. Priller, B. Rauchmann, N. Roy, K. Scheffler, A. Schneider, B.H. Schott, A. Spottke, E.J. Spruth, M.-A. Weber, B. Ertl-Wagner, M. Wagner, J. Wiltfang, F. Jessen, S.J. Teipel, Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease, Alzheimer's Res. Therapy 13 (1) (2021).

[52] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One 10 (7) (2015) e0130140.

[53] F. Jessen, A. Spottke, H. Boecker, F. Brosseron, K. Buerger, C. Catak, K. Fliessbach, C. Franke, M. Fuentes, M.T. Heneka, D. Janowitz, I. Kilimann, C. Laske, F. Menne, P.J. Nestor, O. Peters, J. Priller, V. Pross, A. Ramirez, A. Schneider, O. Speck, E.J. Spruth, S.J. Teipel, R. Vukovich, C. Westerteicher, J. Wiltfang, S. Wolfsgruber, M. Wagner, E. Düzel, Design and first baseline data of the DZNE multicenter observational study on predementia Alzheimer's disease (DELCODE), Alzheimer's Res. Therapy 10 (1) (2018).

[54] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: D. Precup, Y.W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, ICML 2017, in: Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 3319–3328, URL http://proceedings.mlr.press/v70/sundararajan17a/sundararajan17a.pdf.

[55] D. Wang, N. Honnorat, P.T. Fox, K. Ritter, S.B. Eickhoff, S. Seshadri, M. Habes, Deep neural network heatmaps capture Alzheimer's disease patterns reported in a large meta-analysis of neuroimaging studies, NeuroImage 269 (2023) 119929.

[56] S. Liu, A.V. Masurkar, H. Rusinek, J. Chen, B. Zhang, W. Zhu, C. Fernandez-Granda, N. Razavian, Generalizable deep learning model for early Alzheimer's disease detection from structural MRIs, Sci. Rep. 12 (1) (2022) 17106.

[57] A. Lukasová, Hierarchical agglomerative clustering procedure, Pattern Recognit. 11 (5) (1979) 365–381.

[58] L.S. Shapley, A value for n-person games, in: H.W. Kuhn, A.W. Tucker (Eds.), in: Contributions To the Theory of Games, vol. 2, (28) Princeton University Press, Princeton, US, 1953, pp. 307–318, http://dx.doi.org/10.1515/9781400881970-018.

[59] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2) (2020) 336–359.

[60] A.F. Agarap, Deep learning using rectified linear units (ReLU), 2018, http://dx.doi.org/10.48550/arXiv.1803.08375, arXiv:1803.08375v2.

[61] J. Ashburner, K.J. Friston, Voxel-based morphometry - the methods, NeuroImage 11 (6) (2000) 805–821.

[62] B. Fischl, FreeSurfer, NeuroImage 62 (2) (2012) 774–781.

[63] G. Van Rossum, F.L. Drake, Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009, URL https://www.python.org/.

[64] R.S. Desikan, F. Ségonne, B. Fischl, B.T. Quinn, B.C. Dickerson, D. Blacker, R.L. Buckner, A.M. Dale, R.P. Maguire, B.T. Hyman, M.S. Albert, R.J. Killiany, An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest, NeuroImage 31 (3) (2006) 968–980.

[65] B. Fischl, D.H. Salat, E. Busa, M.S. Albert, M.E. Dieterich, C. Haselgrove, A. van der Kouwe, R.J. Killiany, D.N. Kennedy, S. Klaveness, A. Montillo, N. Makris, B.R. Rosen, A.M. Dale, Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain, Neuron 33 (3) (2002) 341–355.

[66] E. Westman, C.A. Aguilar, J.-S. Muehlboeck, A. Simmons, Regional magnetic resonance imaging measures for multivariate analysis in Alzheimer's disease and mild cognitive impairment, Brain Topogr. 26 (1) (2012) 9–23.

