



OPEN ACCESS

EDITED BY

Nilton Custodio,
Peruvian Institute of Neurosciences (IPN), Peru

REVIEWED BY

Diego Castillo-Barnes,
University of Malaga, Spain
Carmen Jiménez-Mesa,
University of Granada, Spain

*CORRESPONDENCE

Nicola Amoroso
✉ nicola.amoroso@uniba.it

†These authors have contributed equally to this work

RECEIVED 10 June 2023

ACCEPTED 08 August 2023

PUBLISHED 31 August 2023

CITATION

Amoroso N, Quarto S, La Rocca M, Tangaro S, Monaco A and Bellotti R (2023) An eXplainability Artificial Intelligence approach to brain connectivity in Alzheimer's disease. *Front. Aging Neurosci.* 15:1238065. doi: 10.3389/fnagi.2023.1238065

COPYRIGHT

© 2023 Amoroso, Quarto, La Rocca, Tangaro, Monaco and Bellotti. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

An eXplainability Artificial Intelligence approach to brain connectivity in Alzheimer's disease

Nicola Amoroso^{1,2*}, Silvano Quarto³, Marianna La Rocca^{2,3}, Sabina Tangaro^{2,4}, Alfonso Monaco^{2,3†} and Roberto Bellotti^{2,3†}

¹Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari Aldo Moro, Bari, Italy, ²Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy, ³Dipartimento Interateneo di Fisica, Università degli Studi di Bari Aldo Moro, Bari, Italy, ⁴Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, Bari, Italy

The advent of eXplainable Artificial Intelligence (XAI) has revolutionized the way human experts, especially from non-computational domains, approach artificial intelligence; this is particularly true for clinical applications where the transparency of the results is often compromised by the algorithmic complexity. Here, we investigate how Alzheimer's disease (AD) affects brain connectivity within a cohort of 432 subjects whose T1 brain Magnetic Resonance Imaging data (MRI) were acquired within the Alzheimer's Disease Neuroimaging Initiative (ADNI). In particular, the cohort included 92 patients with AD, 126 normal controls (NC) and 214 subjects with mild cognitive impairment (MCI). We show how graph theory-based models can accurately distinguish these clinical conditions and how Shapley values, borrowed from game theory, can be adopted to make these models intelligible and easy to interpret. Explainability analyses outline the role played by regions like putamen, middle and superior temporal gyrus; from a class-related perspective, it is possible to outline specific regions, such as hippocampus and amygdala for AD and posterior cingulate and precuneus for MCI. The approach is general and could be adopted to outline how brain connectivity affects specific brain regions.

KEYWORDS

Alzheimer's disease, XAI, brain connectivity, explainability, MCI

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disease, which represents the seventh leading cause of mortality in the United States after when COVID-19 appeared at the top of this ranking (Alzheimer's Association, 2022). According to the World Health Organization (<https://www.who.int/news-room/fact-sheets/detail/dementia>), AD is the most common form of dementia (60–70%) which affects 55 million people all over the world and some studies estimated that over 150 million people will develop dementia by 2050 (Nichols et al., 2022). Alzheimer's disease involves the loss of neuronal connections, thus resulting in a connectivity damage that impairs neuronal functionality and eventually leads to their death. Neuronal death has macroscopic effects on the brain; specific brain regions start shrinking, this is what is usually known as brain atrophy (Devanand et al., 2007; Liu et al., 2019; Talwar et al., 2021).

From a clinical point of view, this feature could be highly beneficial because it can be revealed by imaging, especially by brain Magnetic Resonance Imaging (MRI) (Lerch et al., 2008; Vemuri et al., 2008; Julkunen et al., 2010). The relationships between the structure of

the brain and its functional abilities can be investigated (Emre et al., 2007; Solé-Padullés et al., 2009; Frisoni et al., 2010) and, therefore, on the one hand it is possible to assess to which extent the disease severity reflects the structural damage, on the other hand it is possible to develop accurate diagnostic approaches based on the clinical symptoms. To this aim, it becomes extremely important to develop accurate diagnosis support systems which can detect early signs of atrophy, before symptoms appear, and to initiate timely treatments (Breijyeh and Karaman, 2020). In the last decades, several studies have investigated the structural changes in the brain and tried to correlate them to different stages of the disease severity, including preclinical AD, mild cognitive impairment (MCI), and clinically diagnosed AD (Sperling et al., 2011; Alzheimer's Association, 2020).

