npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01123-7

Constructing personalized characterizations of structural brain aberrations in patients with dementia using explainable artificial intelligence

Esten H. Leonardsen[1,2] ✉, Karin Persson[3,4], Edvard Grødem[1,5], Nicola Dinsdale[6], Till Schellhorn[7,8], James M. Roe[1], Didac Vidal-Piñeiro[1], Øystein Sørensen[1], Tobias Kaufmann[2,9,10], Eric Westman[11], Andre Marquand[12], Geir Selbæk[3,4], Ole A. Andreassen[2,13], Thomas Wolfers[1,2,9,10,14], Lars T. Westlye[1,2,13,14] & Yunpeng Wang[1,14]

Deep learning approaches for clinical predictions based on magnetic resonance imaging data have shown great promise as a translational technology for diagnosis and prognosis in neurological disorders, but its clinical impact has been limited. This is partially attributed to the opaqueness of deep learning models, causing insufficient understanding of what underlies their decisions. To overcome this, we trained convolutional neural networks on structural brain scans to differentiate dementia patients from healthy controls, and applied layerwise relevance propagation to procure individual-level explanations of the model predictions. Through extensive validations we demonstrate that deviations recognized by the model corroborate existing knowledge of structural brain aberrations in dementia. By employing the explainable dementia classifier in a longitudinal dataset of patients with mild cognitive impairment, we show that the spatially rich explanations complement the model prediction when forecasting transition to dementia and help characterize the biological manifestation of disease in the individual brain. Overall, our work exemplifies the clinical potential of explainable artificial intelligence in precision medicine.

Since its invention in the 1970s, magnetic resonance imaging (MRI) has provided an opportunity to non-invasively examine the inside of the body. In neuroscience, images acquired with MRI scanners have been used to identify how the brains of patients with various neurological disorders differ from their healthy counterparts. Stereotypically, this has been done by collecting data from a group of patients with a given disorder and a comparable group of healthy controls, on which traditional statistical inference is applied to identify spatial locations of the brain where the groups differ[1]. Typically, these locations are not atomic locations identified by spatial coordinates, but rather morphological regions defined by an atlas, derived from empirical or theoretical insights of how the brain is structured. Differences between groups are described using morphometric properties like thickness or volume of these prespecified regions. A major benefit of this approach is the innate interpretability of the results: on average, patients

[1]Department of Psychology, University of Oslo, Oslo, Norway. [2]Norwegian Centre for Mental Disorders Research (NORMENT), Oslo University Hospital & Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [3]The Norwegian National Centre for Ageing and Health, Vestfold Hospital Trust, Tønsberg, Norway. [4]Department of Geriatric Medicine, Oslo University Hospital, Oslo, Norway. [5]Computational Radiology & Artificial Intelligence (CRAI) Unit, Division of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway. [6]Oxford Machine Learning in NeuroImaging (OMNI) Lab, University of Oxford, Oxford, UK. [7]Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [8]Division of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway. [9]Department of Psychiatry and Psychotherapy, Tübingen Center for Mental Health, University of Tübingen, Tübingen, Germany. [10]German Center for Mental Health (DZPG), Munich, Germany. [11]Division of Clinical Geriatrics, Department of Neurobiology, Care Sciences, and Society, Karolinska Institutet, Stockholm, Sweden. [12]Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Centre, Nijmegen, the Netherlands. [13]KG Jebsen Center for Neurodevelopmental Disorders, University of Oslo, Oslo, Norway. [14]These authors contributed equally: Thomas Wolfers, Lars T. Westlye, Yunpeng Wang. ✉e-mail: estenhl@uio.no

with a given disorder deviate in a specific region of the brain in a comprehensible manner. Furthermore, the high degree of localization offered by modern brain scans allows for accurate characterization of where and how the brain of an individual deviates from an expected, typically healthy, norm[2]. However, the effects which are found are typically small[3] with limited predictive power at the individual level[4,5], which in turn has raised questions about whether these analytical methods are expressive enough to model complex mental or clinical phenomena[6]. As an alternative, new conceptual approaches are proposed, advocating modeling frameworks with increased expressive power that allow for group differences through complex, non-linear interactions between multiple, potentially distant, parts of the brain[7], with a focus on prediction[8]. Such modeling flexibility is naturally achieved with artificial neural networks (ANNs), a class of statistical learning methods that combines aspects of data at multiple levels of abstraction, to accurately solve a predictive task[9]. However, while this often yields high predictive performance, e.g., by demonstrating clinically sufficient case-control classification accuracy for certain conditions, it comes at the cost of interpretation, as the models employ decision rules not trivially understandable by humans[10]. When the goal of the analysis is clinical, supporting the diagnosis and treatment of someone affected by a potential disorder, this opaqueness presents a substantial limitation. Thus, development and empirical validation of new methods within clinical neuroimaging that combine predictive efficacy with individual-level interpretability is imperative, to facilitate trust in how the system is working, and to accurately describe inter-individual heterogeneity.

With more than 55 million individuals afflicted worldwide[11], over 25 million disability-adjusted life years lost[12,13] and a cost exceeding one trillion USD yearly[14], dementia is a prime example of a neurological disorders that incur a monumental global burden. Due to the global aging population the prevalence is expected to nearly triple by 2050[15], inciting a demand for technological solutions to facilitate handling the upcoming surge of patients. Dementia is a complex and progressive clinical condition[16] with multiple causal determinants and moderators. Alzheimer's disease (AD) is the most common form and accounts for 60–80% of all cases[11]. However, the brain pathologies underlying different subtypes of dementia are not disjoint, but often co-occur[17–19], and have neuropathological commonalities[20]. The most prominent is neurodegeneration, occurring in both specific regions like the hippocampus, and globally across the brain[21], and inter-individual variations in the localization of atrophy has been associated with impairments in specific cognitive domains[22,23]. Thus, the biological manifestation of dementia in the brain is heterogeneous[24], resulting in distinctive cognitive and functional deficits[20], highlighting the need for precise and personalized approaches to diagnosis. For patients with mild cognitive impairment (MCI), a potential clinical precursor to dementia, providing individualized characterizations of the underlying etiological disease at an early stage could widen the window for early interventions[25], alleviate uncertainty about the condition, and help with planning for the future[26].

In dementia, ANNs, and particularly convolutional neural networks (CNNs), have been applied to brain MRIs to differentiate patients from controls[27,28], prognosticate outcomes[29], and differentially diagnose subtypes[30]. However, while research utilizing this technology has been influential, clinical translations are scarce[31]. Where techniques for segmenting brain tumors or detecting lesions typically produce segmentation masks that are innately interpretable, predicting a complex diagnosis would entail compressing all information contained in a high-dimensional brain scan into a single number. Using deep learning, the decisions underlying this immense reduction are obfuscated, both from the developer of the system, the clinical personnel using it, and the patient ultimately impacted by the decision. This black box nature is broadly credited for the low levels of adoption in safety-critical domains like medicine[32]. Responding to this limitation, explainable artificial intelligence (XAI) provides methodology to explain the behavior of ANNs[33]. The nature of these explanations varies, e.g., by what type of model is to be explained, what conceptual level the explanation is at, and who it is tailored for[34,35]. In computer vision, XAI typically aims for post-hoc explanations of individual decisions, explaining why a model arrived at a given prediction for a given image. Explanations are often provided in a visual format, as a heatmap indicating how different regions of the image contribute to the prediction[36]. Layerwise Relevance Propagation (LRP) is a variant of such a method, based on propagating relevance from the prediction-space, backwards through all layers of the model to the image-space, to form a relevance map[37]. A major advantage of LRP is its intuitive interpretation: by construction, the total amount of relevance which denotes contribution to the prediction is kept fixed between layers. Thus, the relevance propagated back to an input voxel is directly indicative of the influence of that exact voxel to the prediction. Recently, several studies have applied both LRP and other explainable AI methods to dementia[38], finding that the heatmaps generally highlight regions known to change in dementia[39–42]. However, the possibility of utilizing the fine-grained, individual, heatmaps produced by LRP to accurately characterize individualized disease manifestations has not been explored, despite its potential for supporting clinical decisions towards precision medicine[38,41].

In the present study, we applied techniques from deep learning and XAI on MRI scans of the brain to make explainable and clinically relevant predictions for dementia at the individual level (Fig. 1). Using a state-of-the-art architecture for neuroimaging data, we trained CNNs to differentiate patients diagnosed with dementia from healthy controls based on T1-weighted structural MRIs. We implemented LRP on top of the trained models to form a computational pipeline producing individual-level explanations in the form of relevance maps alongside the model predictions. The relevance maps were validated in a subset of dementia patients, both in a qualitative comparison with existing knowledge of the anatomical distribution of structural aberrations, and in a quantitative, predictive context. Next, we applied the pipeline to a large, longitudinal dataset of MCI patients to create individual morphological records, a proposed data format for tracking and visualizing disease progression. Finally, we investigated the clinical utility of these records for stratifying patients, both in terms of their specific clinical profile, and progression of the disease. To facilitate reproducibility and improve the translational value of our work, the trained models and the complete explainable pipeline is made accessible on GitHub.

Results

We compiled MRI data from multiple sources (Supplementary Table 1) into a dataset of heterogeneous dementia patients ($n = 854$, age range = 47–95, 47% females, Table 1) based on various diagnoses (Probable AD, vascular dementia, other/unspecified dementia) and diagnostic criteria for inclusion (Supplementary Table 2), and a set of controls strictly matched on site, age, and sex of equal size. We trained multiple CNNs to differentiate between the groups, employing a cross-validation approach utilizing all available timepoints for participants in three training folds and a single randomly selected timepoint for participants in separate validation and test folds. When stacking the out-of-sample predictions for all participants from all folds together ($n = 1708$), for each fold using the model with the best validation performance, we observed satisfactory discrimination with a combined area under the receiver operating characteristics curve (AUC) of 0.908 (0.904–0.920 split across folds, Supplementary Fig. 1), and an accuracy of 84.95% (83.04–87.13%, Supplementary Table 3). This is slightly below with what is commonly achieved in similar studies classifying a specific subtype (typically AD) in a single dataset[28].