[67] A. Evans, D. Collins, S. Mills, E. Brown, R. Kelly, T. Peters, 3D statistical neuroanatomical models from 305 MRI volumes, in: Proceedings of the Conference Record Nuclear Science Symposium and Medical Imaging Conference, NSS MIC 1993, vol. 3, IEEE, 1993, pp. 1813–1817, http://dx.doi.org/10.1109/NSSMIC.1993.373602.

[68] P. Refaeilzadeh, L. Tang, H. Liu, Cross-validation, in: L. Liu, M.T. Özsu (Eds.), Encyclopedia of Database Systems, Springer, Boston, MA, USA, 2009, pp. 532–538, http://dx.doi.org/10.1007/978-0-387-39940-9_565.

[69] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[70] S. Xie, R.B. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, IEEE, 2017, pp. 5987–5995, http://dx.doi.org/10.1109/cvpr.2017.634.

[71] The MONAI Consortium, Project MONAI, 2020, http://dx.doi.org/10.5281/zenodo.4323059.

[72] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems (NeurIPS 2019), Curran Associates, Inc., 2019, pp. 8024–8035, URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[73] D. Merkel, Docker: Lightweight Linux containers for consistent development and deployment, Linux J. 2014 (239) (2014) 2.

[74] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.

[75] W.H.L. Pinaya, P.-D. Tudosiu, J. Dafflon, P.F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, M.J. Cardoso, Brain imaging generation with latent diffusion models, in: A. Mukhopadhyay, I. Oksuz, S. Engelhardt, D. Zhu, Y. Yuan (Eds.), Proceedings of the Workshop on Deep Generative Models Co-Located with the International Conference on Medical Image Computing and Computer-Assisted Intervention, DGM 2022, Springer Nature Switzerland, Cham, 2022, pp. 117–126, http://dx.doi.org/10.1007/978-3-031-18576-2_12.

[76] H. Baniecki, W. Kretowicz, P. Piatyszek, J. Wisniewski, P. Biecek, dalex: Responsible machine learning with interactive explainability and fairness in Python, J. Mach. Learn. Res. 22 (214) (2021) 1–7.

[77] R. Achanta, A. Shaji, K.W. Smith, A. Lucchi, P. Fua, S. Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.

[78] A.L. Foundas, C.M. Leonard, S.M. Mahoney, O.F. Agee, K.M. Heilman, Atrophy of the hippocampus, parietal cortex, and insula in Alzheimer's disease: A volumetric magnetic resonance imaging study, Neuropsychiatr. Neuropsycholo. Behav. Neurol. 10 (2) (1997) 81–89.

[79] G.B. Frisoni, N.C. Fox, C.R. Jack, P. Scheltens, P.M. Thompson, The clinical use of structural MRI in Alzheimer disease, Nat. Rev. Neurol. 6 (2) (2010) 67–77.

[80] S.G. Mueller, N. Schuff, K. Yaffe, C. Madison, B.L. Miller, M.W. Weiner, Hippocampal atrophy patterns in mild cognitive impairment and Alzheimer's disease, Hum. Brain Mapp. 31 (9) (2010) 1339–1347.

[81] S.P. Poulin, R. Dautoff, J.C. Morris, L.F. Barrett, B.C. Dickerson, Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity, Psychiatr. Res. Neuroimag. 194 (1) (2011) 7–13.

[82] S.W. Scheff, D.A. Price, F.A. Schmitt, M.A. Scheff, E.J. Mufson, Synaptic loss in the inferior temporal gyrus in mild cognitive impairment and Alzheimer's disease, J. Alzheimer's Dis. 24 (3) (2011) 547–557.

[83] H. Yang, H. Xu, Q. Li, Y. Jin, W. Jiang, J. Wang, Y. Wu, W. Li, C. Yang, X. Li, S. Xiao, F. Shi, T. Wang, Study of brain morphology change in Alzheimer's disease and amnestic mild cognitive impairment compared with normal controls, General Psychiatr. 32 (2) (2019).