MCI condition is particularly interesting because its symptoms are not fully evident and brain damage is not extensive. MCI is characterized by memory loss episodes, difficulty thinking, and the first signs of physical problems (Gauthier et al., 2006; Petersen et al., 2014). In addition, not all MCI subjects convert to AD: according to estimates, 8 out of 10 people with MCI develop AD within 7 years, whereas there are patients which continue to have MCI or convert back to the normal condition years later (Larrieu et al., 2002; National Institute on Aging, 2002; Gauthier et al., 2006; Tábuas-Pereira et al., 2016). It would therefore open up a wide range of possibilities for cures and disease-modifying therapies in the event we could effectively diagnose MCI and distinguish physiological impairment from early symptoms of AD (Huckans et al., 2013; Huang et al., 2020).

Neuroimaging studies have demonstrated their effectiveness to investigate brain changes and identify the first signs of disease (Lebedeva et al., 2017; Zeng et al., 2021). In particular, some studies have investigated the use of graph theory (Bullmore and Sporns, 2009; de Haan et al., 2012; Tijms et al., 2013), the conversion from MCI to AD and algorithms for studying AD (Daiyanu et al., 2015; Teipel et al., 2016; Liu et al., 2017). During the last several decades, machine learning techniques have demonstrated their ability to perform binary and multi-label classification tasks (Gupta et al., 2019; Kim et al., 2021; Sheng et al., 2021; Song et al., 2021); this is especially important when dealing with MCI, which intrinsically represents a heterogeneous clinical condition, often presenting both normal and pathological behavior. However, clinical practitioners have difficulty adopting these techniques due to the fact that they are often considered as black boxes, difficult to interpret.

In this work, we employ a brain connectivity model based on “patches”, whose effectiveness has already been verified for AD classification (La Rocca et al., 2017, 2018; Amoroso et al., 2018b,c, 2019). This overcomes the typical issues of voxelwise and region-based approaches: (i) it removes the computational burden and overfitting concerns associated with voxelwise methods (Goenka and Tiwari, 2022), but also the parametric statistical methods turned out to be over-conservative for voxelwise inference (Eklund et al., 2016; Górriz et al., 2021); (ii) using unsupervised segmentation of the brain, this approach does not require region of interest (ROI) localization based on prior biological knowledge to extract regional features, and it provides a better way to detect

microscopic structural changes in the brain that ROI extracted features cannot capture (Amoroso et al., 2015).

Here, we take a step beyond by investigating to which extent such model can “explain” the effects of AD on brain connectivity: to this aim, we consider a consolidated eXplainable Artificial Intelligence (XAI) approach based on Shapley values (Messalas et al., 2019; Loh et al., 2022). The use of XAI methods to characterize neurodegenerative diseases and, more in general, to equip neuroimaging studies is rapidly increasing (El-Sappagh et al., 2021; Anjomshoae and Pudas, 2022; Lombardi et al., 2022). A lot of neuroimaging fields have seen a highly increased interest in the application of XAI techniques, obtaining benefits from transparency provided by these approaches (Farahani et al., 2022). MRI research is exploiting these techniques to study brain aging both through *ante-hoc* interpretability models, such as stability assessment or latent variable models, and post hoc models, such as feature importance and saliency maps (Galazzo et al., 2022; Qian et al., 2023). XAI techniques were used to study the conversion from MCI to AD by high-density electroencephalography (HD-EEG) to detect which EEG-channels and range of frequencies were most predominant in disease progression (Morabito et al., 2023). In addition, clinical cognitive tests are getting advantage from the application of XAI, which provides insights into the cognitive processes by visualizing and identifying the specific cognitive features that are most influential in determining the test outcomes; using a single test such as the Clock Drawing Test or selecting a subset of cognitive tests to exploit XAI scores for individualized prediction explanations (Beebe-Wang et al., 2021; Jimenez-Mesa et al., 2023).