Relevance maps highlight predictive brain regions in individuals with dementia

Based on the classifiers with the highest AUCs in the validation sets, we built an explainable pipeline for dementia prediction, $LRP_{dementia}$, using composite LRP[43], and a strategy to prioritize regions of the brain that contributed positively towards a prediction of dementia in the explanations. Using this pipeline, we computed out-of-sample relevance maps for all participants by applying the model for which the participant was unseen. Qualitatively, these maps corroborated known anatomical locations with structural aberrations in dementia, while still allowing for inter-individual variation (Supplementary Fig. 2). We confirmed this apparent corroboration

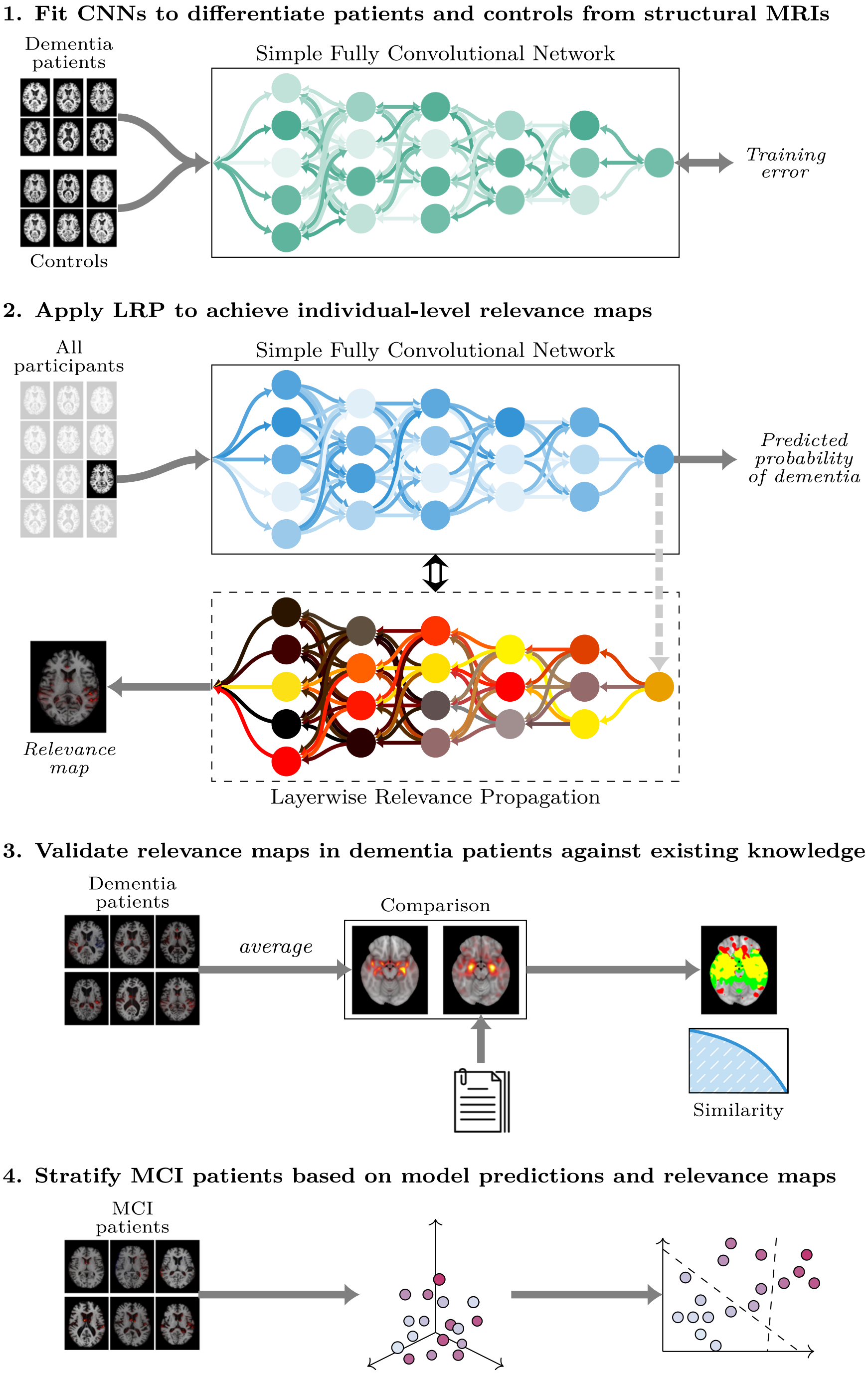
1. Fit CNNs to differentiate patients and controls from structural MRIs
Dementia patients
Controls
Simple Fully Convolutional Network
Training error
2. Apply LRP to achieve individual-level relevance maps
All participants
Simple Fully Convolutional Network
Predicted probability of dementia
Relevance map
Layerwise Relevance Propagation
3. Validate relevance maps in dementia patients against existing knowledge
Dementia patients
average
Comparison
Similarity
4. Stratify MCI patients based on model predictions and relevance maps
MCI patients

Fig. 1 | Overview of the modeling process. The modeling process consisted of four sequential steps. First, we fit multiple Simple Fully Convolutional Networks to classify dementia patients and healthy controls based on structural MRIs. Then we applied the best models to generate out-of-sample predictions and relevance maps for all participants. Next, we validated the relevance maps against existing knowledge using a meta-analysis to generate a statistical reference map. Finally, we employed the full pipeline in an exploratory analysis to stratify patients with mild cognitive impairment (MCI).

quantitatively by comparing a voxel-wise average map $\bar{R}_{dementia}$ (Supplementary Fig. 3), containing positive relevance from all correctly predicted dementia patients, with a statistical reference map G (Supplementary Fig. 4) from an activation likelihood estimation (ALE) meta-analysis[44], methodology established by an earlier study[40]. For sanity checks, we also computed average maps from three alternative pipelines, $\bar{R}_{sex}$, $\bar{R}_{randomized\ weights}$ and $\bar{R}_{randomized\ images}$. The comparisons with the reference map were done by binarizing the maps on both sides of the comparison at various thresholds and measuring the Dice overlap (Fig. 2a). For the three alternative pipelines the amount of overlap decreased monotonically as the binarization threshold rose (Fig. 2b), whereas for $\bar{R}_{dementia}$ it stabilized as the maps grew sparser, indicating its higher similarity with G. This effect was reaffirmed by a normalized cross-correlation[45] of 0.64 for $\bar{R}_{dementia}$, compared to 0.41, 0.40, and 0.12 of $\bar{R}_{sex}$, $\bar{R}_{randomized\ weights}$ and $\bar{R}_{randomized\ images}$, respectively. In addition, we performed a region-wise, qualitative comparison of $\bar{R}_{dementia}$ and G, also yielding general agreement (Fig. 2c), with the most important regions in both maps being the nucleus accumbens, the amygdala, and the parahippocampal gyrus. Next, we tested the importance of the detected regions in a predictive context, by applying an iterative mask-and-predict procedure. For each participant, we produced a baseline dementia-prediction $\hat{y}_0$ and relevance map R_{task} for each pipeline LRP_{task}. We then iteratively masked out the most important regions of the image according to the relevance map and recorded how the prediction changed as a function of the occlusion (Fig. 2d). Using only true positives, the predictions should ideally start out at ~1.0 (empirically found to be 0.89 on average) and trend towards 0.5 (random prediction) as a larger proportion of the image is occluded. The rate of decline is indicative of whether the masked regions contain information essential for the classifier to classify the image correctly.

Over 20 iterations we observed that the predictions based on maps from both $LRP_{dementia}$, LRP_{sex} and $LRP_{randomized\ weights}$ decreased, but $LRP_{dementia}$ at a distinctly steeper rate than the rest (Fig. 2d). To quantify this observation we calculated an area over the perturbation curve (AOPC) of 0.231, 0.009, −0.001 and 0.002 for $LRP_{dementia}$, LRP_{sex}, $LRP_{randomized\ images}$, $LRP_{randomized\ weights}$ respectively. Taken together, these results demonstrate that our pipeline generates maps with relevance in brain regions associated with changes in dementia.

Table 1 | Overview of cohorts

CNN training and cross-validation			
Cohort	**Participants**	**Mean age (± std)**	**Sex (F/M)**
Healthy controls	854	75.13±7.81	401/453
Dementia patients	854	74.82±7.84	401/453
Total	**1708**	**74.98±7.82**	**802/906**
Downstream prognostic and correlational analyses			
Improved MCI	80	71.18±8.14	37/43
Stable MCI	754	74.63±**7.66**	324/430
Progressive MCI	304	75.60±**7.46**	124/180
Total	**1138**	**74.67±7.73**	**485/653**

Key characteristics of the cohorts used for training and testing the predictive models, and the exploratory analyses utilizing their predictions.

Output from the explainable dementia pipeline has prognostic value for MCI patients

For the MCI patients (n = 1256, timepoints = 6448), previously unseen by all models, we built an averaging ensemble to procure a singular out-of-sample prediction and relevance map per patient per timepoint. Put together, we let this represent a morphological record (illustrated in Fig. 4) visualizing the absolute quantity (indicated by the prediction) and location (indicated by the relevance map) of dementia-related pathology detected by the models over time. Qualitatively, both predictions and maps were relatively stable within a participant over time, while allowing enough variation to compose what resembled a trajectory. To investigate the prognostic value of our proposed morphological records we divided the MCI patients into three subgroups based on their trajectories in the follow-up period: those who saw improvement of their condition (n = 80), those with a stable diagnosis throughout (sMCI, n = 754), and those who progressed into dementia (pMCI, n = 304). The remaining (n = 118) had either a non-MCI diagnosis at the first timepoint, or a more complex diagnostic trajectory (e.g., MCI- > AD- > CN) and were excluded from subsequent analyses. We observed that the predictions in the first group were generally very low (mean $\hat{y}$ = 0.13, Supplementary Fig. 5a), indicating that the models detected little, if any, evidence of dementia in these participants. For the stable

