An Unsupervised XAI Framework for Dementia Detection with Context Enrichment

Devesh Singh[1,36], Yusuf Brima[1], Fedor Levin[1], Martin Becker[2], Bjarne Hiller[2], Andreas Hermann[1,3], Irene Villar-Munoz[4,5], Lukas Beichert[6], Alexander Bernhardt[7,8], Katharina Buerger[7,9], Michaela Butryn[10,11], Peter Dechent[31], Emrah Düzel[10,11], Michael Ewers[7,9], Klaus Fliessbach[12,13], Silka D. Freiesleben[4,5,37], Wenzel Glanz[10,11], Stefan Hetzer[32], Daniel Janowitz[9], Doreen Görß[1,14], Ingo Kilimann[1,14], Okka Kimmich[12], Christoph Laske[15,16], Johannes Levin[7,8,17], Andrea Lohse[39], Falk Luesebrink[10], Matthias Munk[15,18], Robert Perneczky[7,17,19,20], Oliver Peters[4,5], Lukas Preis[5,37], Josef Priller[4,5,21,22,38], Johannes Prudlo[1,23], Diana Prychynenko[4,5], Boris S. Rauchmann[19,24,25], Ayda Rostamzadeh[26], Nina Roy-Kluth[12], Klaus Scheffler[34], Anja Schneider[12,13], Louise Droste zu Senden[37], Björn H. Schott[27,28,35], Annika Spottke[12,33], Matthis Synofzik[6,15], Jens Wiltfang[27,28,29], Frank Jessen[12,26,30], Marc-André Weber[36], Stefan J. Teipel[1,14], Martin Dyrba*[1], and the ADNI[#], AIBL[+], FTLDNI[x] study groups

[1] German Center for Neurodegenerative Diseases (DZNE), Rostock/Greifswald, Germany
[2] Institute for Visual and Analytic Computing, University of Rostock, Germany
[3] Translational Neurodegeneration Section "Albrecht Kossel", Department of Neurology, University Hospital Rostock, Rostock, Germany
[4] German Center for Neurodegenerative Diseases (DZNE), Berlin, Germany
[5] Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Department of Psychiatry and Neuroscience, Hindenburgdamm 30, 12203 Berlin, Germany.
[6] Division Translational Genomics of Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research and Center of Neurology, University of Tübingen, Tübingen, Germany.
[7] German Center for Neurodegenerative Diseases (DZNE), Munich, Germany
[8] Department of Neurology, University Hospital of Munich, Ludwig-Maximilians-Universität (LMU) Munich, Munich, Germany
[9] Institute for Stroke & Dementia Research, University Hospital, LMU Munich, Germany
[10] German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany
[11] Institute for Cognitive Neurology and Dementia Research, Faculty of Medicine, University Hospital Magdeburg, Magdeburg, Germany
[12] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[13] Department for Neurodegenerative Diseases and Gerontopsychiatry, University of Bonn, Bonn, Germany
[14] Department of Psychosomatic Medicine, Rostock University Medical Center, Rostock, Germany
[15] German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany
[16] Section for Dementia Research, Hertie Institute for Clinical Brain Research, Department of Psychiatry and Psychotherapy, University Hospital Tübingen, Tübingen, Germany
[17] Munich Cluster for Systems Neurology (SyNergy), Munich, Germany
[18] Department of Psychiatry and Psychotherapy, University Hospital Tübingen, Tübingen, Germany
[19] Department of Psychiatry and Psychotherapy, University Hospital, LMU Munich, Munich, Germany
[20] Ageing Epidemiology Research Unit, School of Public Health, Faculty of Medicine, Imperial College London, London, United Kingdom
[21] Department of Psychiatry and Psychotherapy, School of Medicine and Health, Technical University of Munich, Germany
[22] University of Edinburgh and UK Dementia Research Institute, Edinburgh, United Kingdom
[23] Department of Neurology, University Medical Centre, Rostock, Germany
[24] Sheffield Institute for Translational Neuroscience, The University of Sheffield, Sheffield, United Kingdom
[25] Department of Neuroradiology, University Hospital, LMU Munich, Germany

[26] Department of Psychiatry, University of Cologne, Medical Faculty, Cologne, Germany
[27] German Center for Neurodegenerative Diseases (DZNE), Goettingen, Germany
[28] Department of Psychiatry and Psychotherapy, University Medical Center Goettingen, Goettingen, Germany
[29] Neurosciences and Signaling Group, Institute of Biomedicine (iBiMED), Department of Medical Sciences, University of Aveiro, Aveiro, Portugal
[30] Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases, Faculty of Medicine, University of Cologne, Cologne, Germany
[31] MR-Research in Neurosciences, Department of Cognitive Neurology, University Medical Center Goettingen, Goettingen, Germany
[32] Berlin Center for Advanced Neuroimaging, Charité University Medicine Berlin, Berlin, Germany
[33] Department of Neurology, University Hospital Bonn, Bonn, Germany
[34] Department for Biomedical Magnetic Resonance, University of Tübingen, Tübingen, Germany
[35] Department of Psychiatry and Psychotherapy, University Hospital Magdeburg, Magdeburg, Germany
[36] Institute of Diagnostic and Interventional Radiology, Pediatric Radiology and Neuroradiology, University Medical Centre Rostock, Rostock, Germany
[37] Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Experimental and Clinical Research Center (ECRC), Lindenberger Weg 80, 13125 Berlin, Germany.
[38] German Center for Mental Health (DZPG), Munich, Germany
[39] Department of Psychiatry and Psychotherapy, Charité – University Medicine Berlin, Berlin, Germany

[#] Data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf
[+] Data obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organization (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed. AIBL researchers are listed at www.aibl.csiro.au.
[x] Data obtained from the Frontotemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI) database. The investigators at NIFD/FTLDNI contributed to the design and implementation of FTLDNI and/or provided data.

***Correspondence:**
Martin Dyrba
martin.dyrba@dzne.de, +493814949482.
German Center for Neurodegenerative Diseases (DZNE), Site Rostock/Greifswald, c/o Zentrum für Nervenheilkunde, Gehlsheimer Str. 20, D-18147 Rostock, Germany.

Devesh Singh
devesh.singh@med.uni-rostock.de
Institute of Diagnostic and Interventional Radiology, Pediatric Radiology and Neuroradiology, University Medical Centre Rostock, Rostock, Germany.
Ernst-Heydemann-Str. 6, 18057 Rostock.

Keywords: Neurodegenerative diseases, Alzheimer's disease, frontotemporal dementia, magnetic resonance imaging, brain volumetry, explainable artificial intelligence (XAI), qualitative evaluation

Abstract

Introduction: Explainable Artificial Intelligence (XAI) methods enhance the diagnostic efficiency of clinical decision support systems by making the predictions of a convolutional neural network's (CNN) on brain imaging more transparent and trustworthy. However, their clinical adoption is limited due to limited validation of the explanation quality. Our study introduces a framework that evaluates XAI methods by integrating neuroanatomical morphological features with CNN-generated relevance maps for disease classification.

Methods: We trained a CNN using brain MRI scans from six cohorts: ADNI, AIBL, DELCODE, DESCRIBE, EDSD, and NIFD (N=3253), including participants that were cognitively normal, with amnestic mild cognitive impairment, dementia due to Alzheimer's disease and frontotemporal dementia. Clustering analysis benchmarked different explanation space configurations by using morphological features as proxy-ground truth. We implemented three post-hoc explanations methods: i) by simplifying model decisions, ii) explanation-by-example, and iii) textual explanations. A qualitative evaluation by clinicians (N=6) was performed to assess their clinical validity.

Results: Clustering performance improved in morphology enriched explanation spaces, improving both homogeneity and completeness of the clusters. Post hoc explanations by model simplification largely delineated converters and stable participants, while explanation-by-example presented possible cognition trajectories. Textual explanations gave rule-based summarization of pathological findings. Clinicians' qualitative evaluation highlighted challenges and opportunities of XAI for different clinical applications.

Conclusion: Our study refines XAI explanation spaces and applies various approaches for generating explanations. Within the context of AI-based decision support system in dementia research we found the explanations methods to be promising towards enhancing diagnostic efficiency, backed up by the clinical assessments.

Introduction

Alzheimer's disease (AD) is a significant and growing burden on global healthcare systems. Estimates suggest a global population of 152.8 million people living with dementia by 2050 [1], for which AD accounts for more than two-thirds of all cases [2]. The increasing prevalence of AD warrants the development of automated clinical decision support systems to improve the efficiency of diagnostic procedures and early disease detection. Deep learning (DL) has emerged as a promising tool in this context, offering state-of-the-art methods for the fast and robust analysis of complex neuroimaging data. However, the integration of DL into clinical practice is often hampered by a lack of transparency and interpretability of its predictions due to its 'black-box' nature[3].

Explainable Artificial Intelligence (XAI) methods offer a potential solution to this challenge by making DL models more human-comprehensible and interpretable. By explaining the decisions made by complex DL systems, XAI aims to bridge the gap between model predictions and clinical insights. This is particularly relevant under the European Union's General Data Protection Regulation (GDPR) and Artificial Intelligence (AI) act, which under the 'right to explanation' requires AI systems to provide explanations of their decision-making processes[4]. Other regulatory and government bodies have also advocated for similar AI capabilities, emphasizing the need for accountable and transparent AI systems in critical domains such as healthcare and medical decision-making[5–9].

Despite the advancements in XAI methods, a research gap remains in validating and assessing the quality of the explanations generated by AI systems[10–12]. Notably, it is time-consuming and expensive to consult experts to provide ‘ground-truth’ explanations and to evaluate the explanations generated by XAI methods. It also requires additional fine-tuning of the XAI methods to improve their correctness and suitability for a specific use case. Furthermore, with regards to the inference process, it is often unclear and depends on the user's experience level in determining—what needs to be explained, how, and in what detail[12].

Additional methods for generating explanations rely on sets of rules, to combine symbolic reasoning such as knowledge graphs, with neural models to provide human-understandable insights into AI decision-making. Rule-based explanations offer structured and semantic explanations which enhances transparency by relying upon a-priori domain knowledge [13,14]. Meanwhile, XAI methods that simplify model predictions reduce cognitive burden on the user by presenting the most useful information, and utilize methods such as network pruning or compression[15,16]. XAI methods like explanations-by-example describe model decision for a query sample by providing information about the most similar sample(s) from the training set [17].

Moreover, XAI-generated explanations, beyond the end use-case of providing insights into the AI system, may also be used as an additional information modality. One recent study highlighted a more separable arrangement of the participants in the ‘explanation space’, i.e., the vector space where the data points are arranged according to the explanation, compared to the original input space [18]. Specifically for convolutional neural networks (CNN) applications, explanation space refers to the representational space derived from the attribution (relevance) heatmaps or its feature-level representations. Some CNN studies further explored this ‘quantification gap’ of explanations and evaluated the overlap between visual explanations, i.e., attribution-based relevance maps and ground truth [19–21]. Other

studies have addressed the consistency and coherence of the relevance mapping techniques with respect to expert-created ground truth segmentations [22]. Some dementia studies quantified the voxel-level overlap between relevance maps and proxy ground truth maps, i.e., AD likelihood maps with relevant regions found through literature meta-analysis [23,24]. Notably, all of these studies have predominantly used supervised machine learning approaches that utilize expert-assessed, expensive-to-obtain, ground truth or other proxy measures to calculate these ground truths.

Here, we propose to extend the common understanding of the explanation space and present a framework that incorporates clinically relevant morphological features - such as cortical thickness and gray matter volumetry - combined with relevance maps to create a *context-enriched explanation space.* Previous multimodal studies have informed our additive approach to extend the explanation space. To date, studies of dementia detection that combine multiple data modalities, such as MRI and PET scans, often outperform unimodal models [25,26]. Previously developed disease state indices for differential diagnosis also utilized different information sources in a generalized additive model [27]. Taken together, we hypothesized that the explanation space that better separates, distributes, and structures disease pathology information would also be a more appropriate space for generating explanations. We assumed that the combination of these information sources would produce more contextually sensitive explanations, which in turn would improve the quality of the explanations. To examine these assumptions, we performed a clustering analysis to explore the distribution of participants in different explanation space configurations. Our unsupervised analysis was intended to act as a proxy measure of the utility of explanations, thus bypassing the dependence of a supervised analysis based on ground truth explanation labels. Our framework generates *post hoc* explanations for a CNN model detecting dementia diseases at both global level, i.e., subgroup membership, and local level, i.e., cognitive trajectory examples or textual prediction explanations.

To comprehensively evaluate our AI-based explanations, we also conducted a qualitative analysis with expert clinicians, assessing each explanation's usefulness for improving patient examinations. Through expert evaluation, we tackled a common issue in developing XAI prototypes for clinical decision support systems, i.e., the lack of user involvement in the co-development process[11,24,28]. Our overall aim was to advance the development and validation of robust XAI methods, and address the gaps in the evaluation of the explanations generated in the context of AD diagnosis.

Methods

The workflow of our study is schematically presented in Figure 1. Our framework provides several ways to generate post-hoc explanations for a CNN model trained to detect dementia diseases, including: i) global-level explanations, such as membership in the stable versus converter subgroups, and ii) local-level explanations for each individual prediction, such as ii-a) example-based explanations of cognitive trajectories or ii-b) textual explanation by pathology summarization. To evaluate clinical validity of the different types of AI-based explanations produced from our framework, we also conducted a qualitative analysis with a focus group of radiologists (N=4) and neurologists (N=2).