We initially demonstrate to which extent the patch-based approach is reliable for characterizing patients, controls and MCI subjects. Then, the classification performance and its reliability are investigated in order to ensure the model is sound. Finally, an overall explanation of the model and an explanation of its decision is provided by means of Shapley values.

2. Materials and methods

2.1. Imaging the brain and building a network model

In this research, we used a dataset composed of 432 brain T1 MRI images, relative to 126 normal control (NC), 214 MCI and 92 AD subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI is a longitudinal multicenter study designed to obtain early diagnosis and monitoring of AD through the study and development of clinical, imaging, biochemical, and genetic biomarkers (<https://adni.loni.usc.edu/about/>). ADNI images were normalized using the MNI152 brain template with $197 \times 233 \times 189$ mm³ size and $1 \times 1 \times 1$ mm³ resolution; from now onward voxel and 1 mm³ will be interchangeably used. In the following Table 1 the number of instances, gender information, age, years of education, and Mini Mental State Examination (MMSE) score are enlisted. This cohort of MRI scans passed quality control (QC) using the Laboratory of Neuro Imaging (LONI) QC System (Kim et al., 2019)

TABLE 1 Demographic information for each class.

	AD	MCI	NC	Total
Number of instances	92	214	126	432
Female/male	43/49	86/128	65/61	194/238
Age (years)	75.82 ± 7.60	75.30 ± 7.13	75.61 ± 5.58	75.50 ± 6.81
Education (years)	15.10 ± 3.26	15.52 ± 3.30	16.05 ± 2.61	15.58 ± 3.11
MMSE	23.45 ± 1.95	26.98 ± 1.75	29.16 ± 0.98	26.86 ± 2.59

The table provides the number of instances, gender information, and also age, years of education MMSE, with mean and standard deviation.