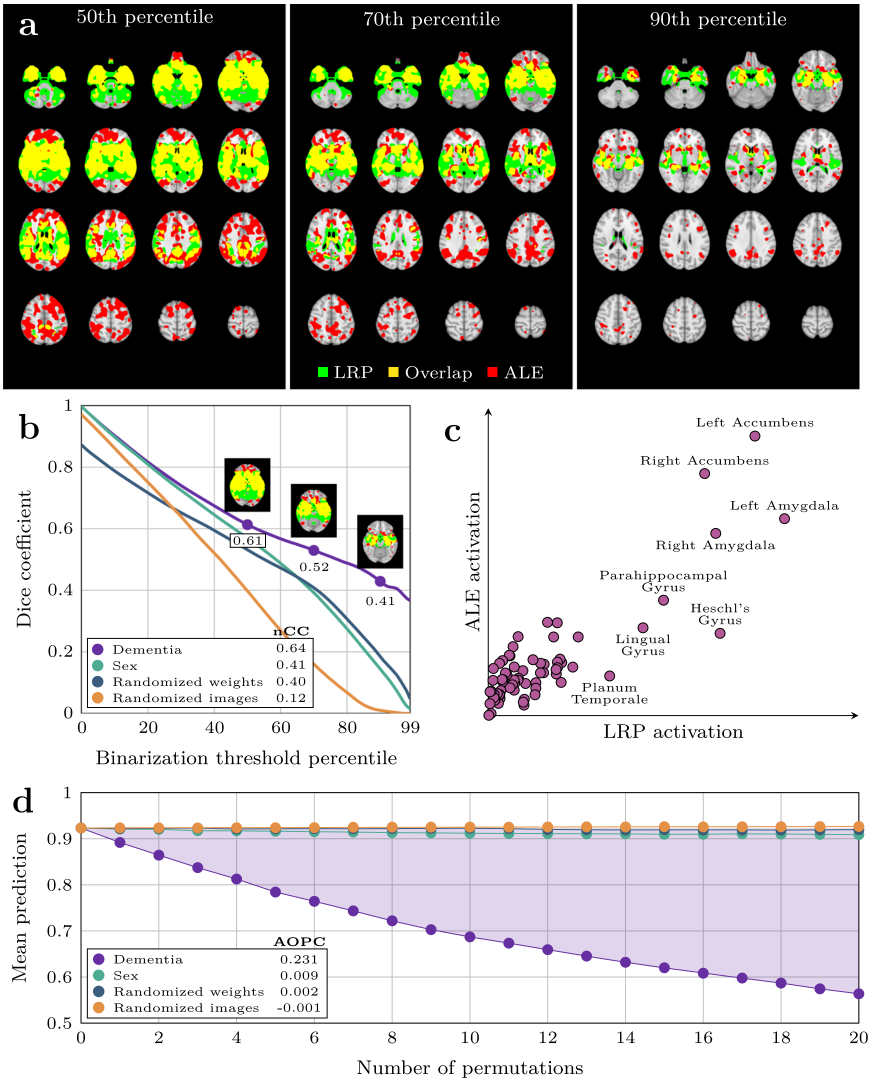

Fig. 2 | Validation of relevance maps from the dementia pipeline compared with three alternative pipelines. **a** Visualization of the comparison between the binarized average relevance map $\bar{\mathbf{R}}_{\mathbf{dementia}}$ from the dementia-pipeline and the binarized statistical reference map **G** from GingerALE, at different thresholds for binarization. **b** Overlap between the four average relevance maps $\bar{\mathbf{R}}$ from our four pipelines and **G** as a function of the binarization threshold. The numbers in the legend denote the normalized Cross Correlation (nCC) for each pipeline. **c** Mean voxel-wise activation in $\bar{\mathbf{R}}_{\mathbf{dementia}}$ and **G**, grouped by brain region. **d** Average participant-wise prediction from the dementia model after iteratively masking out regions of the image according to relevance maps from the four pipelines. Area over the permutation curve (AOPC) for the dementia map is indicated by the shaded area and denoted in the legend for all pipelines.

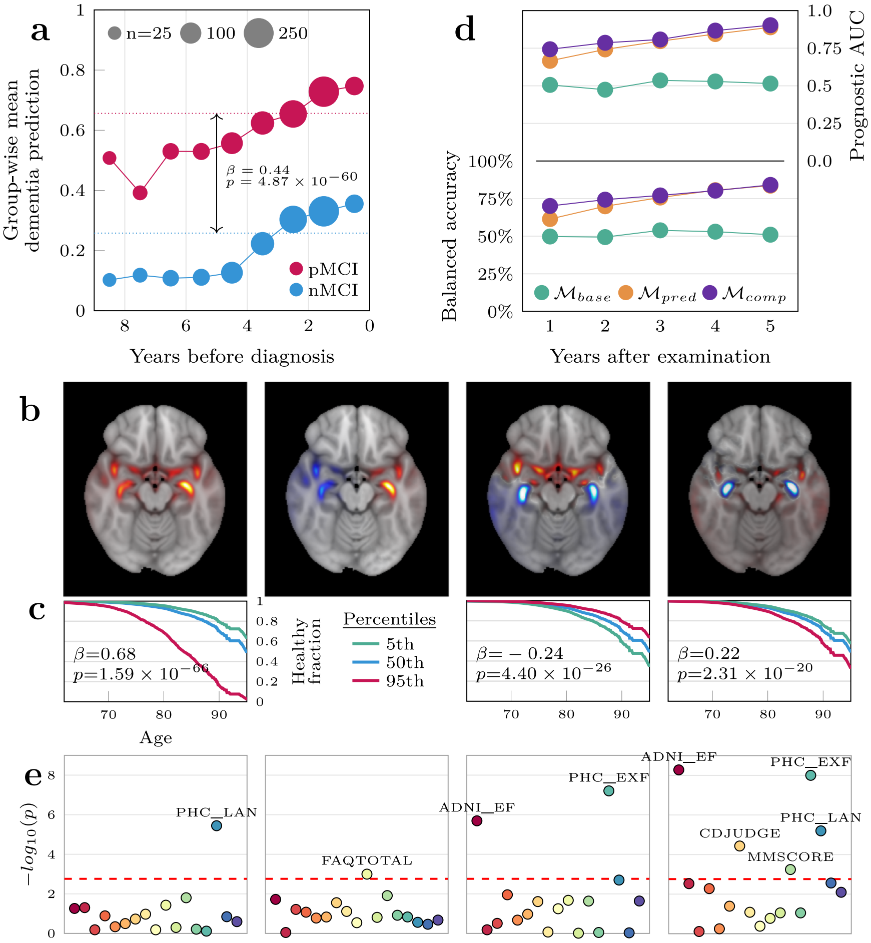

Fig. 3 | Utility of the dementia pipeline for predicting progression and characterizing individual-level deviations in the mild cognitive impairment cohort. **a** Group-wise mean predictions from the dementia-model in the progressive and non-progressive groups in the years before a diagnosis was given. **b** The four first voxel-wise components of the principal component analysis plotted in MNI152-space. **c** Survival curves for the average MCI patient (blue) and fictitious patients at the extreme percentiles of the span for each component. The second component was not significant and is not shown. **d** Predictive performance of the three models predicting progression in the years following the MRI examination. The baseline model ($\mathcal{M}_{\text{base}}$) included only sex and age as covariates, the next model $\mathcal{M}_{\text{pred}}$ included the prediction from the dementia classifier as a predictor, while the final model $\mathcal{M}_{\text{comp}}$ also added the component vectors representing the relevance maps. **e** Significance levels of correlations between the each of the four PCA components and various cognitive measures. The six annotated measures are composite language (PHC_LAN) and executive function (PHC_EXF) scores from the ADSP Phenotype Harmonization Consortium, total score from the Functional Activities Questionnaire (FAQTOTAL), composite executive function score from UW – Neuropsych Summary Scores (ADNI_EF), clinical evaluation of impairment related to judgment and problem-solving (CDJUDGE) from the Clinical Dementia Rating, and an overall measure of cognition from the Mini-Mental State Examination (MMSCORE, commonly referred to as MMSE).

patients the mean prediction was higher (mean $\hat{y} = 0.33$), but still below the classification threshold of 0.5, whereas in the progressive group the model predicted the average patient to already have dementia (mean $\hat{y} = 0.72$). Importantly, this was also true when considering only timepoints before these patients received the clinical diagnosis (mean $\hat{y} = 0.65$, Supplementary Fig. 5b), suggesting that the model found evidence of the disorder before the clinical symptoms surpassed the diagnostic threshold. To formally delineate the differences in predictions leading up to the potential diagnosis, we combined the improving and stable patients into a non-progressive group (nMCI, $n = 834$), and sampled patients to match the progressive group based on their visiting histories, leading up to a terminal diagnosis timepoint (or a constructed non-diagnosis timepoint in the non-progressive group). In this matched dataset ($n = 550$) we applied a linear mixed model controlling for age and sex and observed that the group difference was even greater than what we previously observed ($\beta = 0.47$, $p = 6.05 \times 10^{-71}$, Fig. 3a, Supplementary Table 4). Furthermore, we observed a significant difference in longitudinal slopes ($\beta = 0.05$ increase in prediction per year, $p = 8.14 \times 10^{-17}$) indicating a greater rate of brain change detected by the model in those who would be diagnosed with dementia at a later point in time.

The large group differences in the dementia predictions leading up to a potential diagnosis suggest this as a biomarker with innate prognostic value, yet the most salient part of our morphological records were the relevance maps. Thus, we performed exploratory analyses based on these to further differentiate the non-progressive and progressive groups and characterize both inter- and intra-group heterogeneity. However, given the high dimensionality of the maps and the relatively small number of patients, we first applied a principal component analysis (PCA) to relevance maps from all MCI patients, effectively compressing their information content into a smaller set of characteristic variables encoding facets of the maps, enabling the subsequent analyses. We retained the 64 components that explained the largest amount of variance and observed that they qualitatively clustered into three overarching categories. The first component was a generic component detecting general presence of relevance, resembling the average map from dementia patients, and thus made up a cluster by itself. The next cluster was comprised of the subsequent three components that captured high-level, abstract patterns of relevance, namely differences in lateralization, along the sagittal axis and in subcortical regions (Fig. 3b). The final cluster consisted of the remaining 60 components that captured specific, intricate patterns of presence/non-presence of relevance in regions revealed in the preceding analyses (Supplementary Fig. 6). To investigate the potential of using the relevance maps for prognosis, we first performed a survival analysis using a Cox proportional hazards model where getting a diagnosis was considered the terminal event.

Specifically, we modeled the fraction of the population without a diagnosis as a function of age and used the subject-wise loadings of c_t from the PCA as predictors. After Benjamini-Hochberg correction, 37 of these components were significantly associated with staying undiagnosed (Fig. 3c and Supplementary Table 5). However, we observed a correlation between the singular dementia prediction $\hat{y}$ and the absolute magnitudes of these components (Supplementary Fig. 7), indicating that the associations in the survival analysis could be induced by differences in the prediction rather than variability in the relevance maps. To mitigate this concern, we fit an equivalent model while stratifying on $\hat{y}$, observing that 29 associations remained significant, and that all coefficients had the same sign. Nonetheless, this analysis did not account for the predictions and relevance maps changing within a participant over time, so we reframed the question in a purely predictive setting, constructed to bear resemblance to a clinical

Table 2 | Predictive performance of the three models predicting progression five years into the future

Model	AUC	Balanced accuracy	PPV	Sensitivity	Specificity
$\mathcal{M}_{base}$	0.515	51.05%	0.14	0.09	0.93
$\mathcal{M}_{pred}$	0.889	83.61%	0.91	0.83	0.84
$\mathcal{M}_{comp}$	0.903	84.1%	0.92	0.82	0.86

The baseline model $\mathcal{M}_{base}$ used only age and sex as covariates. $\mathcal{M}_{pred}$ also added the prediction from the dementia model at the current timepoint as a predictor, while $\mathcal{M}_{comp}$ additionally included the component vector c_t encoding information from the relevance maps.

scenario, using the same participants (nMCI = 834, pMCI = 304, total $n = 1138$). For each MCI patient p at each timepoint t we asked whether we were able to predict, at yearly intervals y up to five years into the future, whether p had progressed into dementia, using information from $LRP_{dementia}$ available at t. Importantly, all timepoints for all these participants were unseen by the dementia-model, yielding out of sample predictions and relevance maps from $LRP_{dementia}$, and we employed nested cross-validation to ensure the progression predictions were also out-of-sample. First, we fit a baseline model $\mathcal{M}_{base}$ with age and sex as predictors, showing no predictive efficacy at any timepoint (all AUCs ≈ 0.5, Supplementary Table 6), indicating that the dataset was not biased with respect to these variables. When adding the prediction from the dementia model $\hat{y}_t$ as a predictor in model $\mathcal{M}_{pred}$ we saw large improvements in prognostic efficacy at all yearly intervals, culminating with a fold-wise mean AUC of 0.889 after five years (Fig. 3d). In the final model, $\mathcal{M}_{comp}$, also including the component vector c_t as predictors, we saw further improvements for all years, peaking at 0.903 after five years ($p = 0.035$ when compared to $\mathcal{M}_{pred}$ in a Wilcoxon signed-rank test across the outer folds). Overall, our best performing model predicted progression to dementia after five years with an AUC of 0.903, an accuracy of 84.1%, a positive predicted value of 0.92, a sensitivity of 0.82 and a specificity of 0.86 (Table 2).