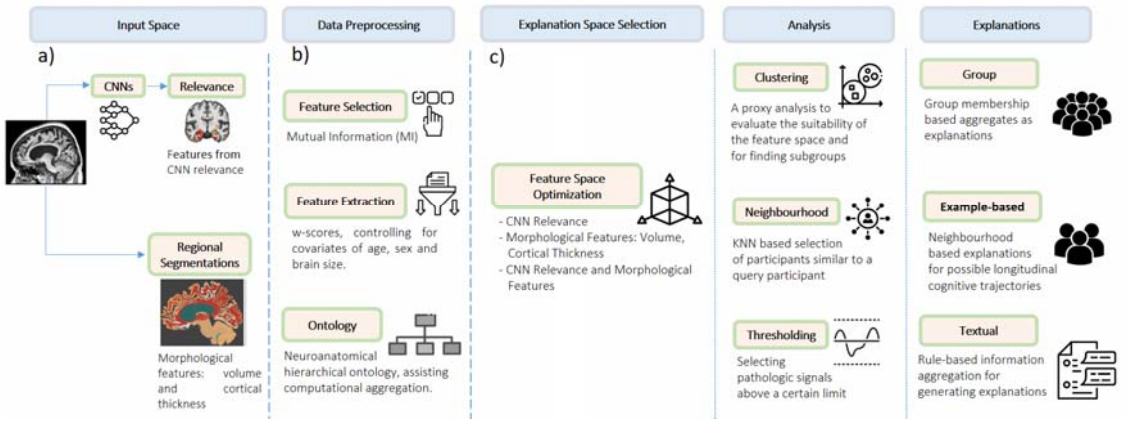


Figure 1: Study design for creating explanations for CNNs detecting dementia diseases from MRI scans. Here we illustrate a) the input space with trained CNN's relevance maps and brain segmentation, b) the preprocessing steps of - feature selection and extraction, and c) the explanation generation from the context-enriched explanation space and features extracted, utilizing different analysis methods.

Neuroimaging datasets

In this study, we collected T1-weighted brain MRI scans (N=3253) from publicly available neuroimaging data cohorts. The data scans were pooled from the following data cohorts: i) the Alzheimer's Disease Neuroimaging Initiative (ADNI), study phases ADNI2/GO and ADNI3, ii) the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) [29], iii) the DZNE Longitudinal Study on Cognitive Impairment and Dementia (DELCODE) [30], iv) the European DTI Study on Dementia (EDSD) [31], v) the DZNE Clinical Registry Study on Frontotemporal Dementia (DESCRIBE-FTD), and vi) the Frontotemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI) which is also known as Neuroimaging Initiative in Frontotemporal Dementia (NIFD).

It should be noted that mild cognitive impairment (MCI) can arise from various underlying conditions, however, the ADNI, AIBL, and DELCODE cohorts apply inclusion and exclusion criteria to focus primarily on amnestic MCI, i.e., individuals with memory impairment. Other conditions, such as depression or substance abuse, were excluded. Summary statistics for the data used are presented in Table 1. See Supplementary Table A.1.1 for statistics reported for each cohort.

These datasets were initially intensity corrected using the N4ITK algorithm for bias field correction. Then, HD-BET was applied for skull stripping[32]. ANTs SyNQuick registration tool was used to linearly warp all images to the MNI reference spare and ANTs AtroposN4 was applied for tissue segmentation into CSF, white matter, and gray matter. The normalized gray matter maps served as input for the CNN model. Subsequently, based on the native space images, FastSurfer version 2.0.4 was used to perform brain segmentation into 100 anatomically defined regions-of-interest (ROIs) and cortical surface reconstructions to measure regional volume and average cortical thickness[33,34]. FastSurfer follows the Desikan–Killiany–Tourville (DKT) atlas protocol for producing the anatomical segments[35,36]. Finally,

the linear deformations from ANTs were applied to the FastSurfer segmentation maps to extract CNN relevance scores per region.

TABLE 1
Sample Statistics Per Disease Diagnosis Stage and Subtype

	CN	**MCI**	**AD**	**FTD**
Sex (M/F)	732/963	448/369	266/282	112/81
Age	70.15 ± 7.58	72.76 ± 7.38	74.14 ± 7.69	63.35 ± 7.92
MMSE	29.02 ± 1.52	27.36 ± 2.3	22.15 ± 4.18	23.56 ± 7.10

CN: cognitively normal, MCI: mild cognitive impairment, AD: dementia due to Alzheimer's disease (the typical presentation), FTD: Frontotemporal dementia where phenotypes include behavioral variant, semantic variant, and progressive nonfluent aphasia, MMSE: mini-mental state examination score, F: female, M: male. Numbers are reported as (mean ± s.d.).

Relevance segmentation, aggregation, and abstraction

We trained a multi-class CNN model based on the DenseNet architecture as the backbone [37,38]. A three-way classification setup was used that classified cognitively normal (CN), Alzheimer's disease (AD; pooled patients with dementia due to AD and patients with amnestic mild cognitive impairment (MCI)), and phenotypes of frontotemporal dementia (FTD) - including behavioral variant (bvFTD), semantic dementia (SD), and progressive nonfluent aphasia (PNFA) - participants. See supplementary section A.2 for further model training details.

We used the Layer-wise Relevance Propagation (LRP) attribution method that generates a heatmap of input regions that the model found useful for differentiating each class [39]. We chose the composite alpha-beta LRP rule as it highlights relevant input features with high specificity and avoids the dispersed distribution of relevance scores across multiple input regions [40–43], unlike other methods such as GradCam [44] or Occlusion Maps [45]. While the LRP rule is sensitive to the parametric choice of alpha and beta hyper-parameters, it does not require defining a base image used by methods like the Integrated Gradients [43,46]. LRP relevance maps have also been used in several other previous dementia studies [47–49].

The 3D LRP relevance maps were segmented using region of interest (ROI) segmentations generated by the FastSurfer segmentation tool. Within each ROI, we calculated the relevance density, i.e., the relevances were summed up and divided by the volume of the respective ROI. The relevance density metric has previously been found to be better associated with disease features than the total sum or other relevance aggregation mechanisms [50,51]. Based on the hierarchical ontology structure developed in our previous work[52], relevance was aggregated and summarized across different levels of neuroanatomical abstraction, such as lobes and hemispheres; which added 24 higher-order (parent-level) aggregation concepts.

Data preprocessing

Feature extraction: W-scores were calculated for each pathology feature, region, and

participant, which quantified the relative deviations from the normative expectations. W-scores are an extension of Z-scores that adjust for covariates; in our study, we controlled for age, sex, brain size, and magnetic field strength, as these variables are widely known to influence brain volume and cortical thickness measures [53–55].

$$W_{j\in S;\, i\in ROIs} = \frac{Observed\ Feature_{i,j} - Expected\ Feature_{i,j}}{Std.Dev\ (residuals_control_{i,j})} \qquad (1)$$

here *features*, i.e., S = {CNN relevance, cortical thickness, volume}. The expected feature is the prediction from a linear regression model that accounts for the confounding covariates and was trained only on the cognitively normal control participants. The $residuals_controls_{i,j}$ are the residuals from the cognitively normal controls.

Feature selection: W-score features per region (X), i.e., the CNN's relevances, the volumetric measure, and the cortical thickness measure, were compared with the disease diagnosis labels (Y), by calculating the mutual information I(X, Y) between them, defined as:

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\ \log\frac{p(x,y)}{p(x)p(y)} \qquad (2)$$

where in equation 2, p(X, Y) is the joint probability distribution for random variables X and Y, while p(X) is the marginal probability distribution for the random variable X. Mutual information quantifies the dependence between individual features and the disease diagnosis, and lies between [0, +∞) where a mutual information of 0 indicates two independent variables. Mutual information (MI) is one of the most widely used methods for feature selection in machine learning, as it effectively quantifies the dependency between features and target variables. Its ability to capture both linear and non-linear relationships makes it particularly valuable in high-dimensional data analysis [56]. The features with mutual information above the threshold of 0.1, chosen heuristically, were selected for further analysis.

Enriched explanation space

We set up a clustering analysis as a proxy measure to calibrate the suitability of the various explanation spaces. Different variations of the explanation space were explored: a) including only the relevance features from the CNN, i.e., the basic *explanation spac*e, b) including only the morphological features - volumetry and/or cortical thickness, and c) including both the relevance features from the CNN and the morphological features, i.e., the *context-enriched explanation space*. The derived clusters were evaluated using broadly two sets of metrics. First, the external validation metrics, measuring agreement between the predicted cluster labels and the ground truth disease diagnosis - homogeneity, completeness and v-measure [57], adjusted mutual information [58], adjusted rand score [59,60], and Fowlkes score [61]. V-measure is the harmonic mean of the homogeneity and completeness scores. Second, the internal validation metrics, measuring the separation between the clusters within the space and requiring no external ground truth labels - average silhouette coefficient [62] and Davies Bouldin score [63].

Deriving Explanations from the Enriched Explanation Space

Group-level explanations: We utilized the agglomerative hierarchical clustering with Ward's linkage to create the group-level, ***feature simplification*** explanations. The hierarchical clustering separates different subgroups of participants. Ward linkage criterion minimizes the within-cluster variance [64] and has been found useful in other dementia studies [65,66]. We chose the Euclidean distance as the metric for calculating the distance between the clusters in the explanation space. For the participants grouped within a cluster, a repeated-measures linear mixed-effect model was fitted to cognition trajectories for Mini-Mental State Examination (MMSE) and global Clinical Dementia Rating (CDR) scores. The models included fixed effects for age at baseline, sex, and interaction terms between baseline cognitive diagnosis (cognitively normal - CN, mild cognitive impairment - MCI, or Alzheimer's disease dementia - AD), cluster index, and time (months). We also specified random intercepts for each participant to account for individual variability in baseline cognition. We additionally performed the Kaplan-Meier survival analysis to compare the time to dementia conversion between the clusters. The conversion event was marked by the change of the CDR global score, i.e., conversion from unimpaired cognition (CDR=0) to MCI due to AD (CDR= 0.5), and conversion from MCI (CDR=0.5) to mild AD dementia (CDR=1), beyond which any further increase in CDR score (>1) was not considered. For each participant, longitudinal data was included for up to six years, and participants were right-censored when they did not convert.

Example-based explanations: The example-based explanations were generated using a meta-classifier abstracting over the details of the explanation space and presenting the likely cognition progression trajectories. The use of a simple meta-classifier is a common practice in building decision support systems to assist experts [67]. We chose the k nearest neighbor (KNN) classifier as the meta-classifier. The size of the neighborhood k=10 was set heuristically. The nearest neighbor for a query sample in the enriched explanation space represents a small group of examples, i.e., participants with similar pathology. Hence, using this notion of similarity, we then present exemplary cognition trajectories of the neighbors, here, the MMSE and CDR scores obtained from follow-up visits up to six years, where available.

Textual explanations: Previous studies that applied knowledge-based approaches [68–70] have established a more structured and knowledge-engineered usage of clinical information for decision support. We previously created a computational neuroanatomy ontology that enhanced the aggregation of pathologic information[52]. Based on this framework, we developed a post-hoc, ***rule-based*** explanation method where the information sources, here CNN relevances, volume, and cortical thickness features, could be integrated to generate ***textual explanations*** for a single participant.

The ontology's hierarchical structure opens up space for the computational aggregation of different pathological features, at multiple abstraction levels. More importantly, the structured setup also allows for more sophisticated logical reasoning, for example, the inclusion or exclusion of entities. We developed a rule-based method that dynamically chooses anatomical entities for which all three (logical and) pathological features indicated abnormal levels, more specifically, the w-score exceeding 2 standard deviations from the norm. In cases where many regions at a lower hierarchy were selected, then only the higher hierarchy region was selected for presentation. This reduces the load of information presented to the end user.

The selected regions were reported to the clinical users as template-based textual explanations. When the average pathology w-score across all applicable features remains between 2 and 3, a region is classified as ‘mild’ pathology. Scores between 3 and 4 indicate ‘moderate’ pathology, while, scores exceeding 4 standard deviations indicate ‘strong’ pathology. This threshold-based logic was empirically derived through the analysis of w-scores in our dataset. This logic facilitates the categorization of average pathology severity for each neuroanatomic region.

Qualitative interviews with the experts

We interviewed neurologists (N=2) working in a memory clinic and radiologists (N=4) with an average of 10+ years of experience. This ensured expert feedback while avoiding input from newly trained professionals. The semi-structured interview was opened with introducing the experts to the high-level perspectives present in explainable artificial intelligence (XAI) field, i.e., the different values and goals, such as causality, confidence, informativeness, and trustworthiness, pursued by the XAI methods [71]. They were also introduced to the taxonomies for grouping various XAI methods [16].

Furthermore, the semi-structured interview consisted of the following steps: introducing a case study sample, the various types of explanations generated for the sample presented one by one, and prompting the experts regarding the different usability aspects of the explanations. Figure 2 illustrates the process of semi-structured expert interviews. All interviews were carried out in accordance with relevant guidelines and regulations. For further details, please refer to the included ethics statement.

This order was chosen to let the experts present their opinions about each method’s value for a use-case, building on the opinions to define the more concrete strengths and challenges, and eventually to state the possible future works for the explanation types. Approximately an hour to an hour and a half was spent to go through all the explanation types and discuss them individually. The focus group interviews were conducted one-on-one or with at most two experts together. In total, 4 interview sessions were conducted. The interviews were conducted between February and March 2025. The focus group interviews helped us qualitatively evaluate the opportunities and challenges of applying XAI methods in clinical decision systems.