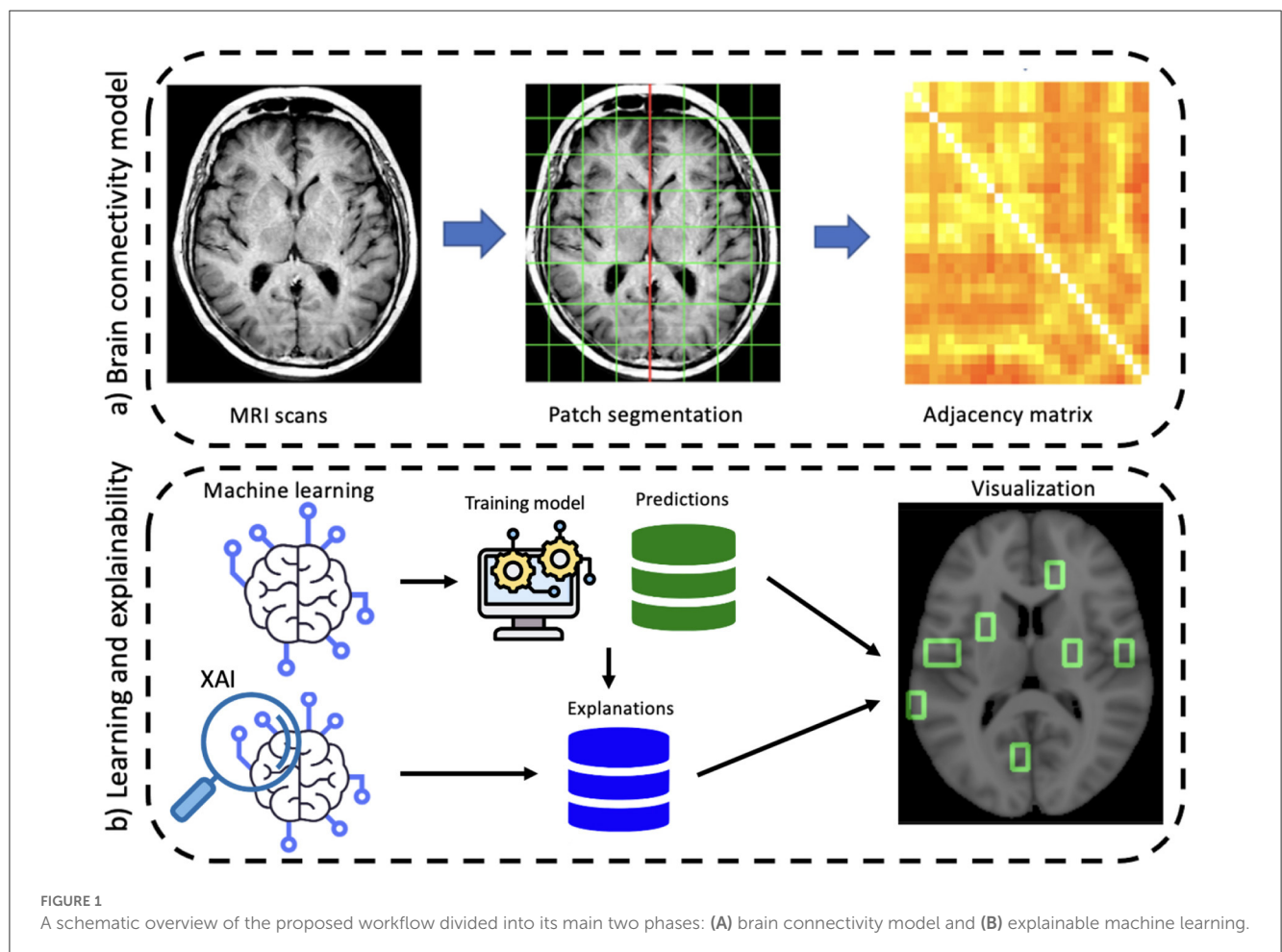


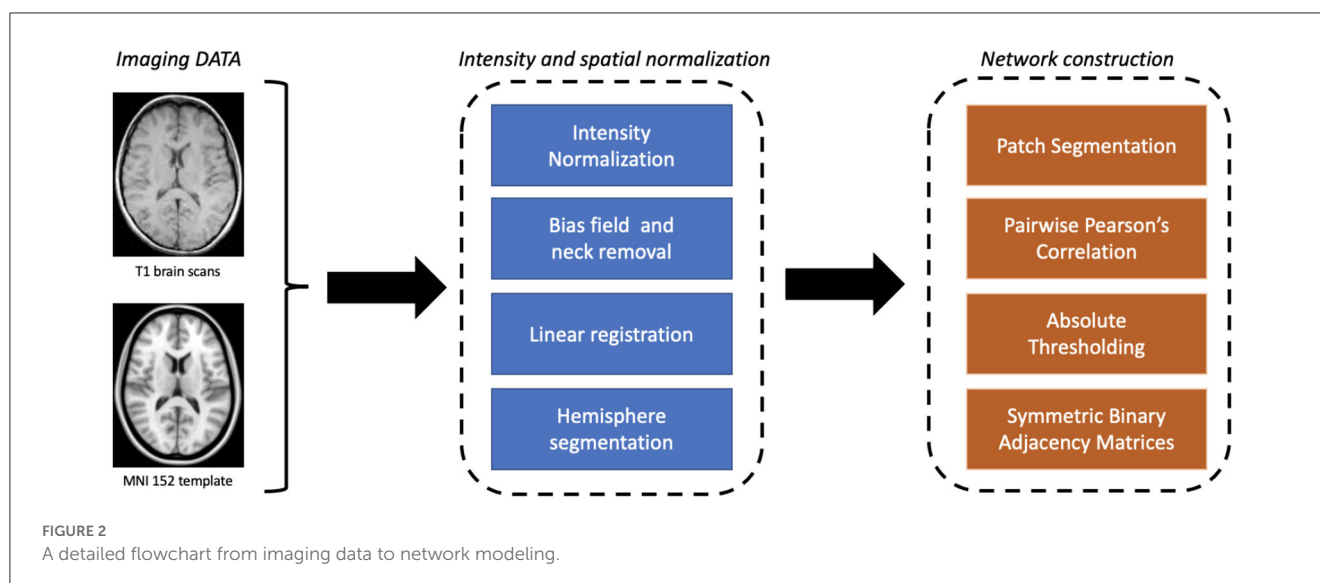
FIGURE 1 A schematic overview of the proposed workflow divided into its main two phases: (A) brain connectivity model and (B) explainable machine learning.

so that only high quality images were considered without signal alterations.