Facets of the relevance maps are associated with cognitive impairments in distinct domains

Finally, we tested whether common features found in the relevance maps, represented by the PCA component, were correlated with impairments in distinct cognitive and functional domains. We extracted 17 summary measures from 7 neuropsychological tests (Supplementary Tables 7 and 8), performed approximately at the same time as an MRI examination, and tested for associations with the subject-wise loadings of c_t in 733 MCI patients using linear models. After FDR correction, while correcting for age, sex and $\hat{y}$, we found 48 significant correlations between 18 unique components and 14 of the cognitive measures (Fig. 3e). Component 30 and the aggregate score from the Functional Activities Questionnaire (FAQTOTAL) had the highest number of significant hits among the components and measures respectively, both with six passing the threshold. Most importantly, the components showed distinct patterns of associations with the different cognitive measures. To ensure the significant associations were not driven by collinearity between components c_i and $\hat{y}$, we ran an equivalent analysis without including $\hat{y}$ as a predictor, observing that only 5/48 of the previously significant hits had coefficients with the opposite sign. To summarize, the spatial features captured in our relevance maps, and subsequently in our component vectors, were associated with distinct patterns of performance on neuropsychological tests relevant for characterizing phenotypic heterogeneity in dementia patients (Supplementary Fig. 8).

Discussion

Given the huge burden of the disease and an expected increase in prevalence, innovative technological solutions for clinical decision-making in dementia diagnostics and prognostics are urgently needed. Although commonly referred to as a homogenous condition or split into a few subtypes based on etiology or pathophysiology[17], dementia patients exhibit unique and complex deficiencies, disease trajectories, and cognitive deficits. To explore the potential of brain MRI and XAI to characterize heterogeneity in the brain underpinnings of dementia, we trained neural networks to differentiate dementia patients from healthy individuals and derived relevance maps using Layerwise Relevance Propagation to explain the individual-level decisions of the classifier. The relevance maps were specific to the individual, spanned regions that were predictive of dementia and corroborated existing knowledge of the anatomical distribution of structural aberrations. In a cohort of MCI patients, it enabled characterization and differentiation of individual-level disease manifestations and trajectories linked to cognitive performance in multiple domains. While further validations in clinical contexts are needed, our XAI pipeline for dementia demonstrates how advanced predictive technology can be employed by clinicians to monitor and characterize disease development for individual patients.

There is a multitude of XAI techniques available for explaining the decisions of an image classifier, many of which have yielded promising results for dementia classification[38]. We employed LRP due to its straightforward interpretation as well as earlier studies indicating robustness[46] and specificity[42], properties we consider integral in a clinical decision support system. But while procuring explanations that are *ipso facto* meaningful is an important step towards adoption of AI in clinical neuroimaging, it is not in itself sufficient. There is a host of predictive models that are trivially explainable, but not understandable[47], and there is genuine concern that XAI will lead to another level of systems that are formally well-defined, but opaque and obscure, and thus practically useless[48]. Thus, empirical explorations are imperative to investigate the nature of these explanations, examine how they may be useful and build essential trust[49]. In our validation, we observed that the explanatory maps produced by the dementia pipeline were more predictive and showed distinctly more agreement with existing knowledge of pathology than those produced by the three alternative pipelines. Given limitations that have been exposed in such methods earlier[50,51] these validations are crucial, and observing that our results both corroborate earlier evidence[40] and extend upon it, provides confidence that the explanations derived from the model are meaningful. However, we emphasize that the ultimate validation should happen in actual implementations of the technology in end-user systems, with clinical personnel applying it in clinical scenarios on realistic data.

We continued beyond validating the relevance maps by proposing them as a potential epistemic and clinical tool to characterize individual facets of dementia. To this end, we explored if the maps contributed to predicting imminent progression from MCI to dementia, and correlated them with different cognitive measures, extending upon the current literature[38]. In both analyses we found evidence, although modest, that the maps are informative beyond the predictions of the model. To illustrate the potential of the pipeline for clinical decision making we compiled its output into a proposed morphological record (visualized for a single patient in Fig. 4) that can help clinicians localize morphological abnormalities during a diagnostic process. Identifying subtle pathophysiology through deep phenotyping could have a huge potential for charting the heterogeneity of dementia, providing precise biological targets to guide future research. Furthermore, for the individual patient, it can support personalized diagnosis to identify appropriate disease-modifying treatments, and in the future, hopefully, accurate therapeutic interventions.

The regions with the highest density of relevance in our maps were the nucleus accumbens, amygdala and the parahippocampal gyrus, all of which are strongly affected in dementia[52–54]. While the two latter corroborate the established involvement of the medial temporal lobe[55], it is surprising that the hippocampus does not appear in our analyses, as it has frequently in similar studies[38,41,42]. While this could be caused by actual localization of pathology[56] we consider it more likely to be related to the internal machinery of the model. Specifically, the CNN relies on spatial context to identify brain regions before assessing their integrity, utilizing filters that span areas of the image larger than those containing the region itself. In the backwards pass, LRP uses these filters, and thus the localization of relevance is not necessarily voxel precise. Furthermore, we believe the model broadly can be seen as an atrophy detector, which necessarily entails looking for gaps surrounding

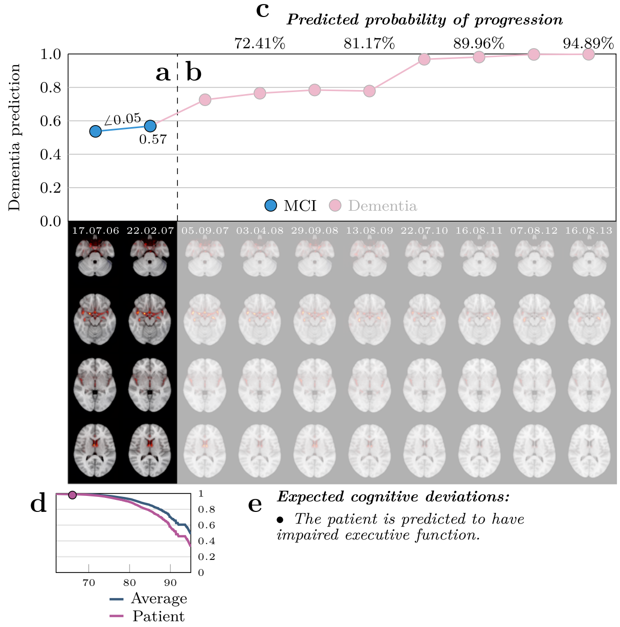

Fig. 4 | A visualization of the proposed morphological record for a randomly selected progressive MCI patient that was held out of all models and analyses. **a** The top half shows the prediction from the dementia model at each visit, while the bottom part displays the relevance map underlying the prediction. The opaque sections (including **c**, **d**, and **e**) contain information accessible at the imagined current timepoint (22.02.07) to support a clinician in a diagnostic procedure. The angle (∠) represents the change in dementia prediction per year based on the first two visits. **b** Translucent regions reveal the morphological record for the remaining follow ups in the dataset, thus depicting the future. The ground truth diagnostic trajectory is encoded by the color of the markers. **c** Predicted probabilities of progression at future follow-ups based on the prediction and relevance map at the current timepoint. **d** Survival curve of the patient compared to the average MCI patient calculated from the prediction and relevance map. The marker indicates the location of the patient at the current timepoint. **e** A list of cognitive domains where the patient is predicted to significantly differ from the average based on the prediction and relevance map.

regions instead of directly at the regions themselves. Therefore, while the relevance maps provide important information, they depend on contextual information and thus rely on interpretation from clinicians to maximize their utility in clinical practice.

We focused our analyses mainly on the relevance maps, but the results with largest, immediate, potential for clinical utility were the predictions from the dementia classifier. Other studies have shown the efficacy of machine learning models in differentiating dementia patients and healthy controls[28], but it is intriguing that we see a large discrepancy in the predictions of the progressive and non-progressive MCI patients many years before the dementia diagnosis is given. This corroborates findings from theory-driven studies[57] and a recent deep learning study[27], implying detectable structural brain changes many years before the clinical diagnosis is given. This gives hope for advanced technology to contribute to early detection and diagnosis through MRI-based risk scores, in our case supported by a visual explanation. If curative treatments prove efficacious and become accessible, early identification of eligible patients could be imperative[58]. Furthermore, timely access to interventions have shown efficiency in slowing the progress of cognitive decline[59], in addition to improving the quality of life for those afflicted and their caregivers[26,60]. Widely accessible technology that allows for early detection with high precision could play a key role in the collective response to the impending surge of patients and provide an early window of opportunity for more effective treatments.

While our results show a great potential for explainable AI, and particularly LRP, as a translational technology to detect and characterize dementia, there are limitations to our study. First, there are technical caveats to be aware of. Most importantly, there is an absolute dependence between the predictions of our model and the relevance maps. In our case, when we qualitatively assessed the relevance maps of the false negatives, they were indistinguishable from the true negatives. This emphasizes the fact that when the model is wrong, this is not evident from the explanations. Next, while the maps contain information sufficient to explain the prediction, they are not necessarily complete. Thus, they don't contain all evidence in the MRI pointing towards a diagnosis, a property which could prove essential for personalization. We have addressed this problem through pragmatic solutions, namely ensembling and targeted augmentations, but theoretical development of the core methodology might be necessary to theoretically guarantee complete maps. Beyond the fundamental aspects of LRP, there are weaknesses to the present study that should be acknowledged. First, the dataset with dementia patients portrayed as heterogeneous mostly consists of ADNI and OASIS data, and thus patients with a probable AD diagnosis (although clinically determined). Thus, while we consider it likely, it is not necessarily true that the dimension of variability spanning from healthy controls to dementia patients portrayed by our model has the expressive power to extrapolate to other aetiologies. To overcome this in actual clinical implementations, we encourage the use of datasets that are organically collected from subsets of the population that are experiencing early cognitive impairments, for instance from memory clinics. Furthermore, it is not trivial to determine whether a clinical, broad, dementia-label is an ideal predictive target for models in clinical scenarios. Both ADNI and AIBL contain rich biomarker information with multiple variables known to be associated with dementia, such as amyloid positivity. It would be intriguing to see studies methodologically similar to ours with a biological predictive target, and we encourage investigations into whether this supports and complements the results we have observed here. Another limitation of the present study is out-of-sample generalization, especially related to scanners and acquisition protocols. Although we utilize data from many sites, which we have earlier shown to somewhat address this problem[61], in combination with transfer learning, we did not explicitly test this by e.g., leaving sites out for validation. Again, we advise that clinical implementations should be based on realistic data, and thus at least be finetuned towards data coming from the relevant site, scanner, and protocol implemented in the clinic[62]. This also includes training models with class frequencies matching those observed in clinical settings, instead of naively balancing classes as we have done here. Next, we want to explicitly mention the cyclicality of our mask-and-predict validation. In a sense it trivially follows that regions that are considered important by a model are also the ones that are driving the predictions, and thus it is no surprise that the relevance maps coming from the dementia model are more important to the dementia model than the maps coming from e.g., the sex model. We addressed this by alternating the models for test and validation, but fully avoiding this circularity would require disjunct datasets, and more and larger cohorts. Finally, we highlight the potential drawbacks of including the improving MCI patients alongside the stable in the progression models. We believe this accurately depicts a realistic clinical scenario, where diagnostic and prognostic procedures happen based on currently available clinical information. However, that these patients improve could indicate that their condition is not caused by stable biological aberrations. This could oversimplify the subsequent predictive task, inflating our performance measures. In summary, the predictive value we observed for the individual patient must be interpreted with caution. However, our extensive validation approach as well as our thorough explanation of the method and its limitations, and training on large datasets, provide a first step towards making explainable AI relevant for clinical decision support in neurological disorders. Nonetheless, it also reveals a complicated balance between validating against existing knowledge and allowing for new discoveries. In our case, confirming whether small details revealed in the relevance maps are important aspects of individualization or simply intra-individual noise requires datasets with a label-resolution beyond what currently exists. Thus, we reiterate our belief that the continuation of our work should happen at the intersection between clinical practice and research[63], by continuously collecting and labeling data to develop and validate technology in realistic settings.