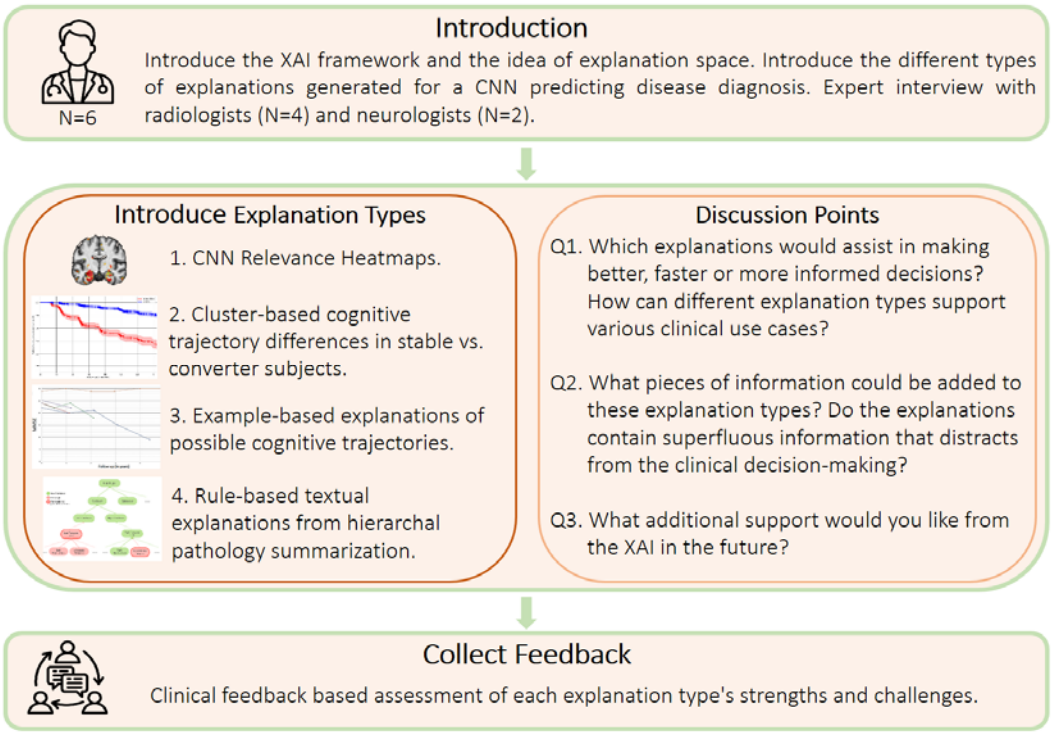


Figure 2. Semi-structured expert interview flowchart. We collected the clinical feedback on different aspects of XAI explanation types to assist clinical decision-making. The interviews approximately lasted for an hour. Clinical experts highlighted which explanations enhance the decision-making process by making CNN more adoptable, what information should be added or removed, and future improvements for XAI support. These interviews served as a basis for qualitatively evaluating the opportunities and challenges of applying XAI methods.

Results

Explanation Space Selection and Subcluster Identification

From our CNN model trained for three-way classification between CN, AD, and FTD, the CN node was chosen to acquire the relevance, as the relevance scores generated from it represent the deviation from the normal group. This means that the relevance of an input voxel reflects its contribution to a subject being classified (or not classified) as cognitively normal, thereby highlighting patterns associated with pathological aging.

After applying a heuristically set threshold of 0.1 on the mutual information criterion, 81 features remained. The selected features included K=19 (23.5%) relevance features, K=46 (56.7%) volume features, and K=16 (19.8%) cortical thickness features. Notably, for the cortical regions left entorhinal, left inferior and superior temporal, and left temporal lobe, as well as for the subcortical regions left hippocampus, left putamen, and left amygdala, all respective features had mutual information above the threshold. See Supplementary section A.3 for more information.

Using the selected features, agglomerative clustering with ward linkage was conducted to compare different variations of the explanation space while separating the AD and CN participants. The results of the clustering analysis are presented in Table 2. According to the V-measure, the enriched explanation space provided the highest score of 0.43.

TABLE 2
Clustering Performance Across Explanation Spaces

Explanation space	**V-measure**	**AMI**	**Homogeneity score**	**Completeness score**	**ARI**	**FMI**	**Silhouette score**	**DBI**
Relevance (Rel)	0.39	0.39	0.34	0.47	0.51	0.85	**0.48**	**1.01**
Volumetry (Vol)	0.13	0.13	0.15	0.12	0.02	0.59	0.20	1.56
Cortical Thickness (CortThk)	0.30	0.30	0.31	0.29	0.44	0.78	0.38	1.03
Vol + CortThk	0.32	0.32	0.26	0.42	0.40	0.83	0.35	1.21
Rel + Vol + CortThk	**0.43**	**0.43**	**0.40**	**0.48**	**0.57**	**0.86**	0.35	1.37

Explanations spaces: the *basic explanation space* only includes the relevance features, while the *context-enriched explanation space* includes volumetry and cortical thickness features with the relevance features. The metrics include V-measure, Adjusted Mutual Information (AMI), Homogeneity Score, Completeness Score, Adjusted Rand Score (ARI), Fowlkes-Mallows Index (FMI), Silhouette Score, and Davies-Bouldin Index (DBI). All measures ranged between [0,1], except - ARI [-0.5,1], Silhouette Score [-1,1], and DBI [0,∞). Higher values (except for DBI, where lower is better) indicate better clustering quality.

Figure 3 illustrates the cluster map of the dataset in the context-enriched explanation space, providing a hierarchical visualization (on the Y-axis) of relationships between data points. From the heatmap intensity, we found a relative segregation of the disease diagnoses between the two main clusters, where darker regions on the heatmap represent more pathologic patients being clustered together. The number of clusters was heuristically set to two to balance between cohesion and separation while maintaining clinical interpretability. Although we explored 3-4 clusters scenarios based on the splits found via the dendrogram, they did not provide additional meaningful insights. While a relatively homogeneous cluster with FTD patients emerged (in 3 cluster scenario), it offered limited new information with respect to the further explanations drawn from the framework.

As visualized by the pie charts in the left dendrogram in Figure 3, one cluster mainly consists of healthy controls or participants with low levels of pathology, this cluster is here on termed as the *stable* cluster. The second cluster consists of participants with more advanced pathology, i.e, a high amount of atrophy in dementia patients, this cluster is here on termed as

the *converter* cluster. Table 3 presents the confusion matrix comparing clustering outcomes with ground truth labels of participants' baseline disease diagnosis.

TABLE 3
Confusion Matrix for the Clustering Outcome

Clusters / Disease Diagnosis	**CN**	**MCI**	**AD**	**FTD**
Stable Cluster (N=2084)	1383 (66.4%)	612 (29.4%)	63 (3%)	26 (1.2%)
Converter Cluster (N=1318)	219 (16.6%)	503 (38.2%)	462 (35.1%)	134 (10.2%)

The proportion of the participant's baseline disease diagnosis stage and subtype within the found clusters - termed *stable* or *converter* - are presented as percentage points. CN: cognitively normal, MCI: mild cognitive impairment, AD: dementia due to Alzheimer's disease (the typical presentation), FTD: Frontotemporal dementia where phenotypes include behavioral variant, semantic variant, and progressive nonfluent aphasia.

Using Fleiss' Kappa (κ), a score for inter-rater reliability was calculated to evaluate the stability (agreement) for the clustering-based binary classification task of stable vs. converter. We performed a 4-fold cross-validation, yielding a κ score 0.77, indicating substantial agreement, i.e., a relatively stable clustering outcome that is independent of the data folds used for initialization.

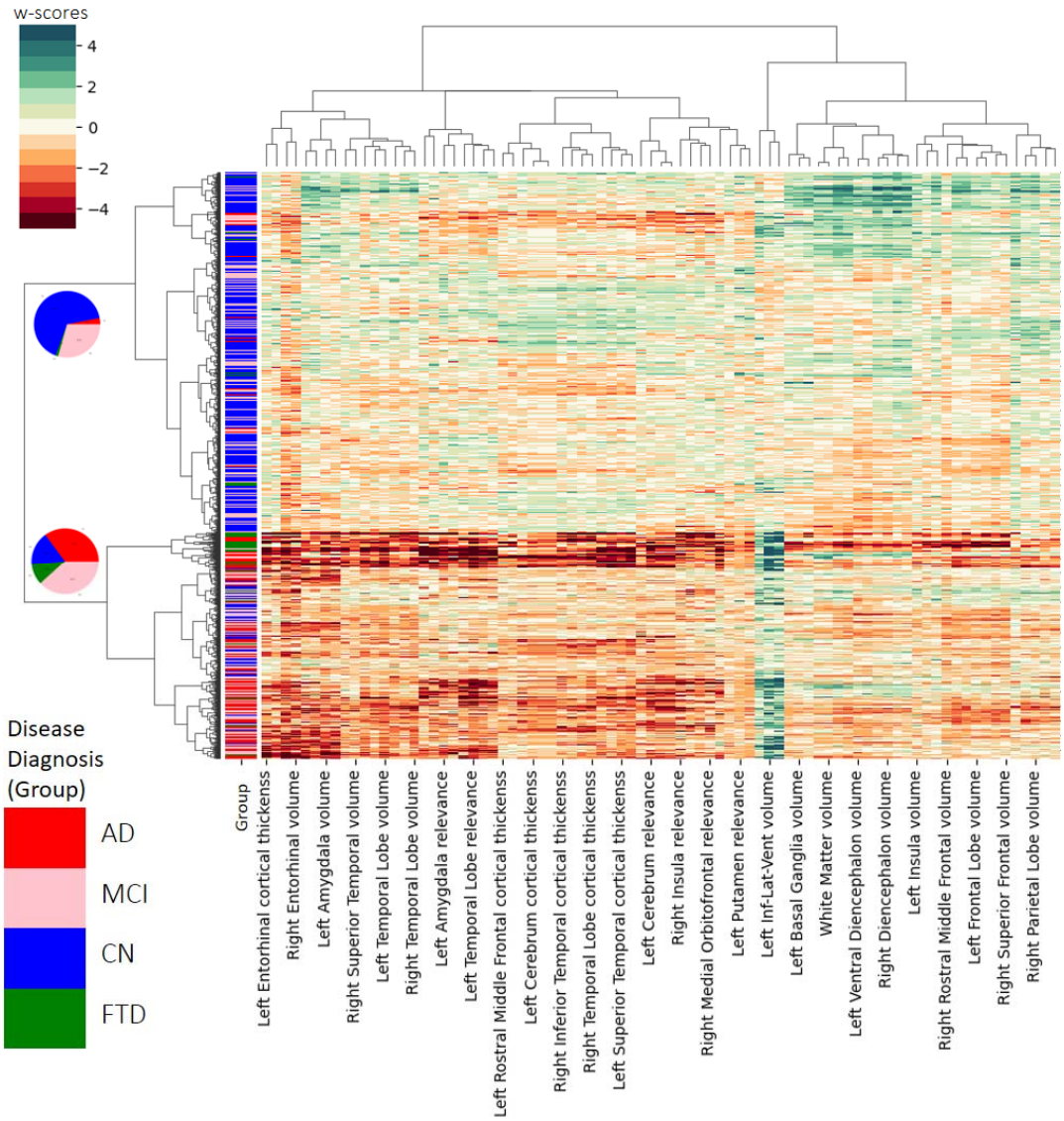


Figure 3. Cluster Map. Hierarchical clustering dendrogram (on the Y-axis) resulting from Ward's hierarchical clustering analysis of individual w-scores profiles of participants computed from three features - CNN relevances, volume, and cortical thickness measures. Four disease diagnoses were considered: cognitively normal (CN, color-coded as blue), mild cognitive impairment (MCI, color-coded as pink), dementia due to Alzheimer's disease (AD, color-coded as red), and frontotemporal dementia (FTD, color-coded as green). The pie charts visualize the relative homogeneity, with respect to the disease diagnoses, of the two clusters. The W-score features are visualized using a custom color scale to indicate the extent of the deviation, where a gradual intensification of color (either red or green) signifies increasing pathological observations. For a vector graphic rendering, please refer to the GitHub version of the plot.

Simplified, Group-Level Explanations

Based on the two clusters identified in the context-enriched explanation space, the longitudinal cognitive trajectories were explored as simplified explanations of the CNN model's predictions. Analysis of the longitudinal MMSE scores showed that the two

identified clusters separate participants which would remain relatively stable or decline at an accelerated rate, i.e., converters. Specifically, as seen in Figure 4, the MMSE score of the high-risk converter group exhibited cognitive decline at a rate of 0.54 points per year, whereas the low-risk stable group declined at a rate of 0.02 points per year. Details for the mixed-effects modeling and the analysis of the Clinical Dementia Rating (CDR) global score can be found in the supplementary section A.4.

From the Kaplan-Meier survival analysis, we also see a similar separation (Figure 5), where 80% of the participants in the stable cluster remain free of conversion for 60 months (or 5 years), with a conversion rate of approximately 3.3% per year, while in the converter cluster approximately 10% of the participants convert per year.

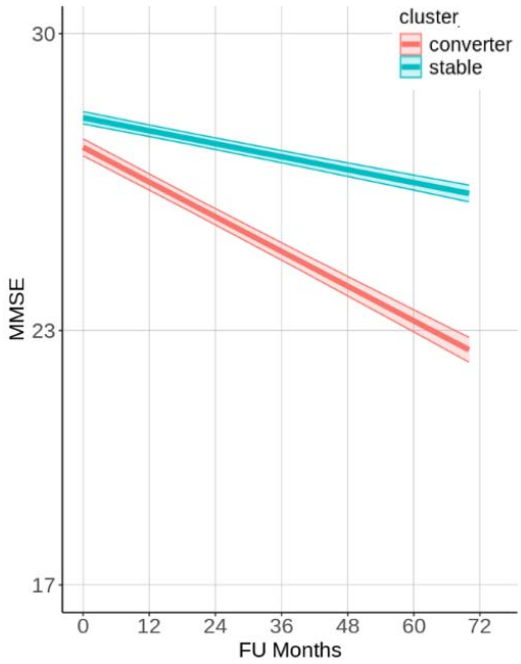


Figure 4: Longitudinal cognitive trajectories of different clusters of participants identified in the context-enriched explanation space. Values on MMSE cognitive test are obtained from mixed effects regression models which included the age, sex, baseline disease diagnosis, and the interaction between cluster membership and follow-up time in months (FU Months), as well as the interaction between baseline disease diagnosis and follow-up months. The model also included random intercepts for each participant to account for repeated measurements. The shaded regions represent 95% confidence intervals.