The proposed approach involves two main stages: the construction of the network model and the learning and explainability phase (see Figure 1). First, a graph theory-based model is employed to study structural brain connectivity and identify early signs of Alzheimer’s disease. Accordingly, each patient’s brain is modeled as a complex network. Brain MRIs are parceled into rectangular boxes called “patches” of fixed dimensions that represent the network’s nodes; the links are obtained in terms of nodes’ pairwise similarity measured by absolute Pearson’s correlation. As a difference with previous works where multiplex networks were adopted, here we consider single-subject networks

to achieve a simpler description and, therefore, a more interpretable model. In fact, using single-subject networks, network features can be directly related to single nodes and then to specific brain regions.

Image processing was the first algorithmic step. Using Oxford FMRIB software library (FSL) (Jenkinson et al., 2012), images were skull-stripped and spatially normalized as well in intensity to mitigate data heterogeneity, an aspect of fundamental importance for studying a multi-center database such as the one provided by ADNI. First bias field correction and skull stripping was performed using FSL Brain Extraction Tool (BET) (Smith, 2002). Thereafter, spatial normalization was performed to ensure co-registration to the MNI152 template using FSL Linear Registration Tool (FLIRT)



(Jenkinson et al., 2002); in particular, an affine registration was adopted with default parameters.

After registration, by using the medial longitudinal fissure, normalized brains were divided into the two hemispheres; each hemisphere was then covered by an equal number of patches of fixed dimensions $l_1 \times l_2 \times l_3$ (patches overlapping the template with $<10\%$ of voxels were neglected). For the patch dimension, we considered the $10 \times 15 \times 20 \text{ mm}^3$ configuration, corresponding to a total volume of $3,000 \text{ mm}^3$, for a total of 549 patches (see Figure 2).

Accordingly, we built an undirected weighted network for each MRI scan with 549 nodes. Pearson's correlation coefficient was chosen for links for three main reasons: (i) it has an affordable computational cost; (ii) it is easy to implement, and (iii) it is easy to interpret in terms of brain atrophy. Absolute values were considered to take into account left/right symmetry of the brain; finally, a threshold was used to remove weak correlations (<0.3) that could raise noisy connections.

Once structural connectivity had been modeled in terms of a complex network, we exploited such characterization to evaluate some network metrics and outline the effects of cerebral atrophy on brain connectivity. Furthermore, since network metrics can be easily divided in local and global ones, to acquire a detailed description of how AD differently affects distinct anatomical districts, nodal centrality measures were preferred. In particular, three graph metrics were used: strength, betweenness, and eigenvector centrality. The strength of a node is a local centrality measure defined by the sum of edge weights of each node; as a difference, a global centrality measure such as eigenvector centrality takes into account also the influence of a node in a network based on the amount of nodes with high number of connections which it is connected to. Finally, a dynamic centrality measure is considered: the betweenness; betweenness is the ratio between the shortest paths connecting two nodes passing through a specific node and all available shortest paths. Accordingly, it is a measure capturing the information flow within a network (Amoroso et al., 2021; Bellantuono et al., 2021; Sheng et al., 2021).

After calculating these three metrics for each node of the adjacency matrices, a matrix representation $M \times n$ of the data was obtained, where M is the number of enrolled patients and n is the total number of features (number of metrics used times number of nodes). Here, the resulting matrix had dimensions of $432 \times 1,647$.