To conclude, while there are still challenges to overcome, our study provides an empirical foundation and a roadmap for implementations of brain MRI based explainable AI in personalized clinical decision support systems. Specifically, we show that deep neural networks trained on a heterogenous set of brain MRI scans can predict dementia, and that their predictions can be made human interpretable. Furthermore, our pipeline allows us to reason about structural brain aberrations in individuals showing early signs of cognitive impairment by providing personalized characterizations which can subsequently be used for precise phenotyping and prognosis, thus fulfilling a realistic clinical purpose.

Methods

Data

The data used here was obtained from previously published, publicly accessible studies. All of these collected informed consents from their participants and received approval from their respective institutional review board or relevant research ethics committee. The present study was performed with approval from the Norwegian Regional Committees for Medical and Health Research Ethics (REK) and conducted in accordance with the Helsinki Declaration.

To train the dementia models we compiled a case-control dataset from seven different sources (Supplementary Table 1), consisting of patients with a dementia diagnosis and healthy controls from the same scanning sites. Because of the different diagnostic criteria used in the original datasets, we applied different rules to achieve a singular, heterogeneous dementia label (Supplementary Table 2). We extracted all participants with a dementia-diagnosis at all timepoints to comprise the patient group ($n = 854$). Then, for each unique proxy site (In ADNI, due to a large number of scanners and acquisition protocols, and the work put into unifying them, we used field strength as a proxy for site), sex, and age-bin spanning 10 years, we sampled an equal number of healthy controls to form the matched control set (total $n = 1708$, Table 1). Lastly, before modeling, we split the data into five equally sized folds stratified on diagnosis, site, sex, and age, such that all timepoints for a single participant resided in the same fold.

For the MCI dataset, we started with all participants from all ADNI waves with an MCI diagnosis (subjective memory complaint, MMSE between 24 and 30, CDR > 0.5 with memory box > 0.5, Weschler Memory Scale-Revised <9 for 16 years of education, <5 for 8–15 years of education and <3 for 0–7 years of education)[64], on at least one timepoint. These were 12661 images from 6448 visits for 1256 participants, none of which were used for model training. This selection criterion ensured all participants had an MCI diagnosis at one point in time, though it did not limit us to only those timepoints. Thus, in addition to those with a consistent, stable, MCI diagnosis (sMCI), we had a variety of diagnostic trajectories, including those transitioning from normal cognition to MCI, MCI to AD (pMCI) and various other combinations. Before the subsequent analyses we discarded all participants without an MCI diagnosis initially, and everyone with ambiguous trajectories (e.g., MCI- > CN- > AD), leaving 5607 visits from 1138 participants.

From these two datasets, we extracted T1-weighted structural MRI data for each participant at each timepoint to use as inputs for the subsequent predictive models. Prior to modeling, the raw images were minimally processed using a previously developed pipeline[58] relying on FreeSurfer v5.3 and FSL v6.0[65] to perform skullstripping[66] and linear registration to MNI152-space[67] with six degrees of freedom. Consequently, the processed images consisted of normalized voxel values from the raw images, registered to a common spatial template and contained minimal non-brain tissue.

Modeling

All dementia models were variants of the PAC2019-winning simple fully convolutional network architecture[68,69], modified to have a single output neuron with a sigmoid activation. The architecture is a simple, VGG-like convolutional neural network with six convolutional blocks and ~3 million parameters. We initialized the model with weights from a publicly accessible brain age model previously shown to have superior generalization capabilities when dealing with unseen scanning sites and protocols[61]. The models were trained on a single Nvidia A100 GPU with 40 GB of memory, Tensorflow 2.6[70] through the Keras interface[71]. We used a vanilla stochastic gradient descent (SGD) optimizer with a learning rate defined by the hyperparameter settings (see next section), optimizing the binary cross-entropy loss. All models ran for 160 epochs with a batch size of 6, and for each run the epoch with the lowest validation loss was chosen. Varying slightly depending on the hyperparameters, a single model trained in ~4 h.

For each hold-out test fold we trained models on three of the remaining folds and validated on the fourth, akin to a cross-validation with an additional out-of-sample test set, to achieve out-of-sample predictions for all 1708 participants while allowing for hyperparameter tuning. The hyperparameters we optimized were dropout $d \in \{0.25,\ 0.5\}$ and weight decay $w \in \left\{10^{-2},\ 10^{-3}\right\}$. Additionally, we tested stepwise, one-cycle and multi-cycle learning rate schedules and a light and a heavy augmenter. Initial values for the learning rate were set manually based on a learning rate sweep[72], though kept conservative to preserve the learned features from the pretraining. The hyperparameter search was implemented as a naive grid-search over the total 24 different configurations (Supplementary Fig. 9). We selected the model procuring the best AUC in the validation set to produce out-of-sample predictions for the outer hold-out fold. In the final evaluation of the models, we compiled predictions for all participants, for each using the model where they belonged to the hold-out test set. Our main method for measuring performance was the AUC, but we also report accuracy, which, due to our matching procedure, is equivalent to balanced accuracy.

Relevance maps

We built a pipeline $LRP_{dementia}$ for generating relevance maps by implementing LRP[37] on top of the trained classifier. LRP is a technique for explaining single decisions made by the model, and thus, when running the pipeline on input X a relevance map R is generated alongside the prediction $\hat{y}$. R is a three-dimensional volume, representing a visual explanation for $\hat{y}$, where each voxel $r_{i,j,k} \in R$ has a spatial position i, j, k corresponding to the location of an input voxel $x_{i,j,k} \in X$. Furthermore, the intensity of $r_{i,j,k}$ can be directly interpreted as how much voxel $x_{i,j,k}$ contributes to $\hat{y}$, such that $\sum_{r \in R} r = \hat{y}$. In the original LRP-formulation, relevance r is propagated backwards between subsequent layers Z_l and Z_{l+1} with artificial neurons $a_m \in Z_l$ and $a_n \in Z_{l+1}$ such that $r(a_m)$ is proportional to how much a_m contributes to the activations of all a_n in the forward pass (Eq. (1)).

$$r(a_m) = \sum_j \frac{a_m w_{mn}}{\sum_o a_o w_{on}} r(a_n), \tag{1}$$

where w_{mn} denotes the weight between a_m and a_n

We controlled the influence of different aspects of the explanations using a composite LRP strategy[43], combining different formulations of the LRP formula for the different layers in the model to enhance specific aspects of the relevance maps. Specifically, we employed a combination of alpha-beta and epsilon rules that have previously shown to produce meaningful results for dementia classifiers[41,42]. For the prediction layer, we retained the most salient explanations through an LRP_ϵ-rule (Eq. (2)).

$$r_\epsilon(a_m) = \sum_n \left(\frac{a_m w_{mn}}{\epsilon + \sum_o a_o w_{on}} \right) r(a_n) \tag{2}$$

For the central convolutional layers, we upweighted positive relevance (e.g., features increasing the prediction, corresponding to evidence for a diagnosis) with $LRP_{\alpha\beta}$-rules (Eq. (3)).

$$r_{\alpha\beta}(a_m) = \sum_n \left(\alpha \frac{(a_m w_{mn})^+}{\sum_o (a_o w_{on})^+} - \beta \frac{(a_m w_{mn})^-}{\sum_o (a_o w_{on})^-} \right) r(a_n), \tag{3}$$

where $(\cdot)^+$ and $(\cdot)^-$ denote positive and negative contributions respectively

For the input layer and the subsequent convolutional layer, we employed LRP_b to smooth finer details of the relevance maps (Eq. (4)).

$$r_b(a_m) = \sum_n \frac{1}{|o|} \tag{4}$$

where $|\boldsymbol{o}|$ denotes the number of nodes connected to $\boldsymbol{a_n}$.

The resulting relevance maps produced by the pipeline were full brain volumes with the same dimensionality as the MRI data (167 × 212 × 160 voxels) containing mostly (see below) positive relevance.

Notation-wise we generally consider the relevance map $R(X)$ for an image X to be a function of the model m_{task}, where *task* indicates which task the model was trained for, the LRP strategy $\mathrm{LRP}_{\mathrm{composite}}$ and the image X (Eq. (5)).

$$R(X) = f\left(m_{task}, \ \mathrm{LRP}_{\mathrm{composite}}, \mathrm{X}\right) \tag{5}$$

Because the composite LRP strategy described above is kept fixed in our pipeline, this can be contracted (Eq. (6)).

$$R(X) = f\left(m_{task}, X\right) \tag{6}$$

Furthermore, the model-specifier *task* can also annotate the map for a further simplification (Eq. (7)).

$$R_{task}(X) = f(X) \tag{7}$$

Thus, LRP_{task} is used to annotate the full pipeline for a given task, while $R_{task}(X)$ denotes a single relevance map generated by this pipeline for image X. When the task is given by the context, we sometimes simplify this further to $R(X)$, and when a general image is considered, we simply use R to denote its relevance map.

While we generally discuss our pipeline as a singular one, there were in reality five approximately equivalent pipelines (corresponding to the models trained for the five test folds), and which one is used depends on what image was used as input. Specifically, for each participant diagnosed with dementia, the pipeline is chosen where the participant was part of the hold-out test set while training the model, and both the relevance maps and the predictions are thus always out-of-sample. For participants used in the MCI analysis, which are all out-of-sample for all models, we created an ensemble by averaging the predictions and the voxel-wise relevance across all models.

Before implementing the LRP procedure we made two slight modifications to the models to facilitate the backwards relevance propagation, both leaving the functional interface of the model unchanged. First, we removed the sigmoid activation in the final layer, so that the output of the model changed from a bounded continuous variable $\hat{y} \in [0, 1]$ to an unbounded prediction $\hat{y}_\sigma \in [-\infty, \infty]$. In this space a raw prediction of $\hat{y}_\sigma = 0$ is equivalent to a sigmoid-transformed prediction of $\hat{y} = 0.5$, and thus $\hat{y}_\sigma < 0$ means that the model predicts control status for the given participant, and oppositely $\hat{y}_\sigma > 0$ implies that the model predicts a dementia diagnosis. Furthermore, this means that all positive relevance $r \in R, \ r > 0$ can be interpreted as visual evidence in favor of a dementia diagnosis. Secondly, we modified the model by fusing all batch normalization layers with their preceding convolutional layers, adjusting their weights and biases to match the shift and scaling previously performed by the normalization layer[73,74].