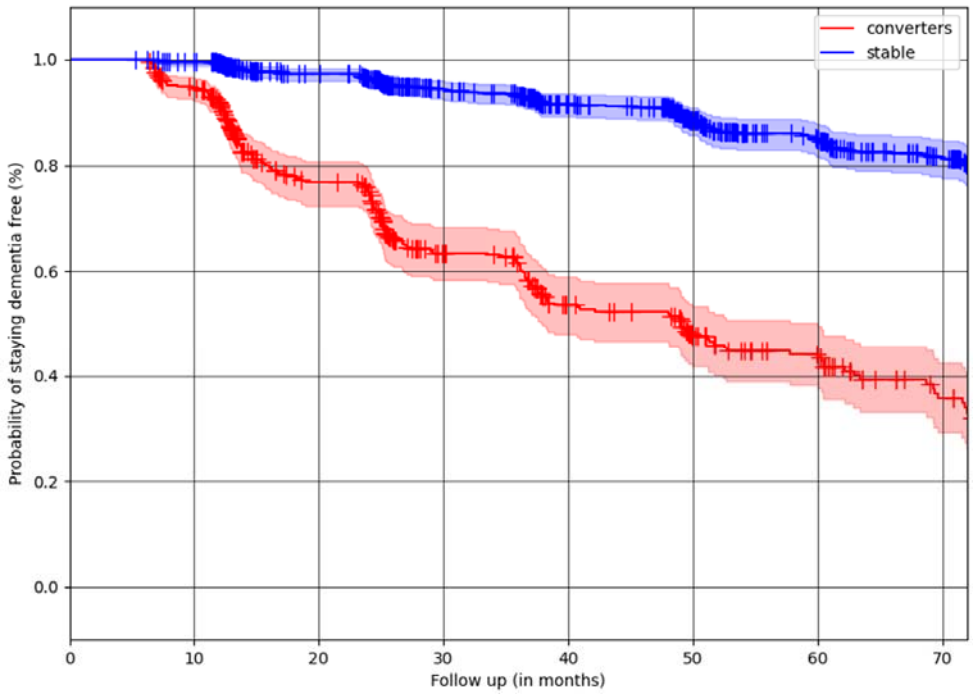


Figure 5: Kaplan-Meier curves illustrating the time to conversion across identified clusters. These survival curves represent the proportion of participants within each cluster who progressed either from CN to MCI and/or from MCI to dementia. Participants who did not develop dementia during the observation period were censored.

Explanation by examples

Within the context-enriched explanation space, the longitudinal cognitive trajectories of participants with similar pathology to a query sample illustrate the possible trajectories over 72 months (6 years). A k nearest neighbor (KNN) model was employed, to find the participants that present the most similar pathology, with the neighborhood window heuristically set to k=10.

Figure 6 shows the cognitive trajectories on the MMSE cognitive test, based on the nearest neighbors of one arbitrarily selected individual from the DELCODE data cohort with the clinical diagnosis of MCI. Supplementary Figure A.5.1 illustrates the explanation-by-example cognitive trajectories for the CDR score, and Supplementary section A.6 illustrates MMSE and CDR explanation-by-example plots where each cognitive trajectory is shown in a unique color for more detail. The participant is a woman in her late sixties who has received almost ten years of formal education, and who has a baseline MMSE score of 24. In the figure legend, the ten nearest neighbors (with their baseline diagnosis and pseudonymised patient ID) are listed in the order of increasing Euclidian distance from this query sample, i.e., the most similar participant in the dataset is listed first. In a clinical setting, the trajectories could serve as illustrations of possible future cognitive development for the query participant.

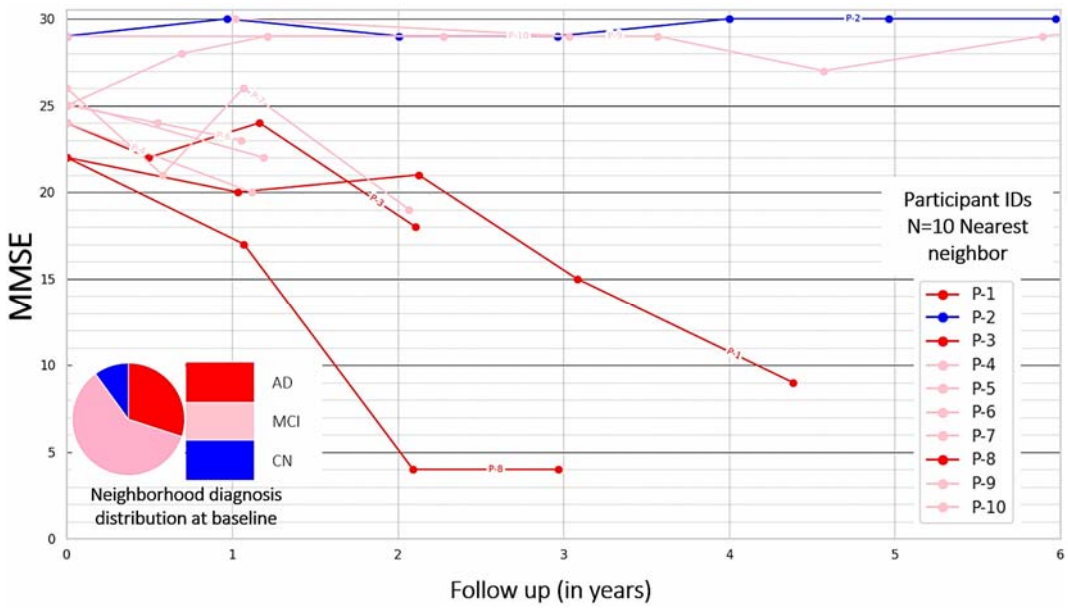


Figure 6: Explanation-by-examples: Within the context-enriched explanation space, the longitudinal cognitive trajectories of k=10 nearest neighbors of a query participant, from the DELCODE cohort, are shown. Scores on the MMSE cognitive test were observed on follow-up examinations for up to 6 years. Patient IDs of the nearest neighbors are pseudonymised, and the nearest neighbors are listed in the order of increasing Euclidian distance from the query sample, illustrating possible future cognition trajectories for the query participant. The cognition trajectories are additionally color-coded by the baseline disease diagnosis.

Rule-based textual explanations

The knowledge-driven, ontology-based explanation method generated structured textual explanations for individual participants. By combining CNN relevances, volumetric, and cortical thickness measures, the rule-based mechanism generated hierarchical summaries of neuroanatomical abnormalities, reducing redundancy by prioritizing higher-order regions. In Figure 7 (a), we see an illustration of the hierarchical selection mechanism. Figure 7 (b,c) illustrates the template-based textual report generated for the same query participant from the DELCODE cohort. Figure 7 (b) lists all the pathologic regions identified, including the left superior temporal, left middle temporal, left temporal lobe, left inferior temporal, and left inferior lateral ventricle. Meanwhile in Figure 7 (c) generated a template-based summary, presenting pathologic information specifically for the left temporal lobe and left inferior lateral ventricle.

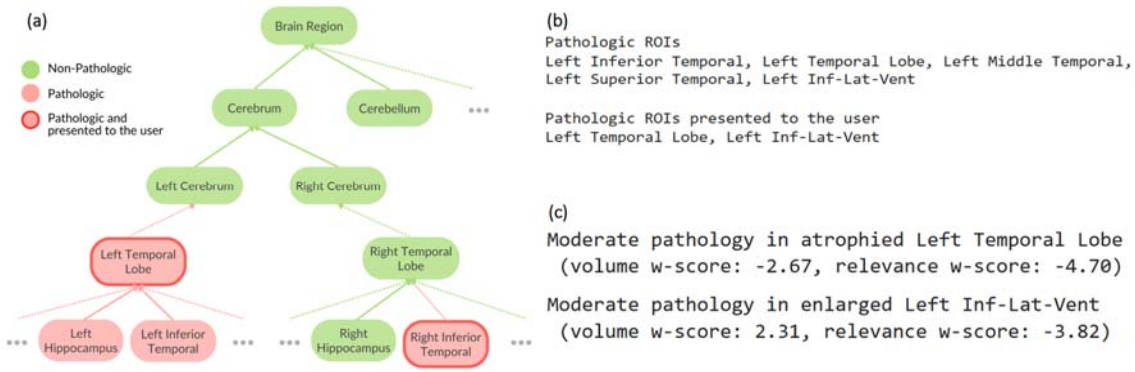


Figure 7: Rule-based textual explanation: (a) an hypothetical exemplary visual illustration of the rule-based mechanism of selecting neuroanatomical regions for which a pathologic threshold is reached for all the features - cortical thickness, volumetry, and relevance; and then narrowing down and optimizing the pathologic regions presented to the clinical user to reduce the information load. For a query participant from the DELCODE cohort, we show (b) a list of all the pathologic and presented regions, and (c) a template-based summary generated, listing w-scores from all relevant features.

Qualitative evaluation of the explanation types

Neurologists (N=2) from the memory clinic found the simplified, group-level explanations to be particularly useful, as an aid to communicate with other clinical experts. They described a scenario where their risk assessment capabilities of XAI methods could possibly help in evaluating an individual's eligibility for clinical trials. They also reported valuing the succinctness of these explanations, emphasizing the importance of limiting the presented information to 3–5 key facts to prevent cognitive overload. For patient interactions in memory clinics, explanation-by-example methods were seen as beneficial in facilitating personalized discussions, particularly to encourage healthier lifestyle choices such as quitting smoking, increasing social engagement, and exercising. However, neurologists also expressed reservations about using the explanation-by-example method with laypersons, as it could cause unnecessary anxiety to their patients, and acknowledged the inherent uncertainty in predicting an individual's future cognitive development.

Radiologists (N=4) favored textual explanations, as these aligned well with their clinical workflow of reporting pathological findings across different regions of interest. They reported being in favor of XAI systems that could pre-identify relevant areas, potentially saving time by highlighting key regions before manual assessment. However, they found relevance heatmaps to be of limited utility, as these visualizations did not directly support their need for regional pathological descriptions. Radiologists in our study also requested that the XAI methods should align with disease diagnosis guidelines, e.g., from the German Society for Neurology (Deutsche Gesellschaft für Neurology), and should automatically highlight relevant brain regions based on the suspected pathology.

Beyond XAI method's clinical validity, both radiologists and neurologists advocated for AI systems capable of integrating longitudinal patient data while accounting for comorbidities beyond neurodegeneration, such as depression, microbleeds, white matter lesions, and medical history, which may influence a patient's current disease presentation. Neurologists highlighted the need for multi-disease diagnostic capabilities to assess the likelihood of different pathologies. Additionally, they expect XAI methods to quantify certainty and confidence intervals of their suggestions. Neurologists also requested an extension to the

explanation-by-example approach, to incorporate multimodal data—including PET-Tau, blood-based biomarkers, and genetic makeup, to be more confident of the projected trajectories.

Discussion

In this study, we introduced a framework that offers a novel unsupervised approach to XAI by extending the scope of conventional relevance heatmaps. Our study extends the basic explanation space by including the regional morphological information, i.e., cortical thickness and volumetry measures, creating the *context-enriched explanation space.* Within this new space, we quantified the information present in the relevance heatmaps and provided evidence for relatively better clustering outcomes with respect to the disease diagnosis labels. We also explored three different methods of generating explanations for the model's predictions, namely: (i) group-based clustering of stable and converter participants, leading to simplified explanations, (ii) neighborhood-based examples of cognitive trajectories, and (iii) rule-based textual reports of pathologic regions. To the best of our knowledge, only a few studies have quantitatively compared relevance heatmaps between different dementia diagnosis groups[23,24,50]. Our study is the first of its kind to examine clinicians' feedback for the generated explanation types.

While our framework offers an enriched feature space that integrates model-derived relevance maps with regional morphological measures, it does not directly capture the entire decision-making process of the CNN. Instead, it validates and contextualizes the information embedded in the relevance maps by situating them in a clinically interpretable feature space. Within the broader challenges of XAI in imaging, aside from certain preliminary approaches such as topographic activation maps[72,73], no current method, to the best of our knowledge, is capable of exhaustively reconstructing or tracing the internal reasoning of CNNs or other deep models. Our contribution should therefore be seen as a complementary approach to the existing relevance heatmaps generating methods, in that it enhances interpretability by bridging abstract heatmaps with clinically meaningful features, while acknowledging the limitations in tracing deep models' reasoning.

Enriching explanation space and explanations generation

Recent studies have provided a quantitative interpretability framework by measuring the agreement between the generated relevance maps and meta-learned disease likelihood maps, i.e., a proxy-ground truth [23,24]. However, these were supervised approaches with only one fixed ground truth for all patients, i.e., the regional disease likelihood. Our study on the other hand, adopts an unsupervised approach that uses the morphological features as proxy ground truth features, which are unique to each patient. This allows for validation of the relevance maps based on the pathologic features tailored to each patient.

Based on the results presented in Table 2, we found that the inclusion of contextual information enhances the homogeneity of the clusters. Clustering in the enriched explanation space leads to better alignment with disease diagnosis labels. There is an improvement in the homogeneity (from 0.34 to 0.4) and V-measure (from 0.39 to 0.43), when comparing clustering outcome in enriched explanation space to basic explanation space. Our findings suggests that contextual features create relatively more coherent clusters, where now participants with the same disease diagnosis are clustered together. As a result, this refines

the explanation space itself, making it more representative of the underlying disease pathology. However, the improved homogeneity comes at the cost of cluster separability, as shown by the lower silhouette score and increased DBI, suggesting a trade-off between interpretability and structural distinction in the explanation space.

To further assess the added value of CNN-derived features, we conducted an additional clustering experiment using only volumetric and cortical thickness measures as the feature space (see Table 2). The resulting clustering outcomes in this explanation space were subpar in terms of cluster homogeneity (0.26) and had limited ability to distinguish between disease stages (V-measure of 0.32). These findings suggest that while morphological features provide supportive contextual information by enriching the explanation space.

Clustering, unlike supervised overlap quantifications, also serves as a flexible approach for integrating diverse information sources, making it adaptable for future applications incorporating various pathological measures, e.g., by adding FDG-PET or tau-PET scans [74,75]. This would allow for explanations to be generated from multi-modal data sources, possibly better capturing the interaction between various clinical factors and making the explanations more inclusive of diverse clinical contexts.

Group-based explanations

The identified subclusters in the context-enriched explanation space provided meaningful differentiation in longitudinal cognitive trajectories, reinforcing the importance of the CNN model's attribution maps when grouped with morphological features. Participants in the stable subcluster demonstrated a significantly lower risk of progression, as evidenced by mixed-effects modeling (Figure 4) and Kaplan-Meier analysis (Figure 5), respectively. On the other hand, in the converter subcluster, participants were more likely to have a rapid cognitive decline. These findings highlight the potential of the clustering model to stratify participants' disease progression risk, using structural MRI scans and CNN models trained on it, aiding in early identification and intervention planning. These explanations serve as simplified interpretations for generated relevance maps from CNN's predictions and, without overly highlighting individual morphological or relevance features.