2.2. Learning AD patterns

We have used the previously defined matrix representation to train a machine learning model. The main objectives we set during this learning phase were to achieve a sound representation of AD and verify the robustness and reliability of the connectivity network model. To this aim, a three-label classification was performed for AD, MCI, and NC subjects. All these analyses were carried out within a 10-fold cross-validation framework; the procedure was iterated 50 times to estimate performances and uncertainties; besides, to ensure balanced cross-validation splits between training and test, stratified sampling was used.

To ensure that the observed performance was due to the informative content provided by the network features independently from the classification model used, we compared the performance of several classifiers: Random Forest (RF) (Breiman, 2001), Support Vector machine (SVM) (Cortes and Vapnik, 1995), eXtreme Gradient Boosting (XGBoost) (Chen et al., 2015), Naive Bayes (NB) (Rish et al., 2001), and Logistic Regression (LR) (Ng and Jordan, 2001) were used.

2.3. From classification to explanations

After classification performance accuracy and reliability had been assessed, explainability analyses were carried out by means of SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017). Accordingly, it is possible to evaluate to which extent each feature affects the model's predictions. By averaging the impact of a feature

on the whole dataset, a feature importance map can be achieved; at the same time, this explanation allows for the identification of the factors which determine the classifier's decisions for each subject. Also, this analysis provides additional information compared with standard feature importance approaches in that it does not determine the importance of one feature but also it explains how its value, combined with the other features' values, leads to the decision taken by the algorithm. Moreover, it is possible to compare whether the impact of features on a single decision is coherent with the overall feature importance (the one related to the three-class classification model) or not. Finally, an important aspect to remark is that SHAP can be adopted with different machine learning algorithms (Strumbelj and Kononenko, 2010), thus making the proposed approach algorithm-independent.

A fundamental aspect of SHAP concerns their computational burden, as the computational load increases exponentially with the number of features. To overcome this issue, a specific experimental design is proposed. Explainability analyses were nested in the repeated 10-fold cross-validation. Using a RF classifier, for each cross-validation round a subset of important features was selected by means of the combined use of Mean Decrease Accuracy (MDA) and statistical significance.

This procedure can be described in the following two steps: (i) firstly, the features exceeding the 95th percentile of the MDA distribution were selected within each cross-validation round; (ii) then, these newly selected features were further reduced considering statistical significance: only the features with the 5% significance using the one-sided binomial test were kept. On average, the number of features selected during each cross-validation round was ~ 50 . During each cross-validation round, this final set of important features was used to train a second classification model on the same training set with reduced number of features, in order to evaluate its informative content and to carry out the explainability analyses. Shapley values were calculated for each of the test subjects of each cross-validation round. The same analysis was carried out for each available class by considering three one-vs.-all classifiers to outline the presence of class-specific behaviors (Whitwell et al., 2007; Byun et al., 2015; Cabral et al., 2015). The mentioned analyses were carried out with R, with the *DALEX* and *shapviz* packages (Biecek, 2018; Biecek and Burzykowski, 2020; Michael Mayer, 2022). Finally, to validate the results from a clinical perspective, we determined which specific brain regions were associated with the selected important features (as each feature was directly related to a patch). To this aim, the Talairach atlas (Lancaster et al., 1997, 2000) was used.

3. Results

3.1. Robustness and reliability

Firstly, we investigated the validity and reliability of the brain connectivity model. To this aim, a 10-fold cross-validation analysis was repeated 50 times, results are shown in Figure 3.

These findings demonstrate that, except for LR, the adopted models are consistent and the informative power of the network features can reach satisfactory values: in terms of accuracy, RF