After generation, the relevance maps are in the same stereotaxic space as their corresponding, linearly registered, input MRIs. To ensure intra-individual comparisons were done in the same space we non-linearly registered the maps to MNI152- space before subsequent statistical analyses were run. First, we registered the preprocessed MRIs X used as inputs to the 1 mm MNI152 template packaged with FSL using fnirt with splineorder=2. We then applied the transformation computed for X to $R(X)$ using applywarp. We also restrained our relevance maps to contain strictly positive relevance, evidence in favor of a dementia prediction, by clipping them to a minimum value of 0. Furthermore, to remove edge-effects from our analyses, we enforce that there is no relevance in non-brain tissue by nullifying all relevance outside the brain (Eq. (8)).

$$\forall(i, j, k)\left[x_{i,j,k} = 0 \ \Rightarrow \ r_{i,j,k} = 0\right] \tag{8}$$

All visualized relevance maps are plotted after non-linear registration, overlayed on the MNI152-template. As the maps are three-dimensional, we generally plot a collection of distributed axial slices. The relevance is colored by the nibabel v3.2.2[75] cold_hot colourmap. Since the absolute relevance values vary between maps, all maps are normalized to the intensity range [0, 1] in the visualizations.

Validating the relevance maps

Earlier studies have shown that interpretability techniques in general are prone to generate visual explanations that do not capture salient parts of the input[50,51]. To investigate the extent of this for our pipeline $LRP_{dementia}$ we employed two analyses to assess the sanity of the relevance maps. The first was an established task-specific technique comparing the relevance maps to existing knowledge of the pathology of dementia[40]. The second was a purely quantitative analysis examining how important the regions found by the pipeline are for the dementia prediction $\hat{y}$. In both cases we contrasted the relevance maps generated from the main pipeline with three alternative pipelines representing variants of a null hypothesis, all expected to produce relevance maps with no significant association with dementia.

$LRP_{random\ images}$ represents the simplest alternative pipeline, and is built around the dementia model, but with an additional preprocessing step scrambling the input (Eq. (9)).

$$R_{random\ images}(X) \ = \ R_{dementia}(\mathcal{X}), \tag{9}$$

where $\mathcal{X} = \mathcal{N}\left(\bar{X}, \ \sigma_X\right)$

$LRP_{random\ images}$ is expected to generate relevance maps where the relevance is evenly distributed across the entire image. In the next pipeline $LRP_{random\ weights}$ we replaced the dementia-model with a model with random weights (Eq. (10)).

$$R_{random\ weights}(X) = R\left(m_\theta, \ X\right) \tag{10}$$

m_θ has not been trained for any task, and thus has random weights initialized by the default Keras "*Glorot Uniform*" weight-initializer. This pipeline is expected to produce relevance maps which correlate with the raw voxel intensities, e.g., high intensity in the input should entail more (absolute) relevance, thereby reflecting aspects of morphology. The final and most realistic alternative pipeline was LRP_{sex}, where we replaced the dementia-model with a binary sex-classifier (Eq. (11)).

$$R_{sex}(X) = R\left(m_{sex}, \ X\right) \tag{11}$$

The sex-classifier was trained to differentiate males from females in one of the splits from the dementia-dataset, achieving an out-of-sample AUC of 0.956 and a balanced accuracy of 89.40%. We did not do any hyperparameter optimization for this model but used the best configuration from the dementia cross-validation in the same fold. The heatmaps from this pipeline should reflect regions where there is intra-individual variation in morphology, which are predictive of sex but with minimal association with dementia.

As a proxy for existing knowledge in the literature, we performed an ALE meta-analysis using Sleuth v3.0.4[76] and GingerALE v3.0.2[44]. We used

Sleuth to search for relevant articles with the query

Imaging Modality is MRI
AND
Context is disease
AND
Diagnosis is Dementia OR Alzheimer's Disease OR Lewy Body Dementia OR Frontotemporal
Dementia OR Non-Aphasic Frontotemporal Dementia

in the Voxel-based morphometry database, yielding 394 experiments from 124 articles. These experiments contained 3972 foci, 280 of which were outside the MNI152 mask, leaving 3692 to be loaded into GingerALE. Then the reference map G, with voxels $g_{i,j,k}$, was generated by an ALE meta-analysis using the default parameters: Cluster-level FWE = 0.01, Threshold Permutations = 1000, P value = 0.001. The reference map is visualized in Supplementary Fig. 4.

We performed four pairwise comparisons to estimate the amount of overlap between each of the pipelines and G. For each pipeline the comparison was performed by computing an average map $\bar{R}$, binarizing both it and G, and computing the Dice overlap between the two. The employed approach closely resembles the method of Wang et al.[40], but with multiple thresholds of binarization also for G, and allowed us to plot similarity as a function of the threshold. For each pipeline, we first computed an average relevance map $\bar{R}$ across all true positives (e.g., dementia patients that were correctly predicted to have a diagnosis by the dementia-model, $n = 697$), by computing their voxel-wise average. Next, we binarized both the average map (Eq. (12)) and the reference map (Eq. (13)) by thresholding them at multiple percentiles $p \in [0,\ 100)$.

$$\bar{R}_p = \begin{cases} 1, & r_{i,j,k} > percentile(\bar{R},\ p) \\ 0, & else \end{cases} \tag{12}$$

$$G_p = \begin{cases} 1, & g_{i,j,k} > percentile(G,\ p) \\ 0, & else \end{cases} \tag{13}$$

Then, for each percentile p we calculate the Sørensen-Dice coefficient SDC_p between the two (Eq. (14)).

$$SDC_p\left(\bar{R}_p,\ G_p\right) = \frac{\sum_{i,j,k} r_{i,j,k}\, g_{i,j,k}}{\sum_{i,j,k} r_{i,j,k} + \sum_{i,j,k} g_{i,j,k}},\ r \in \bar{R},\ g \in G \tag{14}$$

Additionally, to have a singular numerical basis for comparison, we computed the normalized cross-correlation[45] between the (non-binarized) average maps $\bar{R}$ and the reference map G (Eq. (15)).

$$nCC\left(\bar{R},\ G\right) = \frac{\sum_{i,j,k}(r_{i,j,k} - \bar{r})(g_{i,j,k} - \bar{g})}{\sqrt{\sum_{i,j,k}(r_{i,j,k} - \bar{r})^2 * \sum_{i,j,k}(g_{i,j,k} - \bar{g})^2}},\ r \in \bar{R},\ g \in G \tag{15}$$

To facilitate an intuitive understanding of what parts of the brain the dementia-model is focusing on, we also performed a similar, region-wise comparison. This was done by extracting a subset of voxels from the average relevance map $\bar{R}_{dementia}$ belonging to each region ρ (Eq. (16)) from the Harvard-Oxford cortical and subcortical atlases[77].

$$\bar{R}_\rho = \left\{ r_{i,j,k} \,|\, \left(i,\ j,\ k\right) \in \rho \right\}, \tag{16}$$
$$\text{where } \rho \text{ is a predefined region}$$

We did the same for G and let the mean activation per region for both constitute a tuple (Eq. (17)) plotted in Fig. 2c.

$$\left(\frac{\sum_{r \in R_\rho} r}{|R_\rho|},\ \frac{\sum_{g \in G_\rho} g}{|G_\rho|} \right) \tag{17}$$

However, as it is non-trivial to determine which aggregation method corresponds to the most understandable and intuitive interpretation, we also created plots for tuples of sums (Eq. (18)) and maximum values (Eq. (19)) per region in Supplementary Fig. 10.

$$\left(\sum_{r \in R_\rho} r,\ \sum_{g \in G_\rho} g \right) \tag{18}$$

$$\left(\max_{r \in R_\rho} r,\ \max_{g \in G_\rho} g \right) \tag{19}$$

To quantify the importance of the spatial locations captured by the various LRP pipelines for predicting dementia, we implemented a procedure for iteratively occluding parts of the image based on the relevance maps and observing how the prediction from the dementia model changed[78]. Still using the true positives, for each pipeline LRP_{task} for each MRI X_0 we generated a baseline dementia-prediction $\hat{y}_0$ and relevance map R_{task}. Then we located the voxel with the highest amount of relevance in R_{task} and replaced a 15 × 15 × 15 cube centered around the voxel with random uniform noise $\mathcal{U}(0,\ 1)$, effectively concealing all information contained in this region. Next, we ran the modified image X^1_{task} through the dementia-model to see how the prediction $\hat{y}^1_{task}$ changed as a function of the occlusion. Note that injecting a box of random noise into the image is not trivially equivalent to removing information, however we specifically applied the same modification in the random box-augmentation during training and are thus hopeful that the model is invariant to the injection beyond the information removal. We iteratively applied this modify-and-predict procedure, also masking out the regions from the relevant maps between each iteration to minimize overlap of occlusion windows, for 20 iterations, producing a list of predictions $\left[\hat{y}_0,\ \hat{y}^1_{task},\ \hat{y}^2_{task},\ \ldots,\ \hat{y}^{19}_{task}\right]$ plotted for each pipeline in Fig. 2d (averaged across all true positives). The rate of decline in these traces indicates the importance of the regions found in the respective relevance maps. We quantified the differences between the pipelines LRP_{task} by calculating the area over their perturbation curves[78] (AOPCs, Eq. (20)).

$$AOPC_{task} = \frac{1}{20}\left(\sum_{i=1}^{20} \hat{y}_0 - \hat{y}^i_{task} \right) \tag{20}$$

Exploratory analyses in the MCI cohort

In the exploratory MCI analyses, we used $LRP_{dementia}$ to generate predictions and relevance maps for participants from ADNI who were given an MCI diagnosis at inclusion. We first compiled the predictions and relevance maps (and the corresponding timestamps) for each participant at all timepoints into a single data structure we called a morphological record. We then tried to utilize this data structure to differentiate three groups: stable MCI patients (sMCI), progressive MCI patients (pMCI), and those who saw improvement in their cognition throughout the data collection phase. The remaining participants, e.g., those who either passed through all three diagnostic stages, or bounced between diagnoses, were excluded. Furthermore, we combined the stable and improving cohorts into a non-progressive group (nMCI) to facilitate binary group comparisons in the subsequent analyses.