Explanation-by-examples

We used the K-nearest neighbor (KNN) model within the context-enriched explanation space to provide a dynamic method for generating example-based explanations of possible cognitive trajectories. Rather than relying on identified hierarchical sub-clusters, KNN allows for a dynamic selection of the neighborhood. By identifying participants with the most similar pathology, this method enables personalized projections of possible cognitive trajectories without making any modeling assumptions, as outlined by an earlier study [24].

The choice of KNN over alternative meta-models was intentional, as it abstracts away complex aggregation details that could obscure interpretability for a clinical user. For instance, explanations offered by interpreting a regression model's parameters, i.e., beta coefficients may less intuitive and might not provide needed decision support functionalities. A similar argument for a lack of intuitive clarity could also be made about marginal contribution scores calculated via Shapley values. Instead, the chosen neighborhood-based approach offers a more accessible way to present likely disease trajectories by linking a participant's current pathology profile to other participants tracked longitudinally. More importantly, the objective of the KNN model was not to develop a meta-

classifier superior to the original CNN, but rather to offer example-based explanations that enhance interpretability. The notion of neighborhood plays a key role here, providing transparent and participant-specific insights into the rationale for predicting disease progression.

Rule-based textual explanations

Moving away from data-driven explanations towards knowledge-driven explanations, the ontology-based explanation method provides a structured approach to generating individualized textual summaries of neuroanatomical abnormalities. Rule-based summarization reduces cognitive overload on clinicians by hierarchically aggregating pathological findings using a-priori neuroanatomical knowledge. The generated template-based textual reports provide an intuitive means of communicating the model's decisions to the clinical users. Unlike purely data-driven deep learning models, which often lack transparency, this approach integrates CNN relevance with morphological features in a rule-based manner, enhancing clinical usability.

Both rule-based explanations and explanation-by-examples generate so-called *local* explanations, which means they show individual properties of a single participant's data. In contrast, methods that generate *global* explanations, which target the overall behavior of the whole model, might overlook the subtleties of individual cases. Local interpretations are often found to assist in making context-sensitive decisions [76,77], which is crucial in domains such as medical diagnosis.

Qualitative evaluation of the explanation types

In our the focus-group interviews, we aimed to facilitate the collaboration between method developers and healthcare professionals. Martin et al.[11] highlight the need for clinical stakeholders in evaluating XAI for dementia and radiology. Limited expert involvement hinders adoption and reduces the effectiveness of XAI methods, as clinicians ensure that the explanations align with their workflows and aid decision-making.

We report that the neurologists in our study favored group-level explanations for expert communication and risk assessment, but they were cautious about using explanation-by-example with patients. The radiologists in our sample preferred textual explanations for their workflow. Also, they viewed relevance heatmaps to be less useful for pathology reporting. These distinctions between the two professional groups underscores their differing priorities, with neurologists focusing on both current and future patient care, while radiologists concentrate more on the accurate description of pathological imaging findings to support diagnosis and treatment planning. Future XAI development in neurodegenerative research will benefit by accounting for these varying needs across clinical specialties and use cases.

Limitations and future work

As the current work is based on a data-driven methodology, one key limitation is that the explanation space is inherently dependent on the CNN model and the relevance heatmap generation XAI method used for its creation. This implies that the subsequent quantification of explanations' quality relies upon the performance of the underlying CNN model and relevance attribution method, which necessitates the use of a well-trained and generalizable CNN model to derive relevance attributions from. In the current study to mitigate this issue,

during cross-validation we select the CNN model from the fold with the best performance metrics. Additionally, the cluster quality metrics within various explanation spaces (Table 2), are relative measures and should therefore be interpreted in relation to one another rather than as absolute indicators. Although, the qualitative evaluation of XAI methods highlighted key considerations for future research and development, there remains certain limitations. We acknowledge a small sample size of experts in our study. The selection of neurologists in was non-random, which may introduce selection bias, as those experts may already favor XAI adoption. These drawbacks would be rectified in our future work.

A limitation of our study is that mutual information based feature selection and downstream analysis used the same dataset. Although cross-validation with Fleiss' Kappa showed stable outcomes, future work would be further strengthened by validation on independent datasets. Moving forward, future studies should also explore different XAI methods for relevance map generation and compare them head-to-head with the LRP method presented in our current study. More pertinently, our future research will focus on assessing the model's and the generated explanation's confidence and certainty, for we assume that this would enhance the reliability of the explanations [43,49]. Furthermore, in future work we would like the explanations to be automatically tailored to their intended use case, i.e., distinguishing between communication among clinicians, where explanations are detailed, versus communication between clinicians and laypersons, which requires simplified and layperson-friendly language.

Another line of promising research is leveraging the large language models (LLMs) for knowledge-driven, ontology-based textual explanation refinement. Retrieval augmented generation (RAG) might be particularly suitable, as it improves interpretability by keeping LLMs grounded in the context provided[78]. This approach would minimize the risk of "hallucination" that is often associated with LLMs, while ensuring faithfulness to the underlying domain logic.

Conclusion

This study introduces a framework for generating various types of explanations based on different XAI methods. Our proposed methods enrich the standard explanation space with clinically relevant morphological features. Our results demonstrate that the enriched explanation space yields more clinically meaningful insights, as shown by improved clustering metrics and the ability to distinguish between stable and converter participant subgroups. The explanation-by-example method visualizes exemplary possible cognition trajectories for a query participant for up to 72 months without making further modeling assumptions. The ontology-based textual explanations are dynamically generated in a rule-based manner, creating structured summaries that reduce cognitive overload for clinicians. Furthermore, our qualitative evaluation with clinicians highlighted the practical relevance of different explanation types.

Conflict of Interest

S.Teipel: was serving on advisory boards of Eisai, Lilly, and GE Healthcare. He is member of the independent data safety and monitoring board of the study ENVISION (Biogen). A.Hermann: received honoraria for presentations and participation in advisory boards from Amylyx and IFT Pharma. He has received royalties from Elsevier Press and Kohlhammer. E.Duezel: Paid consultancy work and talks for Roche, Lilly, Eisai, Biogen, neotiv, and UCLC; Holds shares of neotiv. O.Peters: Paid consultancy work and talks for Biogen, Eisai, Grifols, Lilly, Noselab, Prinnovation, Schwabe, and Roche. J.Wiltfang: Paid consultancy and talks for Abbott, Actelion, Amgen, Beijing Yibai Science and Technology Ltd., Biogen, Boehringer Ingelheim, Gloryren, Immungenetics, Janssen Cilag, Lilly, Med Update GmbH, MSD Sharp & Dohme, Noselab, Pfizer, Roche, and Roboscreen; holds patents PCT/EP2011 001724 and PCT/EP 2015 052945 and also supported by an Ilidio Pinho professorship, iBiMED (UIDB/04501/2020) at the University of Aveiro, Portugal. J.Priller: serves on the TSC of the Sinapps2 study and holds patents on EPO variants. F.Jessen: received fees for consultation from Eli Lilly, Novartis, Roche, BioGene, MSD, Piramal, Janssen, and Lundbeck. M.Synofzik: received consultancy honoraria from Ionis, UCB, Prevail, Orphazyme, Servier, Reata, GenOrph, AviadoBio, Biohaven, Zevra, and Lilly, all unrelated to the present manuscript. J.Levin: speaker fees from Bayer Vital, Biogen, EISAI, TEVA, and Roche, consulting fees from Axon Neuroscience and Biogen, author fees from Thieme medical publishers. These financial interests caused no effects on the study design, data collection and analysis, decision to publish, or preparation of the manuscript. All other authors declare no competing interests.

Author Contributions

Devesh Singh: conceptualization, investigation, methodology, software, visualization, writing – original draft; Stefan J. Teipel: writing – review & editing, supervision; Martin Dyrba: writing – review & editing, conceptualization, software, supervision; All other co-authors - Yusuf Brima, Fedor Levin, Martin Becker, Bjarne Hiller, Andreas Hermann, Irene Villar-Munoz, Lukas Beichert, Alexander Bernhardt, Katharina Buerger, Michaela Butryn, Peter Dechen, Emrah Düzel, Michael Ewers, Klaus Fliessbach, Silka D. Freiesleben, Wenzel Glanz, Stefan Hetzer, Daniel Janowitz, Doreen Görß, Ingo Kilimann, Okka Kimmich, Christoph Laske, Johannes Levin, Andrea Lohse, Falk Luesebrink, Matthias Munk, Robert Perneczky, Oliver Peters, Lukas Preis, Josef Priller, Johannes Prudlo, Diana Prychynenko, Boris S. Rauchmann, Ayda Rostamzadeh, Nina Roy-Kluth, Klaus Scheffler, Anja Schneider, Louise Droste, Björn H. Schott, Annika Spottke, Matthis Synofzik, Jens Wiltfang, Frank Jessen, Marc-André Weber: data acquisition, collection and curation, writing – review & editing.

Funding

This study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project ID 454834942, funding code DY151/2-1.

Acknowledgments

We sincerely appreciate the expertise and support of Dr. Großmann, Dr. Jäschke, Dr. Beller and Dr. Streckenbach from the Institute for Diagnostic and Interventional Radiology,

Pediatric and Neuroradiology at the University Hospital of Rostock in the qualitative evaluation of the methods developed.

The data samples were partly collected from the DELCODE study group of the Clinical Research Unit of the German Center for Neurodegenerative Diseases (DZNE). Details and participating sites can be found at www.dzne.de/en/research/studies/clinical-studies/delcode. The data samples were partly collected from the DESCRIBE-FTD study group of the Clinical Research Unit of the German Center for Neurodegenerative Diseases (DZNE). Details and participating sites can be found at https://www.dzne.de/en/research/studies/clinical-studies/describe-ftd/. Data collection and sharing for this project was partially funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering and generous support of other industry partners. A complete listing of ADNI investigators can be found at https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. Data was collected by the AIBL study group. AIBL researchers are listed at https://aibl.csiro.au. European DTI study on dementia EDSD was collected by nine European centers: Amsterdam (Netherlands), Brescia (Italy), Dublin (Ireland), Frankfurt (Germany), Freiburg (Germany), Milano (Italy), Mainz (Germany), Munich (Germany), and Rostock (Germany). FTLDNI was funded through the National Institute of Aging with the goal of identifying neuroimaging modalities and methods for tracking frontotemporal lobar degeneration (FTLD). For up-to-date information on participation and protocol, please visit http://memory.ucsf.edu/research/studies/nifd. Data collection and sharing for this project was funded by the Frontotemporal Lobar Degeneration Neuroimaging Initiative (National Institutes of Health Grant R01 AG032306).

Data Availability Statement

The source code is available via GitHub: https://github.com/martindyrba/xai4dementia-framework

Data used for training/evaluation of the models is available from the respective initiatives ADNI: https://adni.loni.usc.edu/data-samples/, AIBL: https://aibl.org.au/ , DELCODE: https://www.dzne.de/en/research/studies/clinical-studies/delcode, DESCRIBE: https://www.dzne.de/en/research/studies/clinical-studies/describe/, EDSD: https://www.gaaindata.org/partner/EDSD, NIFD/FTLDNI: https://memory.ucsf.edu/research-trials/research/allftd.

Ethics Statement

The studies involving humans were approved by the respective neuroimaging initiatives internal review boards of each of the participating study sites. See https://adni.loni.usc.edu and https://aibl.org.au for details. All initiatives met common ethical standards in the collection of the data such as the Declaration of Helsinki. Analysis of the data was approved by the internal review board of the Rostock University Medical Center, reference number A 2020-0182. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from

the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

The study involved focus-group interviews with clinical experts who are cognitively normal individuals and for whom oral consent was deemed sufficient. First, the group supervisors from the Institute of Diagnostic and Interventional Radiology, Pediatric Radiology and Neuroradiology (IfDIR) and the Klinik für Psychosomatische Medizin und Psychotherapie (KPM-Geron) of the University Medical Centre Rostock were contacted with the interview request, explaining the study's purpose and procedure. Individual experts were contacted for scheduling appointments. Before the start of the interviews, experts provided consent for both the interview and audio recording.

References

1. Emma Nichols, e. a. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. *The Lancet. Public health* 7, e105-e125; 10.1016/S2468-2667(21)00249-8 (2022).
2. ALZHEIMER'S DISEASE INTERNATIONAL. World Alzheimer Report 2022. Life after diagnosis: Navigating treatment, care and support (2022).
3. Watson, D. S. *et al.* Clinical applications of machine learning algorithms: beyond the black box. *BMJ (Clinical research ed.)* 364, l886; 10.1136/bmj.l886 (2019).
4. Goodman, B. & Flaxman, S. European Union Regulations on Algorithmic Decision Making and a "Right to Explanation". *AI Magazine* 38, 50–57; 10.1609/aimag.v38i3.2741 (2017).
5. Cabitza, F. *et al.* Quod erat demonstrandum? - Towards a typology of the concept of explanation for the design of explainable AI. *Expert Systems with Applications* 213, 118888; 10.1016/j.eswa.2022.118888 (2023).
6. The Royal Society. Explainable AI: the basics Policy briefing (2019).
7. OECD. The OECD AI Principles (2024).
8. The International Research Center for AI Ethics and Governance. International Research Center for AI Ethics and Governance. A Cross Cultural and Transdisciplinary Center for Building Responsible and Beneficial AI for Human and Ecology Good.
9. Tavares, J. Application of Artificial Intelligence in Healthcare: The Need for More Interpretable Artificial Intelligence. *Acta medica portuguesa* 37, 411–414; 10.20344/amp.20469 (2024).
10. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy (Basel, Switzerland)* 23; 10.3390/e23010018 (2020).
11. Martin, S. A., Townend, F. J., Barkhof, F. & Cole, J. H. Interpretable machine learning for dementia: A systematic review. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 19, 2135–2149; 10.1002/alz.12948 (2023).
12. Groen, A. M., Kraan, R., Amirkhan, S. F., Daams, J. G. & Maas, M. A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI? *European journal of radiology* 157, 110592; 10.1016/j.ejrad.2022.110592 (2022).
13. Horta, V. A. C. & Mileo, A. Generating Local Textual Explanations for CNNs: A Semantic Approach Based on Knowledge Graphs. In *AIxIA 2021 – Advances in Artificial Intelligence,* edited by S. Bandini, F. Gasparini, V. Mascardi, M. Palmonari & G. Vizzari (Springer International Publishing, Cham, 2022), Vol. 13196, pp. 532–549.