[84] P.M. Thompson, K.M. Hayashi, G.I. de Zubicaray, A.L. Janke, S.E. Rose, J. Semple, M.S. Hong, D.H. Herman, D. Gravano, D.M. Doddrell, A.W. Toga, Mapping hippocampal and ventricular change in Alzheimer disease, NeuroImage 22 (4) (2004) 1754–1766.

[85] H. Yang, H. Xu, Q. Li, Y. Jin, W. Jiang, J. Wang, Y. Wu, W. Li, C. Yang, X. Li, S. Xiao, F. Shi, T. Wang, Study of brain morphology change in Alzheimer's disease and amnestic mild cognitive impairment compared with normal controls, General Psychiatr. 32 (2) (2019).

[86] L. de Gois Vasconcelos, A.P. Jackowski, M.O. de Oliveira, Y.M.R. Flor, A.A.L. Souza, O.F.A. Bueno, S.M.D. Brucki, The thickness of posterior cortical areas is related to executive dysfunction in Alzheimer's disease, Clinics 69 (1) (2014) 28–37.

[87] G. Karas, P. Scheltens, S. Rombouts, R. van Schijndel, M. Klein, B. Jones, W. van der Flier, H. Vrenken, F. Barkhof, Precuneus atrophy in early-onset Alzheimer's disease: A morphometric structural MRI study, Neuroradiology 49 (12) (2007) 967–976.

[88] H.I.L. Jacobs, D.A. Hopkins, H.C. Mayrhofer, E. Bruner, F.W. van Leeuwen, W. Raaijmakers, J.D. Schmahmann, The cerebellum in Alzheimer's disease: Evaluating its role in cognitive decline, Brain 141 (1) (2017) 37–47.