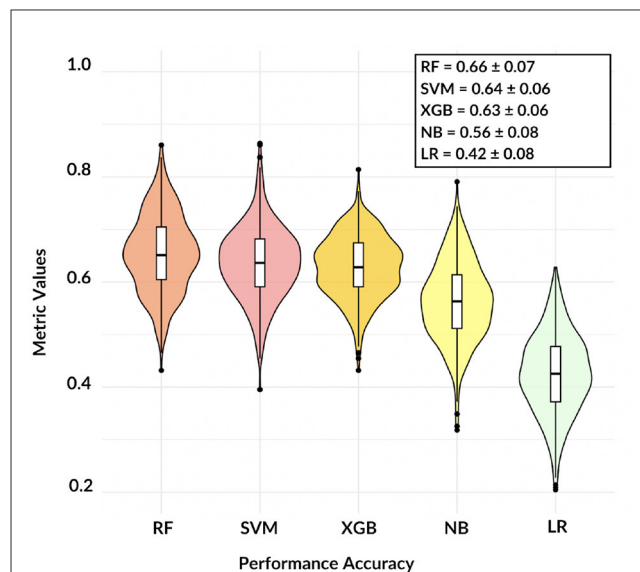


FIGURE 3
The violin plot shows the comparison among five different models, except NB and LR classification remains accurate over 60%.

resulted the best model with a mean three-class accuracy of 0.66 ± 0.07 .

To obtain class-specific evaluations, ROC curves and their area under the curve (AUC) were taken into account (see Figure 4).

The results show an accurate classification for all classes: 0.81 ± 0.02 for AD, 0.77 ± 0.01 for MCI and 0.83 ± 0.01 for NC. The contingency table allows to appreciate how MCI and AD include most of misclassifications, 50 and 46, respectively. We also evaluated the agreement of the predictions of the best performing classifiers: RF and SVM by means of Pearson's correlation. We found a 0.79 correlation for AD, 0.70 for MCI and 0.78 for NC. Hence, for subsequent analyses, only the RF model was considered.

3.2. Characterizing AD, MCI, and NC patterns

Once demonstrated the reliability of the base of knowledge, we investigated whether the features driving classification remained the same despite class-specific differences. In Figure 5, it is possible to observe the overall (related to all three classes) ranking of features; only the top 20 features are shown for display purposes.

The smooth trend demonstrates that it is difficult to find few features dominating over the others. Interestingly, the top three positions were occupied by betweenness; however, the significant presence of both strength and eigenvector centrality features suggested that no metrics prevailed over the others.

Using the three class-specific models, three importance rankings were also constructed; besides, to outline class-related patterns, explainability analyses were carried out using only correctly classified subjects. These class-rankings were compared with the overall ranking by means of the