14. Futia, G. & Vetrò, A. On the Integration of Knowledge Graphs into Deep Learning Models for a More Comprehensible AI—Three Challenges for Future Research. *Information* 11, 122; 10.3390/info11020122 (2020).
15. Ibrahim, R. & Shafiq, M. O. Explainable Convolutional Neural Networks: A Taxonomy, Review, and Future Directions. *ACM Comput. Surv.* 55, 1–37; 10.1145/3563691 (2023).
16. Belle, V. & Papantonis, I. Principles and Practice of Explainable Machine Learning. *Frontiers in big data* 4, 688969; 10.3389/fdata.2021.688969 (2021).
17. Kenny, E. M. & Keane, M. T. Explaining Deep Learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI. *Knowledge-Based Systems* 233, 107530; 10.1016/j.knosys.2021.107530 (2021).
18. Schulz, M.-A., Chapman-Rounds, M., Verma, M., Bzdok, D. & Georgatzis, K. Inferring disease subtypes from clusters in explanation space. *Scientific reports* 10, 12900; 10.1038/s41598-020-68858-7 (2020).
19. Dong, F. *et al.* One step further into the blackbox: a pilot study of how to build more confidence around an AI-based decision system of breast nodule assessment in 2D ultrasound. *European radiology* 31, 4991–5000; 10.1007/s00330-020-07561-7 (2021).
20. Brima, Y. & Atemkeng, M. Saliency-driven explainable deep learning in medical imaging: bridging visual explainability and statistical quantitative analysis. *BioData mining* 17, 18; 10.1186/s13040-024-00370-4 (2024).
21. Rieger, L., Singh, C., Murdoch, W. J. & Yu, B. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge, 2019.
22. Arun, N. *et al.* Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiology. Artificial intelligence* 3, e200267; 10.1148/ryai.2021200267 (2021).
23. Di Wang *et al.* Deep neural network heatmaps capture Alzheimer's disease patterns reported in a large meta-analysis of neuroimaging studies. *NeuroImage* 269, 119929; 10.1016/j.neuroimage.2023.119929 (2023).
24. Leonardsen, E. H. *et al.* Constructing personalized characterizations of structural brain aberrations in patients with dementia using explainable artificial intelligence. *NPJ digital medicine* 7, 110; 10.1038/s41746-024-01123-7 (2024).
25. Huang, Y., Xu, J., Zhou, Y., Tong, T. & Zhuang, X. Diagnosis of Alzheimer's Disease via Multi-Modality 3D Convolutional Neural Network. *Frontiers in neuroscience* 13, 509; 10.3389/fnins.2019.00509 (2019).
26. Luo, M., He, Z., Cui, H., Ward, P. & Chen, Y.-P. P. Dual attention based fusion network for MCI Conversion Prediction. *Computers in biology and medicine* 182, 109039; 10.1016/j.compbiomed.2024.109039 (2024).
27. Tolonen, A. *et al.* Data-Driven Differential Diagnosis of Dementia Using Multiclass Disease State Index Classifier. *Frontiers in aging neuroscience* 10, 111; 10.3389/fnagi.2018.00111 (2018).
28. Holzinger, A., Carrington, A. & Müller, H. Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations. *Kunstliche intelligenz* 34, 193–198; 10.1007/s13218-020-00636-z (2020).
29. Ellis, K. A. *et al.* The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International psychogeriatrics* 21, 672–687; 10.1017/S1041610209009405 (2009).
30. Jessen, F. *et al.* Design and first baseline data of the DZNE multicenter observational study on predementia Alzheimer's disease (DELCODE). *Alzheimer's research & therapy* 10, 15; 10.1186/s13195-017-0314-2 (2018).
31. Brueggen, K. *et al.* The European DTI Study on Dementia - A multicenter DTI and MRI study on Alzheimer's disease and Mild Cognitive Impairment. *NeuroImage* 144, 305–308; 10.1016/j.neuroimage.2016.03.067 (2017).

32. Isensee, F. *et al.* Automated brain extraction of multisequence MRI using artificial neural networks. *Human brain mapping* 40, 4952–4964; 10.1002/hbm.24750 (2019).
33. Henschel, L. *et al.* FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. *NeuroImage* 219, 117012; 10.1016/j.neuroimage.2020.117012 (2020).
34. Henschel, L., Kügler, D. & Reuter, M. FastSurferVINN: Building resolution-independence into deep learning segmentation methods-A solution for HighRes brain MRI. *NeuroImage* 251, 118933; 10.1016/j.neuroimage.2022.118933 (2022).
35. 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; 10.1016/j.neuroimage.2006.01.021 (2006).
36. Klein, A. & Tourville, J. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in neuroscience* 6, 171; 10.3389/fnins.2012.00171 (2012).
37. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks, 2016.
38. Singh, D. & Dyrba, M. Comparison of CNN Architectures for Detecting Alzheimer's Disease using Relevance Maps. In *Bildverarbeitung für die Medizin 2023,* edited by T. M. Deserno*, et al.* (Springer Fachmedien Wiesbaden, Wiesbaden, 2023), pp. 238–243.
39. Bach, S. *et al.* On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PloS one* 10, e0130140; 10.1371/journal.pone.0130140 (2015).
40. Dyrba, M., Pallath, A. H. & Marzban, E. N. Comparison of CNN Visualization Methods to Aid Model Interpretability for Detecting Alzheimer's Disease. In *Bildverarbeitung für die Medizin 2020,* edited by T. Tolxdorff*, et al.* (Springer Fachmedien Wiesbaden, Wiesbaden, 2020), pp. 307–312.
41. 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,* edited by W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen & K.-R. Müller (Springer International Publishing, Cham, 2019), Vol. 11700, pp. 193–209.
42. Kohlbrenner, M. *et al.* Towards Best Practice in Explaining Neural Network Decisions with LRP, 2019.
43. Hiller, B. C. *et al.* Evaluating the Fidelity of Explanations for Convolutional Neural Networks in Alzheimer's Disease Detection. In *Bildverarbeitung für die Medizin 2025,* edited by C. Palm*, et al.* (Springer Fachmedien Wiesbaden, Wiesbaden, 2025), pp. 76–81.
44. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization; 10.48550/arXiv.1610.02391 (2016).
45. Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks, 2013.
46. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks, 2017.
47. Santi, L. A. de, Pasini, E., Santarelli, M. F., Genovesi, D. & Positano, V. An Explainable Convolutional Neural Network for the Early Diagnosis of Alzheimer's Disease from 18F-FDG PET. *Journal of digital imaging* 36, 189–203; 10.1007/s10278-022-00719-3 (2023).
48. Pohl, T., Jakab, M. & Benesova, W. Interpretability of deep neural networks used for the diagnosis of Alzheimer's disease. *Int J Imaging Syst Tech* 32, 673–686; 10.1002/ima.22657 (2022).
49. Dyrba, M. *et al.* Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease. *Alzheimer's research & therapy* 13, 191; 10.1186/s13195-021-00924-2 (2021).
50. 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. *Frontiers in aging neuroscience* 11, 194; 10.3389/fnagi.2019.00194 (2019).
51. Eitel, F. & Ritter, K. Testing the Robustness of Attribution Methods for Convolutional Neural Networks in MRI-Based Alzheimer's Disease Classification. In *Interpretability of Machine*

Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, edited by K. Suzuki*, et al.* (Springer International Publishing, Cham, 2019), Vol. 11797, pp. 3–11.

52. Singh, D. *et al.* A computational ontology framework for the synthesis of multi-level pathology reports from brain MRI scans. *Journal of Alzheimer's disease : JAD,* 13872877251331222; 10.1177/13872877251331222 (2025).
53. Dyrba, M. *et al.* Comparison of Different Hypotheses Regarding the Spread of Alzheimer's Disease Using Markov Random Fields and Multimodal Imaging. *Journal of Alzheimer's disease : JAD* 65, 731–746; 10.3233/JAD-161197 (2018).
54. Boccardi, M. *et al.* The MRI pattern of frontal and temporal brain atrophy in fronto-temporal dementia. *Neurobiology of aging* 24, 95–103; 10.1016/S0197-4580(02)00045-3 (2003).
55. Jack, C. R. *et al.* Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology* 49, 786–794; 10.1212/wnl.49.3.786 (1997).
56. Bi, N., Tan, J., Lai, J.-H. & Suen, C. Y. High-dimensional supervised feature selection via optimized kernel mutual information. *Expert Systems with Applications* 108, 81–95; 10.1016/j.eswa.2018.04.037 (2018).
57. Rosenberg, A. & Hirschberg, J. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure, 410–420 (2007).
58. Nguyen, X., Epps, J. & Bailey, J. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance 11, 2837–2854; 10.5555/1756006.1953024 (2010).
59. Hubert, L. & Arabie, P. Comparing partitions. *Journal of Classification* 2, 193–218; 10.1007/BF01908075 (1985).
60. Chacón, J. E. & Rastrojo, A. I. Minimum adjusted Rand index for two clusterings of a given size. *Adv Data Anal Classif* 17, 125–133; 10.1007/s11634-022-00491-w (2023).
61. Fowlkes, E. B. & Mallows, C. L. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association* 78, 553–569; 10.1080/01621459.1983.10478008 (1983).
62. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53–65; 10.1016/0377-0427(87)90125-7 (1987).
63. Davies, D. L. & Bouldin, D. W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-1, 224–227; 10.1109/TPAMI.1979.4766909 (1979).
64. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58, 236–244; 10.1080/01621459.1963.10500845 (1963).
65. Levin, F. *et al.* Data-driven FDG-PET subtypes of Alzheimer's disease-related neurodegeneration. *Alzheimer's research & therapy* 13, 49; 10.1186/s13195-021-00785-9 (2021).
66. Racine, A. M. *et al.* Biomarker clusters are differentially associated with longitudinal cognitive decline in late midlife. *Brain : a journal of neurology* 139, 2261–2274; 10.1093/brain/aww142 (2016).
67. Polikar, R. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* 6, 21–45; 10.1109/MCAS.2006.1688199 (2006).
68. Dissanayake, P. I., Colicchio, T. K. & Cimino, J. J. Using clinical reasoning ontologies to make smarter clinical decision support systems: a systematic review and data synthesis. *Journal of the American Medical Informatics Association : JAMIA* 27, 159–174; 10.1093/jamia/ocz169 (2020).
69. Malhotra, A. *et al.* ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 10, 238–246; 10.1016/j.jalz.2013.02.009 (2014).
70. Singh, D. & Dyrba, M. Computational Ontology and Visualization Framework for the Visual Comparison of Brain Atrophy Profiles. In *Bildverarbeitung für die Medizin 2024,* edited by A. Maier*, et al.* (Springer Fachmedien Wiesbaden, Wiesbaden, 2024), pp. 149–154.

71. Barredo Arrieta, A. *et al.* Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58, 82–115; 10.1016/j.inffus.2019.12.012 (2020).
72. Krug, V. Neuroscience-inspired analysis and visualization of Deep Neural Networks. Universitäts- und Landesbibliothek Sachsen-Anhalt, 2024.
73. Krug, V., Ratul, R. K., Olson, C. & Stober, S. Visualizing Deep Neural Networks with Topographic Activation Maps. In *HHAI 2023: Augmenting Human Intellect,* edited by P. Lukowicz, S. Mayer, J. Koch, J. Shawe-Taylor & I. Tiddi (IOS Press2023).
74. Shojaie, M. *et al.* PET Imaging of Tau Pathology and Amyloid-β, and MRI for Alzheimer's Disease Feature Fusion and Multimodal Classification. *Journal of Alzheimer's disease : JAD* 84, 1497–1514; 10.3233/JAD-210064 (2021).
75. Song, J. *et al.* An Effective Multimodal Image Fusion Method Using MRI and PET for Alzheimer's Disease Diagnosis. *Frontiers in digital health* 3, 637386; 10.3389/fdgth.2021.637386 (2021).
76. Metta, C., Beretta, A., Pellungrini, R., Rinzivillo, S. & Giannotti, F. Towards Transparent Healthcare: Advancing Local Explanation Methods in Explainable Artificial Intelligence. *Bioengineering (Basel, Switzerland)* 11; 10.3390/bioengineering11040369 (2024).
77. Duell, J., Fan, X. & Seisenberger, M. Towards Polynomial Adaptive Local Explanations for Healthcare Classifiers. In *Foundations of Intelligent Systems,* edited by M. Ceci, S. Flesca, E. Masciari, G. Manco & Z. W. Raś (Springer International Publishing, Cham, 2022), Vol. 13515, pp. 411–420.
78. Ferber, D. *et al.* GPT-4 for Information Retrieval and Comparison of Medical Oncology Guidelines. *NEJM AI* 1; 10.1056/AIcs2300235 (2024).