[89] E. Hoxha, P. Lippiello, F. Zurlo, I. Balbo, R. Santamaria, F. Tempia, M.C. Miniaci, The emerging role of altered cerebellar synaptic processing in Alzheimer's disease, Front. Aging Neurosci. 10 (2018).
[90] D.R. Hinton, A.A. Sadun, J.C. Blanks, C.A. Miller, Optic-nerve degeneration in Alzheimer's disease, N. Engl. J. Med. 315 (8) (1986) 485–487.
[91] C. Nishioka, C. Poh, S.-W. Sun, Diffusion tensor imaging reveals visual pathway damage in patients with mild cognitive impairment and Alzheimer's disease, J. Alzheimer's Dis. 45 (2015) 97–107, 1.
[92] A. Biegon, J. Eberling, B. Richardson, M. Roos, S. Wong, B. Reed, W. Jagust, Human corpus callosum in aging and Alzheimer's disease: A magnetic resonance imaging study, Neurobiol. Aging 15 (4) (1994) 393–397.
[93] O. Narkiewicz, M.J. de Leon, A. Convit, A.E. George, J. Wegiel, J. Morys, M. Bobinski, J. Golomb, D.C. Miller, H.M. Wisniewski, Dilatation of the lateral part of the transverse fissure of the brain in Alzheimer's disease, Acta Neurobiol. Exper. 53 (3) (1993) 457–465.
[94] A.L. Powell, R.S. Mezrich, A.C. Coyne, A. Loesberg, I. Keller, Convex third ventricle: A possible sign for dementia using MRI, J. Geriatric Psychiatr. Neurol. 6 (4) (1993) 217–221.
[95] L.A. van de Mortel, R.M. Thomas, G.A. van Wingen, Grey matter loss at different stages of cognitive decline: A role for the thalamus in developing Alzheimer's disease, J. Alzheimer's Dis. 83 (2021) 705–720.
[96] E. Pardilla-Delgado, H. Torrico-Teave, J.S. Sanchez, L.A. Ramirez-Gomez, A. Baena, Y. Bocanegra, C. Vila-Castelar, J.T. Fox-Fuller, E. Guzmán-Vélez, J.R. Martínez, S. Alvarez, M. Ochoa-Escudero, F. Lopera, Y.T. Quiroz, Associations between subregional thalamic volume and brain pathology in autosomal dominant Alzheimer's disease, Brain Commun. 3 (2) (2021).
[97] S. Lehéricy, E.C. Hirsch, L.B. Hersh, Y. Agid, Cholinergic neuronal loss in the globus pallidus of Alzheimer disease patients, Neurosci. Lett. 123 (2) (1991) 152–155.
[98] L.W. de Jong, K. van der Hiele, I.M. Veer, J.J. Houwing, R.G.J. Westendorp, E.L.E.M. Bollen, P.W. de Bruin, H.A.M. Middelkoop, M.A. van Buchem, J. van der Grond, Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: An MRI study, Brain 131 (12) (2008) 3277–3285.
[99] D. Mascali, M. DiNuzzo, L. Serra, S. Mangia, B. Maraviglia, M. Bozzali, F. Giove, Disruption of semantic network in mild Alzheimer's disease revealed by resting-state fMRI, Neuroscience 371 (2018) 38–48.
[100] Q. Yuan, X. Liang, C. Xue, W. Qi, S. Chen, Y. Song, H. Wu, X. Zhang, C. Xiao, J. Chen, Altered anterior cingulate cortex subregional connectivity associated with cognitions for distinguishing the spectrum of pre-clinical Alzheimer's disease, Front. Aging Neurosci. 14 (2022).
[101] B.F. Jones, J. Barnes, H.B. Uylings, N.C. Fox, C. Frost, M.P. Witter, P. Scheltens, Differential regional atrophy of the cingulate gyrus in Alzheimer disease: A volumetric MRI study, Cerebral Cortex 16 (12) (2006) 1701–1708.
[102] J.L. Whitwell, Progression of atrophy in Alzheimer's disease and related disorders, Neurotox. Res. 18 (3) (2010) 339–346.
[103] Z. Yao, Y. Zhang, L. Lin, Y. Zhou, C. Xu, T. Jiang, Abnormal cortical networks in mild cognitive impairment and Alzheimer's disease, PLoS Comput. Biol. 6 (11) (2010) e1001006.