For the first analysis comparing predictions in the two groups, due to variability in the total number and the frequency of visits between participants, we aimed to create a matched dataset based on visit history from the nMCI and pMCI cohorts to compare the predictions in the two groups with reference to a specific timepoint. We first started with all the progressive patients $p_p \in pMCI$ who got a diagnosis at timepoint t_{n+1}, and, for each patient individually, compiled all previous visits $t_m,\ m \le n$ into a vector h_p representing the time of the visits. The entries d_{t_m} of the vector were the number of days until the diagnosis was given,

$t_{n+1} - t_m$, including $d_{t_{n+1}} = 0$ (Eq. (21)).

$$h_p = \left[d_{t_0}, d_{t_1}, \ldots, d_{t_n}, 0\right] \tag{21}$$

Then, for each of the non-progressive patients $p_n \in nMCI$ who did not have a time of diagnosis (e.g., t_{n+1} is not given) we compiled a set H_p of all possible history vectors h_p by varying which visit was chosen as t_0 and a terminal non-diagnosis timepoint t_{n+1}. Next, we defined a cost-criterion for matching two histories (with an equal number of visits) as the sum of absolute pairwise differences between the vectors (Eq. (22)).

$$cost\left(h_1, h_2\right) = \sum_{m=0}^{n} |d_{t_m}^{h_1} - d_{t_m}^{h_2}| \tag{22}$$

For each pair of progressive and non-progressive patients (p_p, p_n) this allowed us to calculate a best possible match (Eq. (23)), given that the stable patient had a total number of visits equal to or larger than the number of visits for the progressive patient.

$$match(p_p, p_s) = \begin{cases} \min\limits_{h \in H_{p_s}} cost\left(h_{p_p}, h\right) & \exists h \in H_{p_s}\left(|h| = \left|h_{p_p}\right|\right) \\ \infty & else \end{cases} \tag{23}$$

Finally, we compiled the cost of the optimal match from all pairs into a matrix and found the best complete matching by minimizing the total cost across this matrix using the Hungarian algorithm implemented in scipy v1.6.3[79], such that each patient occurs in at most one pair.

We estimated differences in predictions $\hat{y}$ between the two groups using a linear mixed model. Specifically, we modeled $\hat{y}$ at all timepoints before the terminal timepoint t_{n+1} as a function of age, sex (as controlling variables), years to diagnosis, categorical group membership (nMCI, pMCI), and an interaction between years to diagnosis and group. In addition, we had an independent intercept and slope per participant. The model was fit through the formula API of statsmodels v0.13.2[80] using the formula from Eq. (24) on the matched dataset.

$$\begin{aligned} y \sim age + sex + years\ to\ diagnosis + C\left(group\right) \\ + years\ to\ diagnosis : C\left(group\right) \\ + \left(1 + years\ to\ diagnosis | subject\right) \end{aligned} \tag{24}$$

A full overview of coefficients and p values can be found in Supplementary Table 4.

Due to the high dimensionality of the relevance maps, we decomposed them with a principal component analysis (PCA) before the final analyses. To fit the PCA we used the non-linearly registered relevance maps from a randomly selected timepoint for all MCI patients. Before fitting the model, all relevance maps were smoothed with a constant 3 × 3 × 3 blurring kernel using the convolution operation from Tensorflow 2.6 to strengthen the signal-to-noise ratio. The PCA was computed using scikit-learn v1.0.2[81], retaining 64 components (out of 1137 maximally possible) in a component vector $c = \left[c_0, c_1, \ldots, c_{63}\right]$. An axial slice from each of the 64 components visualized in MNI152-space is shown in Supplementary Fig. 6.

We fit Cox proportional hazard models using the component vectors as predictors to assess the association between the relevance maps and progression as a function of age. In addition to the components, representing the maps, we controlled for sex in the model. The p values and coefficients can be found in Supplementary Table 5. To account for covariance between the components and the dementia-prediction $\hat{y}$ we ran an additional model where we divided the patients into ten strata based on $\hat{y}$. Both models were fit using lifelines v0.27.1[82].

To further explore the prognostic efficacy of our pipeline we set up a predictive analysis for predicting progression at multiple, fixed timepoints a given number of months in the future. For each participant p with visits at timepoints t^p, we denoted the last timepoint with an MCI diagnosis t_{neg}^p and the first timepoint with a dementia diagnosis (if present) t_{pos}^p. Using a fixed set of years into the future, $\gamma \in \{1, 2, 3, 4, 5\}$, we constructed a target variable $z_\gamma(t^p)$ encoding progression according to Eq. (25).

$$z_\gamma(t^p) = \begin{cases} 1 & t^p + \gamma \geq t_{pos}^p \\ 0 & t^p + \gamma \leq t_{neg}^p \\ NA & else \end{cases} \tag{25}$$

where the NAs allow for exclusion of all patients where the status at timepoint $t^p + \gamma$ is unknown. For each γ we constructed the target vector z_γ across all timepoints for all participants with $z_\gamma \neq NA$ and split the constituent patients p into five folds stratified on z_γ, sex and age, such that all timepoints from a participant resided in the same fold. Using these folds, we fit logistic regression models to predict z_γ with an l_1-penalty in a nested cross-validation loop, allowing us to both tune the regularization parameter λ and have out-of-sample predictions for all participants. For eligible participants we used all timepoints for training the models, but during testing we sampled a random timepoint per participant to ensure independence between datapoints in the final evaluation. For each γ we fit three models: a baseline model to assess the bias in the dataset with respect to age at the given timepoint t^p and sex (Eq. (26)), a model including the prediction $\hat{y}_{t^p}$ from the dementia classifier at t^p as a predictor (Eq. (27)), and a model including the relevance maps from t^p, represented by the component vector c_{t^p}, as additional predictors (Eq. (28)).

$$\mathcal{M}_{base} := z_\gamma \sim age_{t^p} + sex + age_{t^p} \times sex \tag{26}$$

$$\mathcal{M}_{pred} := z_\gamma \sim age_{t^p} + sex + age_{t^p} \times sex + \hat{y}_{t^p} + age_{t^p} \times \hat{y}_{t^p} \tag{27}$$

$$\mathcal{M}_{comp} := z_\gamma \sim age_{t^p} + sex + age_{t^p} \times sex + \hat{y}_{t^p} + age_{t^p} \times \hat{y}_{t^p} + c_{t^p} \tag{28}$$

All models were fit and tuned using the LogisticRegressionCV interface of sklearn v1.0.2[81]. We compared models by measuring the mean AUC across the five folds (Supplementary Table 6). To evaluate clinical applicability we also report accuracy, positive predictive value, sensitivity, and specificity (Table 2). To determine whether the more complex models represented significant improvements we employed a one-sided Wilcoxon signed-rank test from scipy v1.9.3[79] to do pairwise comparisons between $\mathcal{M}_{base}$ and $\mathcal{M}_{pred}$, and $\mathcal{M}_{pred}$, and $\mathcal{M}_{comp}$ across the five out-of-sample AUCs independently.

To assess whether the relevance maps were associated with specific cognitive functions we associated aspects of them with performance on various cognitive tests. We first extracted test results from seven neuropsychological batteries which spanned all ADNI waves and contained high-level summary scores from the ADNI website (Supplementary Table 7). We then manually extracted 17 summary scores spanning different, but overlapping, cognitive domains (Supplementary Table 8). The component vectors c were used as proxies for the relevance maps, where each represented a template for localization of pathology. We matched 2402 component vectors with test results from 733 MCI patients, forming a basis for the comparison. We then calculated the univariate association between cognitive performance according to each of the 17 with each of the dimensions $c_i \in c$, while including age and sex as covariates for correction. To isolate the effect of the localization we also corrected for dementia-prediction, $\hat{y}$. When a patient had multiple potential matches, a random timepoint was selected, and the final number of datapoints used in the analyses varied from 518 to 675. Correction for multiple testing was done with the Benjamini-Hochberg procedure. To ensure the associations were not confounded by collinearities between c and $\hat{y}$, we also performed an equivalent analysis without correction to observe whether the sign of the coefficients changed.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data used in this study were gathered from various sources, an overview including acknowledgments of their respective funding sources is provided in Supplementary Table 1. Among others, data used in the preparation of this article was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI, see adni.loni.usc.edu for further details), the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL, www.aibl.csiro.au) the AddNeuroMed consortium, and MIRIAD (www.nitrc.org/projects/miriad). The investigators within these studies contributed to the design and implementation of the data collection process but did not participate in the analysis or writing of this report, and this publication is solely the responsibility of the authors. Requests for access will need to be placed with the principal investigators responsible for the individual studies.

Code availability

The trained model and explainable pipeline and the underlying code are available at https://github.com/estenhl/pyment-public. Generic code for generating explanations for 3D CNNs is available at https://github.com/estenhl/keras-explainability.

Received: 10 October 2023; Accepted: 23 April 2024;
Published online: 02 May 2024

References

1. Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci* **20**, 365–377 (2017).
2. Bethlehem, Ra. I. et al. Brain charts for the human lifespan. *Nature* **604**, 525–533 (2022).
3. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
4. Arbabshirani, M. R., Plis, S., Sui, J. & Calhoun, V. D. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *NeuroImage* **145**, 137–165 (2017).
5. Sui, J., Jiang, R., Bustillo, J. & Calhoun, V. Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises. *Biol. Psychiatry* **88**, 818–828 (2020).
6. Davatzikos, C. Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. *NeuroImage* **23**, 17–20 (2004).
7. Westlin, C. et al. Improving the study of brain-behavior relationships by revisiting basic assumptions. *Trends Cogn. Sci.* **27**, 246–257 (2023).
8. Bzdok, D. & Ioannidis, J. P. A. Exploration, Inference, and Prediction in Neuroscience and Biomedicine. *Trends Neurosci.* **42**, 251–262 (2019).
9. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
10. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
11. Gauthier S., Webster C., Servaes S., Morais J. A., Rosa-Neto P. World Alzheimer Report 2022—Life After Diagnosis: Navigating Treatment, Care and Support (Alzheimer's Disease International, 2022).
12. Nichols, E. et al. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **18**, 88–106 (2019).
13. Vos, T. et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* **396**, 1204–1222 (2020).
14. World Health Organization. Global Status Report on the Public Health Response to Dementia (World Health Organization, 2021).
15. Nichols, E. et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. *Lancet Public Health* **7**, e105–e125 (2022).
16. Feldman, H. H. et al. Diagnosis and treatment of dementia: 2. Diagnosis. *CMAJ* **178**, 825–836 (2008).
17. Karantzoulis, S. & Galvin, J. E. Distinguishing Alzheimer's disease from other major forms of dementia. *Expert Rev. Neurother.* **11**, 1579–1591 (2011).
18. Echávarri, C. et al. Co-occurrence of different pathologies in dementia: implications for dementia diagnosis. *J. Alzheimer's Dis.* **30**, 909–917 (2012).
19. Schneider, J. A. Neuropathology of dementia disorders. *CONTINUUM: Lifelong Learn. Neurol.* **28**, 834 (2022).
20. Ryan, J., Fransquet, P., Wrigglesworth, J. & Lacaze, P. Phenotypic heterogeneity in dementia: a challenge for epidemiology and biomarker studies. *Front. Public Health* **6**, 181 (2018).
21. Ikram, M. A. et al. Brain tissue volumes in relation to cognitive function and risk of dementia. *Neurobiol. Aging* **31**, 378–386 (2010).
22. McDonald, C. R. et al. Relationship between regional atrophy rates and cognitive decline in mild cognitive impairment. *Neurobiol. Aging* **33**, 242–253 (2012).
23. Ferreira, D., Nordberg, A. & Westman, E. Biological subtypes of Alzheimer disease: a systematic review and meta-analysis. *Neurology* **94**, 436–448 (2020).
24. Verdi, S., Marquand, A. F., Schott, J. M. & Cole, J. H. Beyond the average patient: how neuroimaging models can address heterogeneity in dementia. *Brain* **144**, 2946–2953 (2021).
25. Rasmussen, J. & Langerman, H. Alzheimer's disease—why we need early diagnosis. *Degener. Neurol. Neuromuscul. Dis.* **9**, 123–130 (2019).
26. Robinson, L., Tang, E. & Taylor, J.-P. Dementia: timely diagnosis and early intervention. *BMJ* **350**, h3029 (2015).
27. Lu, B. et al. A practical Alzheimer's disease classifier via brain imaging-based deep learning on 85,721 samples. *J. Big Data* **9**, 101 (2022).
28. Mirzaei, G. & Adeli, H. Machine learning techniques for diagnosis of Alzheimer disease, mild cognitive disorder, and other types of dementia. *Biomed. Signal Process. Control* **72**, 103293 (2022).
29. Mirabnahrazam, G. et al. Predicting time-to-conversion for dementia of Alzheimer's type using multi-modal deep survival analysis. *Neurobiol. Aging* **121**, 139–156 (2023).
30. Castellazzi, G. et al. A machine learning approach for the differential diagnosis of alzheimer and vascular dementia fed by MRI selected features. *Front. Neuroinformatics* **14**, 25 (2020).
31. Yao, A. D., Cheng, D. L., Pan, I. & Kitamura, F. Deep learning in neuroradiology: a systematic review of current algorithms and approaches for the new wave of imaging technology. *Radiol. Artif. Intell.* **2**, e190026 (2020).
32. Kundu, S. AI in medicine must be explainable. *Nat. Med.* **27**, 1328–1328 (2021).
33. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer Nature, 2019).
34. Barredo Arrieta, A. et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).
35. Samek, W. & Müller, K.-R. Towards explainable artificial intelligence. in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (eds. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. &