Supplementary Material for 'An Unsupervised Explainable AI Framework for Dementia Detection with Context Enrichment'

A.1 Neuroimaging datasets

ADNI	CN	MCI	AD	FTD
Age	72.48 ± 6.99	73.20 ± 7.92	74.94 ± 7.90	
MMSE	29.11 ± 1.13	27.70 ± 1.93	22.68 ± 3.25	
Sex (M/F)	239/341	224/174	142/107	
AIBL				
Age	72.44 ± 6.18	74.33 ± 6.85	73.18 ± 7.26	
MMSE	28.73 ± 1.23	27.04 ± 2.19	21.24 ± 5.31	
Sex (M/F)	188/260	49/47	26/36	
DELCODE				
Age	68.25 ± 5.43	72.46 ± 5.72	74.66 ± 6.24	
MMSE	29.47 ± 0.83	27.80 ± 1.97	23.09 ± 3.25	
Sex (M/F)	120/162	81/72	42/61	
EDSD				
Age	68.59 ± 6.00	71.11 ± 7.41	72.69 ± 8.29	
MMSE	28.61 ± 3.18	26.34 ± 2.99	20.86 ± 5.28	
Sex (M/F)	98/97	94/76	56/78	
NIFD				
Age	62.67 ± 7.25			63.01 ± 7.08
MMSE	29.38 ± 0.77			23.79 ± 7.21
Sex (M/F)	58/74			75/57
DESCRIBE				
Age	60.57 ± 13.40			64.08 ± 9.50
MMSE	29.07 ± 1.15			23.04 ± 6.87
Sex (M/F)	29/29			37/24
NIFD	**CN**	**BV**	**SV**	**PFNA**
Age	62.67 ± 7.25	60.97 ± 6.59	62.65 ± 6.10	67.44 ± 7.29
MMSE	29.38 ± 0.77	23.87 ± 5.66	22.59 ± 9.57	25.03 ± 6.78
Sex (M/F)	58/74	42/21	21/16	12/20
DESCRIBE	**CN**	**BV**	**SV**	**PFNA**
Age	60.57 ± 13.40	64.08 ± 9.50		
MMSE	29.07 ± 1.15	23.04 ± 6.87		
Sex (M/F)	29/29	37/24		

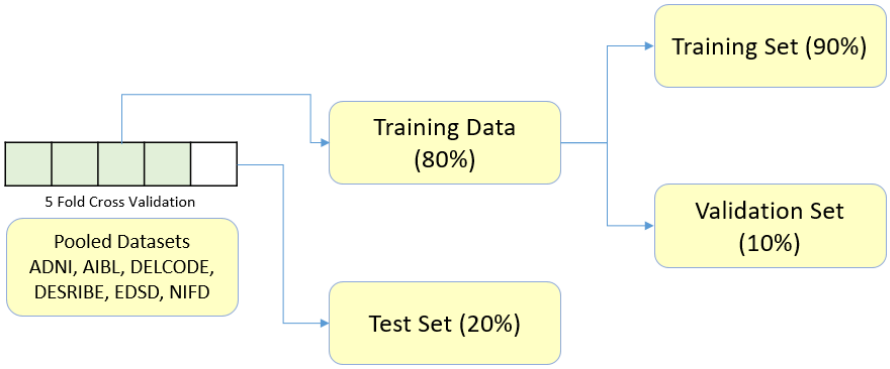
Supplementary Table A.1.1. Patient statistics separated by the diagnosis group. The statistics are reported for each of the seven data cohorts. The patients were pooled from the following study cohorts: ADNI phase 2 and phase 3, AIBL, DELCODE, DESCRIBE, EDSD, and NIFD. CN: cognitively normal, MCI: mild cognitive impairment, AD: dementia due to Alzheimer's disease, FTD: Frontotemporal dementia, where phenotypes include, BV: behavioral variant of FTD, SV: semantic variant of FTD, and PFNA: progressive nonfluent aphasia. MMSE: mini-mental state examination score, F: female, M: male. Numbers are reported as (mean ± sd).

A.2 Model Training

We trained a DenseNet model, using a stratified five-fold cross-validation (see Fig. A.2.1). The models were trained for a three-way classification - AD-vs-CN-vs-FTD. Here Alzheimer's dementia (AD) patients and patients with amnestic mild cognitive impairment (MCI) were merged into one disease-positive class, while multiple phenotypes of frontotemporal dementia (FTD) - behavioral variant (bvFTD), semantic dementia (SD), and progressive nonfluent aphasia (PNFA) were also clubbed under one FTD class. These two classes were compared against the cognitively normal (CN) participants, i.e., the control class.

Categorical cross-entropy was chosen as the loss function. The models were optimized using the Adam optimizer with a learning rate of 0.0001, and other parameter settings were set to default. We trained the models for 100 epochs, using a batch size of 128. To reduce model over-fitting, an early stopping regularization method was applied, monitoring the validation set loss as a performance metric over epochs, with patience of 5 epochs and a minimum change threshold of 0.01. To avoid overfitting, we also weighted the model's error with the label's class weight. During each cross-validation run, only the best-performing model was saved.

For training, the data augmentations were generated using AUCMEDI Python package, where 3D volumes were randomly left/right-flipped with a 50% probability, and rescaled with a 50% probability within the zooming in or out limits of 90% and 110%. These augmentations were only applied during model training and were disabled on validation and test sets.

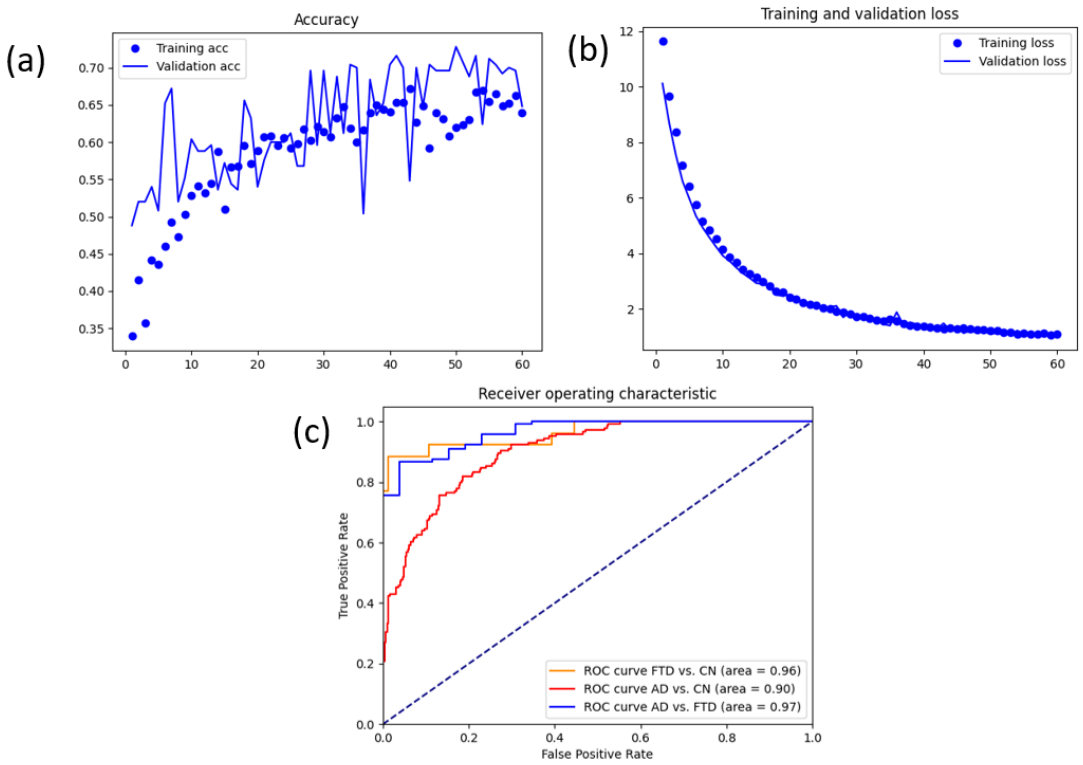


Supplementary Figure A.2.1 Schematics representation of the data splitting for CNN model training.

Based on the results from the 5-fold cross-validation training of the models (see Table A.2.1), we chose fold 1 as the default model for further analysis in our study. The model training results from fold 1 are illustrated in Figure A.2.2.

Fold	Acc ADvsCN	Acc FTDvsCN	Acc ADvsFTD	AUC ADvsCN	AUC FTDvsCN	AUC ADvsFTD
1	78.18	96.89	88.24	0.90	0.96	0.97
2	78.24	92.94	43.79	0.80	0.90	0.78
3	84.16	96.61	61.27	0.89	0.99	0.91
4	83.77	95.76	60.26	0.88	0.94	0.87
5	69.13	92.63	15.12	0.78	0.95	0.84
mean(sd)	78.7±6.07	94.97±2.04	53.73±26.83	0.85±0.05	0.95±0.03	0.87±0.07

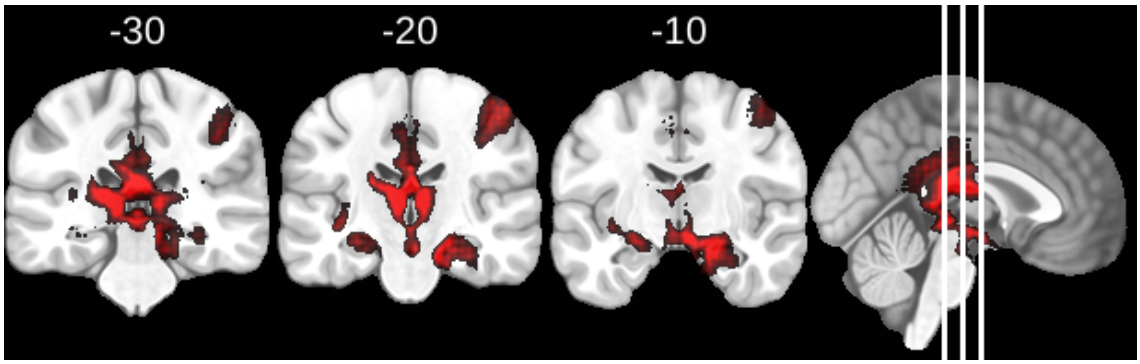
Supplementary Table A.2.1 Performance metrics on the test set. Acc: simple accuracy, AUC: Area under the (ROC) curve. CN: cognitively normal, AD: dementia due to Alzheimer's disease (which due to design choices, also includes the amnestic mild cognitive impairment (MCI) subjects), and FTD: frontotemporal dementia.



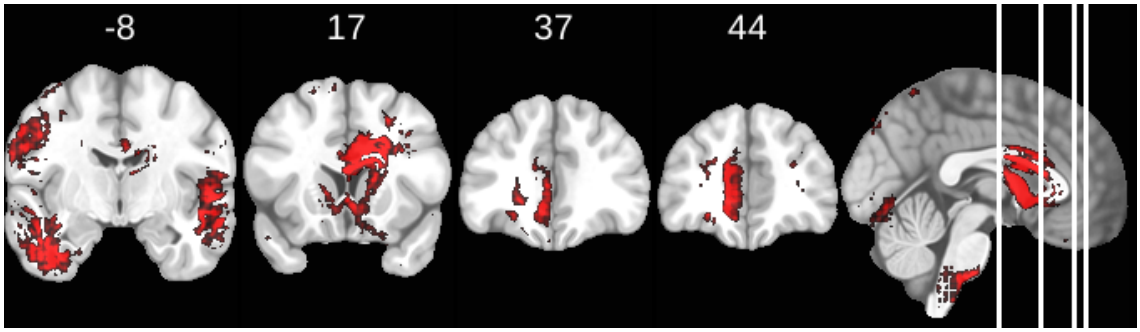
Supplementary Figure A.2.2 Model training results from fold 1: (a) (simple) Accuracy metric on train and validation set, (b) loss metric on train and validation set, and (c) binarized ROC-AUC curves on the test set. CN: cognitively normal, AD: dementia due to Alzheimer's disease (which due to design choices, also includes the amnestic mild cognitive impairment (MCI) subjects), and FTD: frontotemporal dementia.

The mean relevance maps for the test set of fold 1 are visualized below. For relevance attribution, we employed the compositional LRP rule (= 1, β = 0), as established in our previous work for generating clinically meaningful explanations[1]. To enhance the signal-to-noise ratio during visualization, we re-scaled the relevance intensities based on the 99.99th percentile (q = 0.9999) and clipped the resulting values to the range [−1, 1]. A Gaussian smoothing filter with a standard deviation of 0.8 was then applied to further improve interpretability.

The mean relevance maps of Alzheimer's disease (AD) dementia and mild cognitive impairment (MCI) patients appeared visually similar; however, distinct patterns emerged when comparing across disease groups. In the AD group, supplementary figure A.2.3, relevance was concentrated in the hippocampus (slices [-20, -10]) and bilaterally in the thalamus (slices [-30, -20]). In contrast, the frontotemporal dementia (FTD) group, supplementary figure A.2.4, exhibited prominent relevance in the frontal lobes, particularly the right insula and frontal opercular cortex in slice -8, as well as the pregenual anterior cingulate cortex (pACC) in slices [37, 44]. Notably, insular involvement was also reported in our prior study[2], suggesting consistency across different model training strategies and relevance attribution techniques in identifying clinically relevant brain regions.



Supplementary Figure A.2.3: Mean relevance maps for the AD group of the test dataset obtained using the LRPα=1,β=0 relevance propagation method overlaid on MNI brain template. Coronal slices show Y=[-10,-20,-30] mm in MNI reference space are shown. The most relevant input regions are highlighted. Relevance maps were created following proportional scaling of the activations.



Supplementary Figure A.2.4: Mean relevance maps for the FTD group of the test dataset obtained using the LRPα=1,β=0 relevance propagation method overlaid on MNI brain template. Coronal slices show Y=[-8,17,37,44] mm in MNI reference space are shown. The most relevant input regions are highlighted. Relevance maps were created following proportional scaling of the activations.

References

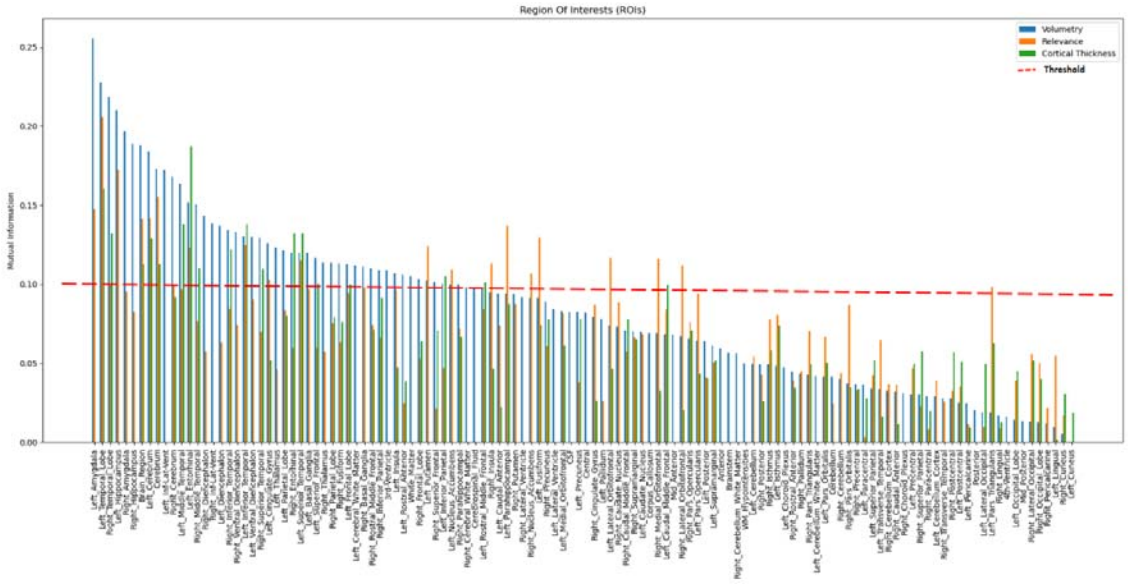
[1] Dyrba, Martin, Arjun H. Pallath, and Eman N. Marzban. "Comparison of CNN visualization methods to aid model interpretability for detecting Alzheimer’s disease." Bildverarbeitung für die Medizin 2020: Algorithmen–Systeme–Anwendungen. Proceedings des Workshops vom 15. bis 17. März 2020 in Berlin. Wiesbaden: Springer Fachmedien Wiesbaden, 2020.