[104] X. Nie, Y. Sun, S. Wan, H. Zhao, R. Liu, X. Li, S. Wu, Z. Nedelska, J. Hort, Z. Qing, Y. Xu, B. Zhang, Subregional structural alterations in hippocampus and nucleus accumbens correlate with the clinical impairment in patients with Alzheimer's disease clinical spectrum: Parallel combining volume and vertex-based approach, Front. Neurol. 8 (2017) 399.
[105] L. deToledo Morrell, T.R. Stoub, M. Bulgakova, R.S. Wilson, D.A. Bennett, S. Leurgans, J. Wuu, D.A. Turner, MRI-derived entorhinal volume is a good predictor of conversion from MCI to AD, Neurobiol. Aging 25 (9) (2004) 1197–1203.
[106] C. Gaser, R. Dahnke, P.M. Thompson, F. Kurth, E. Luders, CAT - a computational anatomy toolbox for the analysis of structural MRI data, bioRxiv, 2022.
[107] K.J. Friston, J.T. Ashburner, S.J. Kiebel, T.E. Nichols, W.D. Penny, Statistical parametric mapping: The analysis of functional brain images, first ed., Elsevier, 2007.
[108] R. Vallat, Pingouin: Statistics in Python, J. Open Source Softw. 3 (31) (2018) 1026.
[109] C. Bonferroni, Teoria Statistica Delle Classi E Calcolo Delle Probabilità, in: Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze, Seeber, 1936.
[110] S. Seabold, J. Perktold, Statsmodels: Econometric and statistical modeling with python, in: Proceedings of the 9th Python in Science Conference, SCIPY 2010, 2010, pp. 92–96, URL https://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf.
[111] M. Inglese, N. Patel, K. Linton-Reid, F. Loreto, Z. Win, R.J. Perry, C. Carswell, M. Grech-Sollars, W.R. Crum, H. Lu, P.A. Malhotra, L.C. Silbert, B. Lind, R. Crissey, J.A. Kaye, R. Carter, S. Dolen, J. Quinn, L.S. Schneider, S. Pawluczyk, M. Becerra, L. Teodoro, K. Dagerman, B.M. Spann, J. Brewer, H. Vanderswag, A. Fleisher, J. Ziolkowski, J.L. Heidebrink, Zbizek-Nulph, J.L. Lord, L. Zbizek-Nulph, R. Petersen, S.S. Mason, C.S. Albers, D. Knopman, K. Johnson, J. Villanueva-Meyer, V. Pavlik, N. Pacini, A. Lamb, J.S. Kass, R.S. Doody, V. Shibley, M. Chowdhury, S. Rountree, M. Dang, Y. Stern, L.S. Honig, A. Mintz, B. Ances, J.C. Morris, D. Winkfield, M. Carroll, G. Stobbs-Cucchi, A. Oliver, M.L. Creech, M.A. Mintun, S. Schneider, D. Geldmacher, M.N. Love, R. Griffith, D. Clark, J. Brockington, D. Marson, H. Grossman, M.A. Goldstein, J. Greenberg, E. Mitsis, R.C. Shah, M. Lamar, A. Sood, K.S. Blanchard, D. Fleischman, K. Arfanakis, P. Samuels, R. Duara, M.T. Greig-Custo, R. Rodriguez, M. Albert, D. Varon, C. Onyike, L. Farrington, S. Rudow, R. Brichko, M.T. Greig, S. Kielb, A. Smith, B.A. Raj, K. Fargher, M. Sadowski, T. Wisniewski, M. Shulman, A. Faustin, J. Rao, K.M. Castro, A. Ulysse, S. Chen, M.O. Sheikh, J. Singleton-Garvin, P.M. Doraiswamy, J.R. Petrella, O. James, T.Z. Wong, S. Borges-Neto, J.H. Karlawish, D.A. Wolk, S. Vaishnavi, C.M. Clark, S.E. Arnold, C.D. Smith, G.A. Jicha, R. El Khouli, F.D. Raslau, O.L. Lopez, M. Zmuda, M. Butters, M. Oakley, D.M. Simpson, A.P. Porsteinsson, K. Martin, N. Kowalski, K.S. Martin, M. Keltz, B.S. Goldstein, K.M. Makino, M.S. Ismail, C. Brand, C. Reist, G. Thai, A. Pierce, B. Yanez, E. Sosa, M. Witbracht, B. Kelley, T. Nguyen, K. Womack, D. Mathews, M. Quiceno, A.I. Levey, J.J. Lah, I. Hajjar, J.S. Cellar, J.M. Burns, R.H. Swerdlow, W.M. Brooks, D.H.S. Silverman, S. Kremen, L. Apostolova, K. Tingus, P.H. Lu, G. Bartzokis, E. Woo, E. Teng, N.R. Graff-Radford, F. Parfitt, K. Poki-Walker, M.R. Farlow, A.M. Hake, B.R. Matthews, J.R. Brosch, S. Herring, C.H. van Dyck, A.P. Mecca, S.P. Good, M.G. MacAvoy, R.E. Carson, P. Varma, H. Chertkow, S. Vaitekunis, C. Hosein, S. Black, B. Stefanovic, C.C. Heyn, G.