Müller, K.-R.) 5–22 (Springer International Publishing, Cham, 2019). https://doi.org/10.1007/978-3-030-28954-6_1.

36. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. Preprint at https://doi.org/10.48550/arXiv.1312.6034 (2014).
37. Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS One* **10**, e0130140 (2015).
38. Martin, S. A., Townend, F. J., Barkhof, F. & Cole, J. H. Interpretable machine learning for dementia: a systematic review. *Alzheimer's Dement.* **19**, 2135–2149 (2023).
39. Böhle, M., Eitel, F., Weygandt, M. & Ritter, K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Front. Aging Neurosci.* **11**, 194 (2019).
40. Wang, D. et al. Deep neural network heatmaps capture Alzheimer's disease patterns reported in a large meta-analysis of neuroimaging studies. *NeuroImage* **269**, 119929 (2023).
41. Dyrba, M. et al. Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps evaluation in Alzheimer's disease. *Alzheimer's Res. Ther.* **13**, 191 (2021).
42. Dyrba, M., et al. 307–312 (Springer Fachmedien, Wiesbaden, 2020). https://doi.org/10.1007/978-3-658-29267-6_68.
43. Kohlbrenner, M. et al. Towards best practice in explaining neural network decisions with LRP. In: *Proc. International Joint Conference on Neural Networks (IJCNN)* 1–7. https://doi.org/10.1109/IJCNN48605.2020.9206975 (2020).
44. Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F. & Fox, P. T. Activation Likelihood Estimation meta-analysis revisited. *Neuroimage* **59**, 2349–2361 (2012).
45. Briechle, K. & Hanebeck, U. D. *Template Matching Using Fast Normalized Cross-Correlation*. in (eds. Casasent, D. P. & Chao, T.-H.) 95–102 (Orlando, FL, 2001). https://doi.org/10.1117/12.421129.
46. Eitel, F. & Ritter, K. Testing the robustness of attribution methods for convolutional neural networks In MRI-based Alzheimer's disease classification. In (eds. Suzuki, K. et al.) *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support* 3–11 (Springer International Publishing, Cham, 2019). https://doi.org/10.1007/978-3-030-33850-3_1.
47. Erasmus, A., Brunet, T. D. P. & Fisher, E. What is interpretability? *Philos. Technol.* **34**, 833–862 (2021).
48. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**, e745–e750 (2021).
49. Amann, J. et al. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digit. Health* **1**, e0000016 (2022).
50. Adebayo, J. et al. Sanity Checks for Saliency Maps. *arXiv:1810.03292 [cs, stat]* (2020).
51. Kindermans, P.-J. et al. The (Un)reliability of saliency methods. in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (eds. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R.) 267–280 (Springer International Publishing, Cham, 2019) https://doi.org/10.1007/978-3-030-28954-6_14.
52. Nie, X. et al. 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**, 399 (2017).
53. Poulin, S. P., Dautoff, R., Morris, J. C., Barrett, L. F. & Dickerson, B. C. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry Res.* **194**, 7–13 (2011).
54. Van Hoesen, G. W., Augustinack, J. C., Dierking, J., Redman, S. J. & Thangavel, R. The parahippocampal gyrus in Alzheimer's disease. Clinical and preclinical neuroanatomical correlates. *Ann. N.Y. Acad. Sci.* **911**, 254–274 (2000).
55. Visser, P. J. et al. Medial temporal lobe atrophy and memory dysfunction as predictors for dementia in subjects with mild cognitive impairment. *J. Neurol.* **246**, 477–485 (1999).
56. Echávarri, C. et al. Atrophy in the parahippocampal gyrus as an early biomarker of Alzheimer's disease. *Brain Struct. Funct.* **215**, 265–271 (2011).
57. Dickerson, B. C. et al. Alzheimer-signature MRI biomarker predicts AD dementia in cognitively normal adults. *Neurology* **76**, 1395–1402 (2011).
58. Rafii, M. S. & Aisen, P. S. Detection and treatment of Alzheimer's disease in its preclinical stage. *Nat. Aging* **3**, 520–531 (2023).
59. Frisoni, G. B. et al. Dementia prevention in memory clinics: recommendations from the European task force for brain health services. *Lancet Reg. Health – Europe* **26**, 100576 (2023).
60. de Vugt, M. E. & Verhey, F. R. J. The impact of early dementia diagnosis and intervention on informal caregivers. *Prog. Neurobiol.* **110**, 54–62 (2013).
61. Leonardsen, E. H. et al. Deep neural networks learn general and clinically relevant representations of the ageing brain. *NeuroImage* **256**, 119210 (2022).
62. Mårtensson, G. et al. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Med. Image Anal.* **66**, 101714 (2020).
63. Herzog, C. On the ethical and epistemological utility of explicable AI in medicine. *Philos. Technol.* **35**, 50 (2022).
64. Petersen, R. C. et al. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology* **74**, 201–209 (2010).
65. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W. & Smith, S. M. FSL. *NeuroImage* **62**, 782–790 (2012).
66. Ségonne, F. et al. A hybrid approach to the skull stripping problem in MRI. *Neuroimage* **22**, 1060–1075 (2004).
67. Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* **5**, 143–156 (2001).
68. Gong, W., Beckmann, C. F., Vedaldi, A., Smith, S. M. & Peng, H. Optimising a simple fully convolutional network for accurate brain age prediction in the PAC 2019 challenge. *Frontiers in Psychiatry* **12**, 627996 (2021).
69. Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A. & Smith, S. M. Accurate brain age prediction with lightweight deep neural networks. *Med. Image Anal.* **68**, 101871 (2021).
70. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. 19 (2015).
71. Chollet, F. & others. Keras. https://github.com/fchollet/keras (2015).
72. Smith, L. N. *Cyclical Learning Rates for Training Neural Networks 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 464-472 (2017).
73. Guillemot, M., Heusele, C., Korichi, R., Schnebert, S. & Chen, L. Breaking Batch Normalization for better explainability of Deep Neural Networks through Layer-wise Relevance Propagation. *arXiv:2002.11018 [cs, stat]* (2020).
74. Montavon, G., Binder, A., Lapuschkin, S., Samek, W. & Müller, K.-R. Layer-wise relevance propagation: an overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (eds. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R.) 193–209 (Springer International Publishing, Cham, 2019). https://doi.org/10.1007/978-3-030-28954-6_10.
75. Brett, M. et al. nipy/nibabel: 3.2.2. Zenodo https://doi.org/10.5281/zenodo.6617121 (2022).
76. Laird, A. R., Lancaster, J. L. & Fox, P. T. BrainMap: the social evolution of a human brain mapping database. *Neuroinformatics* **3**, 65–78 (2005).

77. Desikan, R. S. et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral-based regions of interest. *Neuroimage* **31**, 968–980 (2006).
78. Samek, W., Binder, A., Montavon, G., Lapuschkin, S. & Müller, K.-R. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Trans. Neural Netw. Learn. Syst.* **28**, 2660–2673 (2017).
79. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
80. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. 92–96 (Austin, Texas, 2010). https://doi.org/10.25080/Majora-92bf1922-011.
81. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
82. Davidson-Pilon, C. Lifelines: survival analysis in Python. *J. Open Source Softw.* **4**, 1317 (2019).

Acknowledgements

This work was funded by the UiO:LifeScience Convergence Environment (project: 4MENT), the Research Council of Norway (302854), and the European Research Council under the European Union's Horizon 2020 Research and Innovation program (802998). The Southern and Eastern Norway Regional Health Authority supported the study through funding for K.P. but was not involved in conducting the study or in the preparation of the manuscript. T.W. acknowledges funding from the German Research Foundation (DFG) Emmy Noether: 513851350. The work was performed on the Service for Sensitive Data (TSD) platform, owned by the University of Oslo, operated, and developed by the TSD service group at the University of Oslo IT Department (USIT). We also acknowledge the computational resources provided by UNINETT Sigma2—the National Infrastructure for High-Performance Computing and Data Storage in Norway—with project no. (nn9769k/ns9769k).

Author contributions

Conceptualization: E.H.L., T.W., L.T.W., Y.W. Data curation: K.P., E.W., G.S. Formal analysis: E.H.L. Funding acquisition: O.A.A., Y.W. Investigation: E.H.L., J.M.R., D.V.P., T.K., A.M., O.A.A., T.W., L.T.W., Y.W. Methodology: E.H.L., E.G., N.D., T.S., Ø.S., T.W., L.T.W., Y.W. Project administration: G.S., O.A.A., L.T.W., Y.W. Software: E.H.L. Supervision: T.W., L.T.W., Y.W. Validation: E.H.L. Visualization: E.H.L. Writing—original draft: E.H.L., T.W., L.T.W., Y.W. Writing—review & editing: K.P., E.G., N.D., T.S., J.M.R., D.V.P., Ø.S., T.K., E.W., A.M., G.S., O.A.A.

Competing interests

K.P. report work with Roche BN29553 and Novo Nordisk NN6535-4730 trials; All other authors declare that they have no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01123-7.

Correspondence and requests for materials should be addressed to Esten H. Leonardsen.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.