[2] Gryshchuk, Vadym, et al. "Contrastive self-supervised learning for neurodegenerative disorder classification." Frontiers in Neuroinformatics 19 (2025): 1527582.

A.3 Feature Selection with Mutual Information

All features used in the mutual information analysis were derived from w-scores, representing age, sex, brain size and MRI scanner strength adjusted residualized values. Average cortical thickness measures were only estimated for cortical regions (e.g., superior temporal gyrus, frontal lobe areas) and not for subcortical structures (e.g., hippocampus, thalamus, or basal ganglia), which explains the absence of cortical thickness values in these regions.

To enhance transparency and accessibility, we have uploaded the intermediate results from our analysis pipeline to GitHub: https://github.com/martindyrba/xai4dementia-framework/. The repository includes a CSV file that specifies, for each of the 120 regions, which of the three feature types (CNN relevance, volumetry, or cortical thickness) passed the mutual information threshold and were included in downstream analysis. This is process outcome is also visualized in Supplementary Figure A.3.1.



Supplementary Figure A.3.1: Mutual Information-Based Feature Selection: This figure illustrates the mutual information values computed across three different *w-score* features - CNN relevance, volumetry, and cortical thickness w-scores, representing their shared information content in comparison to the disease diagnosis label. The w-score features were sorted according to their mutual information on the volumetry features. The w-score features with mutual information above the threshold of 0.1 were retained as relevant and were selected for further analysis. For a vector graphic rendering, please refer to the GitHub version of the plot.

A.4 Mixed-Effects Models of Cognitive Trajectories

Mixed-effects model experiments were done to investigate cognitive decline in patients, while accounting for repeated measures and inter-individual variability. The analysis was conducted using data from two groups: the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the DZNE Longitudinal Study on Cognitive Impairment and Dementia (DELCODE) cohorts. These datasets contain repeated cognitive assessments for each patient for up to 6 years, allowing for a longitudinal investigation of cognitive decline.

We tested a series of increasingly complex mixed-effects models. By incrementally adding predictors and interaction terms, we assessed model fit and explanatory power. The best-fitting model was determined using likelihood ratio tests via ANOVA.

Model 1 (Base Model): Includes age, sex, and the interaction between cluster membership and follow-up months (FUMonths) while accounting for repeated measures per participant.

CDR_i (or $MMSE_i$) = $\beta_0 + \beta_1\ age_i + \beta_2\ sex_i + \beta_3\ cluster_i + \beta_4\ FUMonths_i + \beta_5\ (cluster \times FUMonths)_i + u_i + \varepsilon_i$

where $u_i \sim N(0, \sigma_u^2)$ represents the random intercept for each participant, and $\epsilon_i \sim N(0, \sigma^2)$ is the residual error term.

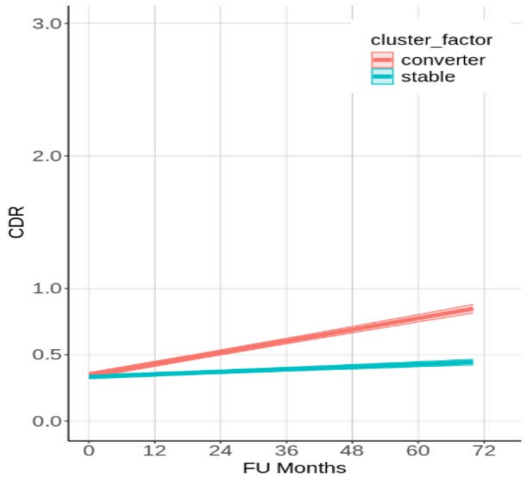
Model 2 (Expanded Diagnosis Model): Adds baseline diagnosis as a fixed effect.

CDR_i (or $MMSE_i$) = $\beta_0 + \beta_1\ age_i + \beta_2\ sex_i + \beta_3\ \mathbf{baseline_diag_i} + \beta_4\ cluster_i + \beta_5\ FUMonths_i + \beta_6\ (cluster \times FUMonths)_i + u_i + \varepsilon_i$

Model 3 (Final Model): Introduces an interaction between baseline diagnosis and follow-up months.

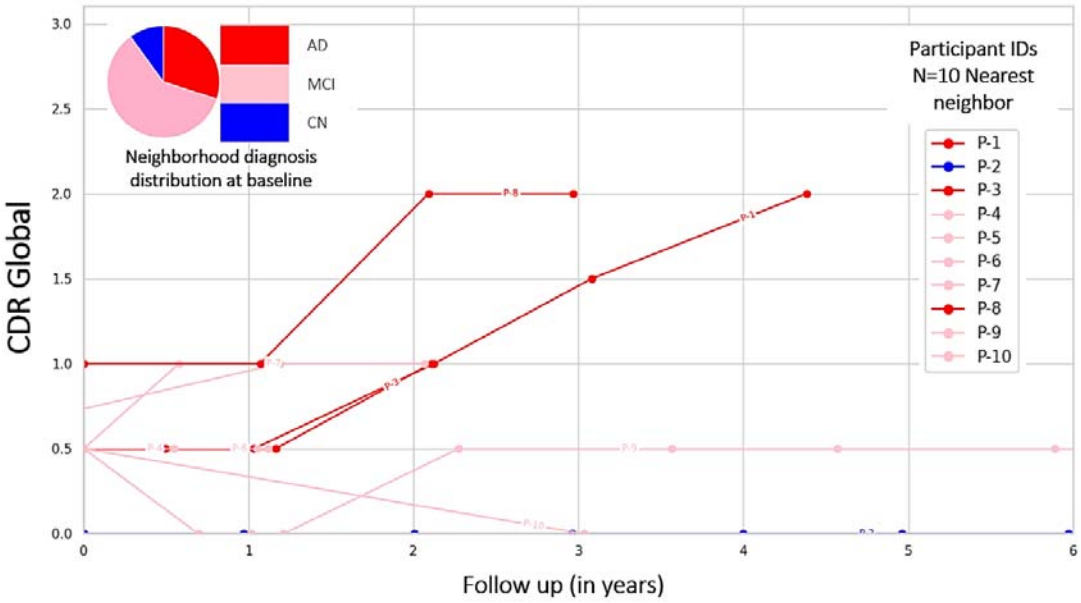
CDR_i (or $MMSE_i$) = $\beta_0 + \beta_1\ age_i + \beta_2\ sex_i + \beta_3\ baseline_diag_i + \beta_4\ cluster_i + \beta_5\ FUMonths_i + \beta_6\ \mathbf{(baseline_diag \times FUMonths)_i} + \beta_7\ (cluster \times FUMonths)_i + u_i + \varepsilon_i$

To compare model fit, ANOVA tests were performed, evaluating the nested model comparisons. The results indicated that Model 3 provided the best fit, suggesting that the interaction between baseline diagnosis and follow-up time significantly improves the model's explanatory power. Specifically, for the CDR global (Figure A.4.1:), the high-risk converter group showed an annual increase of 0.074 points, while the low-risk group remained relatively stable with an increase of 0.007 points per year.



Supplementary Figure A.4.1: Longitudinal cognitive trajectories of different clusters of patients. Values on Clinical Dementia Rating (CDR) global are obtained from mixed effects regression models which included the age, sex, baseline disease diagnosis, and the interaction between cluster membership and follow-up time in months (FU Months), as well as the interaction between baseline disease diagnosis and follow-up months. The model also included random intercepts for each patient to account for repeated measurements. The shaded regions represent 95% confidence intervals.

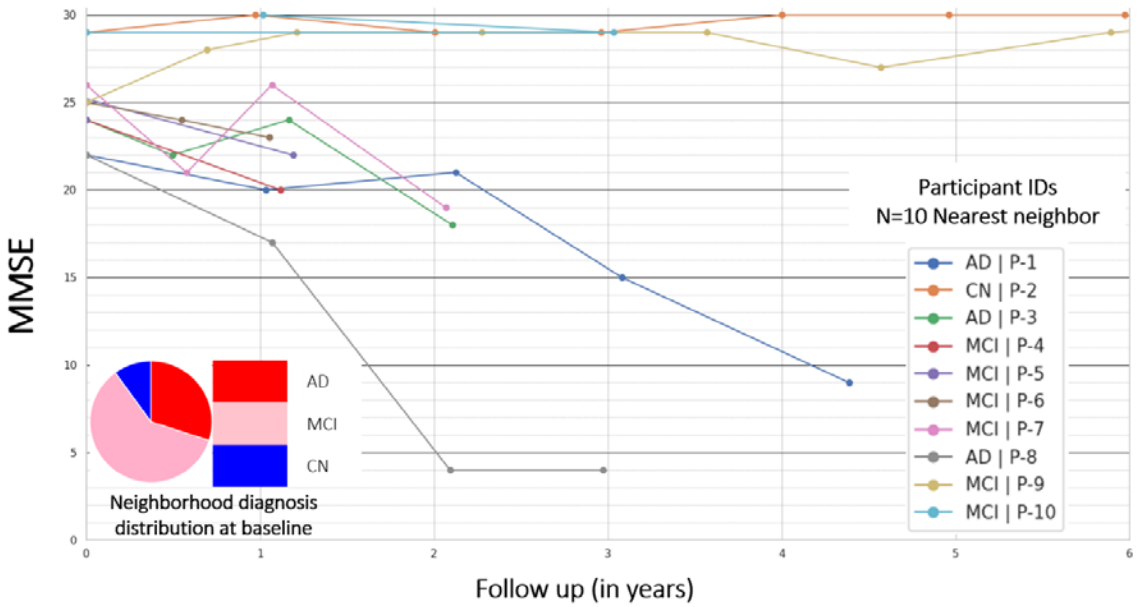
A.5 Explanation-by-example plots for Clinical Dementia Rating (CDR)



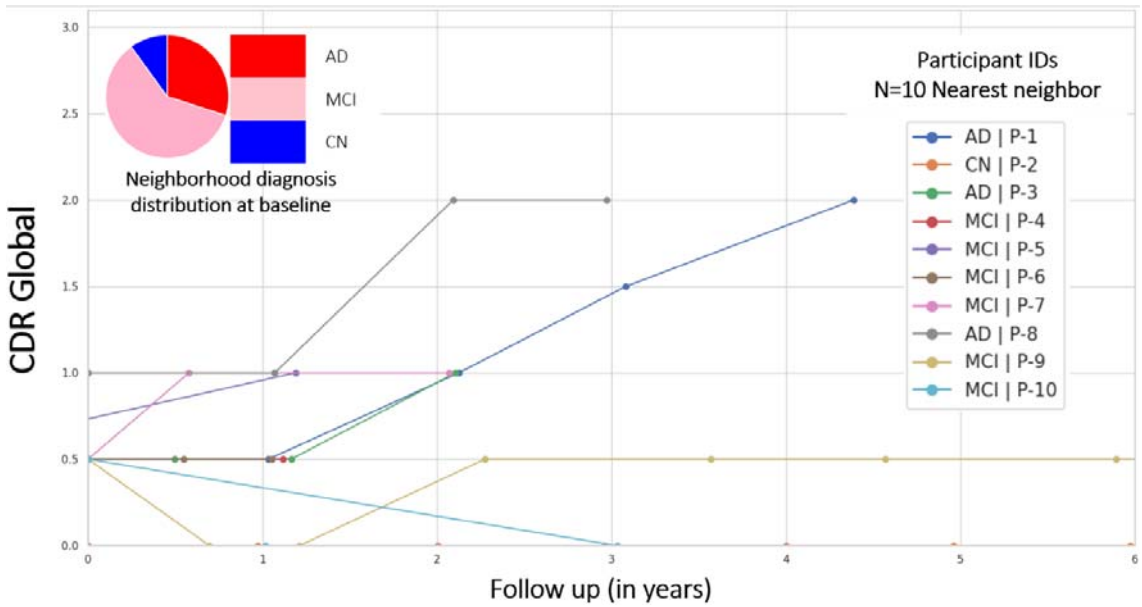
Supplementary Figure A.5.1: Explanation-by-examples: Within the context-enriched explanation space, the longitudinal cognitive trajectories of k=10 nearest neighbors of a query patient, from the DELCODE cohort, are shown. Patient IDs of the nearest neighbors are pseudonymised, and the nearest neighbors are listed in the order of increasing Euclidian distance from the query sample. Scores on the cognitive test Clinical Dementia Rating (CDR) global were observed on follow-up

examinations for up to 6 years. The cognition trajectories are additionally color-coded by the baseline disease diagnosis.

A.6 Explanation-by-example plots, each participant colored individually



Supplementary Figure A.6.1**:** Explanation-by-examples: Longitudinal MMSE trajectories of the *k = 10* nearest neighbors of a query participant from the DELCODE cohort, observed for up to 6 years. Here, each cognitive trajectory is shown in a unique color for more detail.



Supplementary Figure A.6.2**:** Explanation-by-examples: Longitudinal CDR trajectories of the *k = 10* nearest neighbors of a query participant from the DELCODE cohort, observed for up to 6 years. Here, each cognitive trajectory is shown in a unique color for more detail.