-Y.R. Hsiung, E. Kim, B. Mudge, V. Sossi, H. Feldman, M. Assaly, E. Finger, S. Pasternak, I. Rachinsky, A. Kertesz, D. Drost, J. Rogers, I. Grant, B. Muse, E. Rogalski, J.R.M.M. Mesulam, D. Kerwin, C.-K. Wu, N. Johnson, K. Lipowski, S. Weintraub, B. Bonakdarpour, N. Pomara, R. Hernando, A. Sarrael, H.J. Rosen, S. Mackin, C. Nelson, D. Bickford, Y.H. Au, K. Scherer, D. Catalinotto, S. Stark, E. Ong, D. Fernandez, B.L. Miller, H. Rosen, D. Perry, R.S. Turner, K. Johnson, B. Reynolds, K. MCCann, J. Poe, R.A. Sperling, K.A. Johnson, G.A. Marshall, J. Yesavage, J.L. Taylor, S. Chao, J. Coleman, J.D. White, B. Lane, A. Rosen, J. Tinklenberg, C.M. Belden, A. Atri, K.A.C.E. Zamrini, M. Sabbagh, R. Killiany, R. Stern, J. Mez, N. Kowall, A.E. Budson, T.O. Obisesan, O.E. Ntekim, S. Wolday, J.I. Khan, E. Nwulia, S. Nadarajah, A. Lerner, P. Ogrocki, C. Tatsuoka, P. Fatica, E. Fletcher, P. Maillard, J. Olichney, C. DeCarli, O. Carmichael, V. Bates, H. Capote, M. Rainka, M. Borrie, T.-Y. Lee, R. Bartha, S. Johnson, S. Asthana, C.M. Carlsson, A. Perrin, A. Burke, D.W. Scharre, M. Kataki, R. Tarawneh, D. Hart, E.A. Zimmerman, D. Celmins, D.D. Miller, L.L.B. Ponto, K.E. Smith, H. Koleva, H. Shim, K.W. Nam, S.K. Schultz, J.D. Williamson, S. Craft, J. Cleveland, M. Yang, K.M. Sink, B.R. Ott, J. Drake, G. Tremont, L.A. Daiello, J.D. Drake, A. Ritter, C. Bernick, D. Munic, A. O'Connelll, J. Mintzer, A. Wiliams, J. Masdeu, J. Shi, A. Garcia, the Alzheimer's Disease Neuroimaging Initiative, A predictive model using the mesoscopic architecture of the living brain to detect Alzheimer's disease, Commun. Med. 2 (1) (2022) 70.
[112] R.D. Rudelli, M.W. Ambler, H.M. Wisniewski, Morphology and distribution of Alzheimer neuritic (senile) and amyloid plaques in striatum and diencephalon, Acta Neuropathol. 64 (4) (1984) 273–281.
[113] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, M. Reuter, FastSurfer - a fast and accurate deep learning based neuroimaging pipeline, NeuroImage 219 (2020) 117012.
[114] T.-Y. Lin, P. Goyal, R.B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2017) 318–327.
[115] A. Dosovitskiy, L. Beyer, A.I. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: Proceedings of the International Conference on Learning Representations, ICLR 2021, 2021, URL https://openreview.net/pdf?id=YicbFdNTTy.
[116] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R.G. van Stiphout, P. Granton, C.M. Zegers, R. Gillies, R. Boellard, A. Dekker, H.J. Aerts, Radiomics: Extracting more information from medical images using advanced feature analysis, Eur. J. Cancer 48 (4) (2012) 441–446.
[117] D.L. Beekly, E.M. Ramos, G. van Belle, W.D. Deitrich, A.D. Clark, M.E. Jacka, W.A. Kukull, The national Alzheimer's coordinating center (NACC) database: An Alzheimer disease database, Alzheimer Dis. Assoc. Disord. 18 (4) (2004) 270–277.
[118] T.N. Wolf, S. Pölsterl, C. Wachinger, DAFT: A universal module to interweave tabular data and 3D images in CNNs, NeuroImage 260 (2022) 119505.