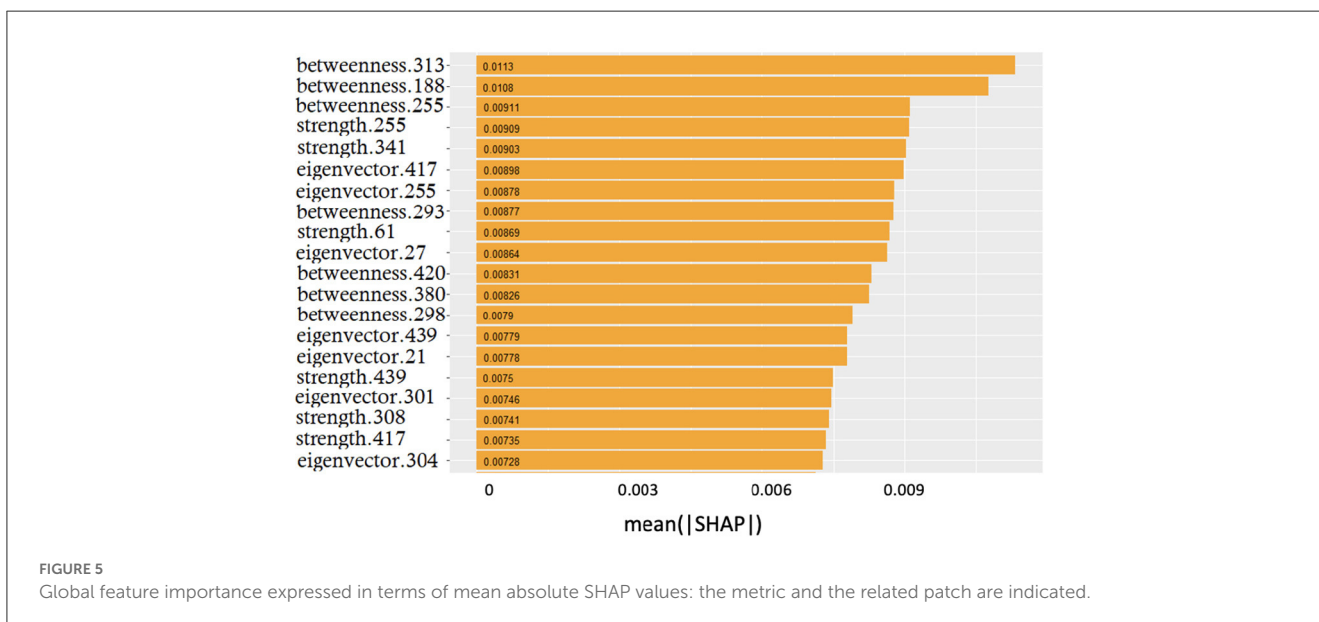
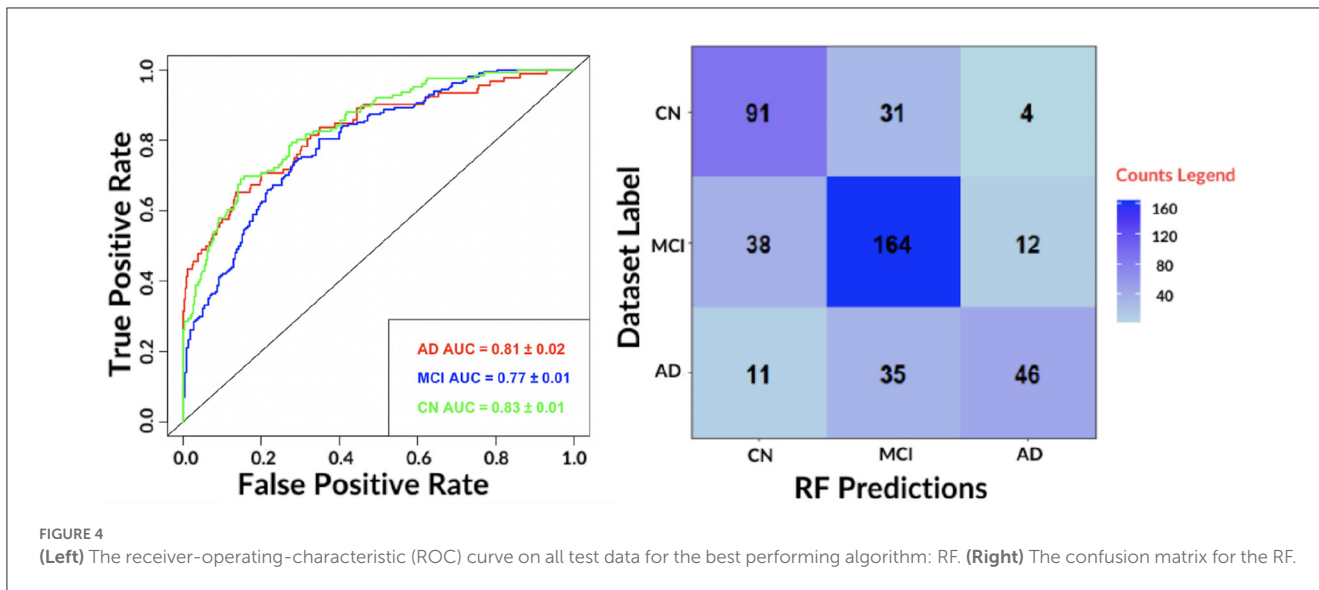


TABLE 2 Agreement between the overall feature importance and the ones retrieved considering the three classes separately.

Overall feature importance vs	Spearman's ρ	p -value
AD feature importance	0.06	0.6
MCI feature importance	0.73	$<2.2 \times 10^{-16}$
NC feature importance	0.59	4×10^{-8}

TABLE 3 Agreement between the feature rankings restricted to correctly classified subjects.

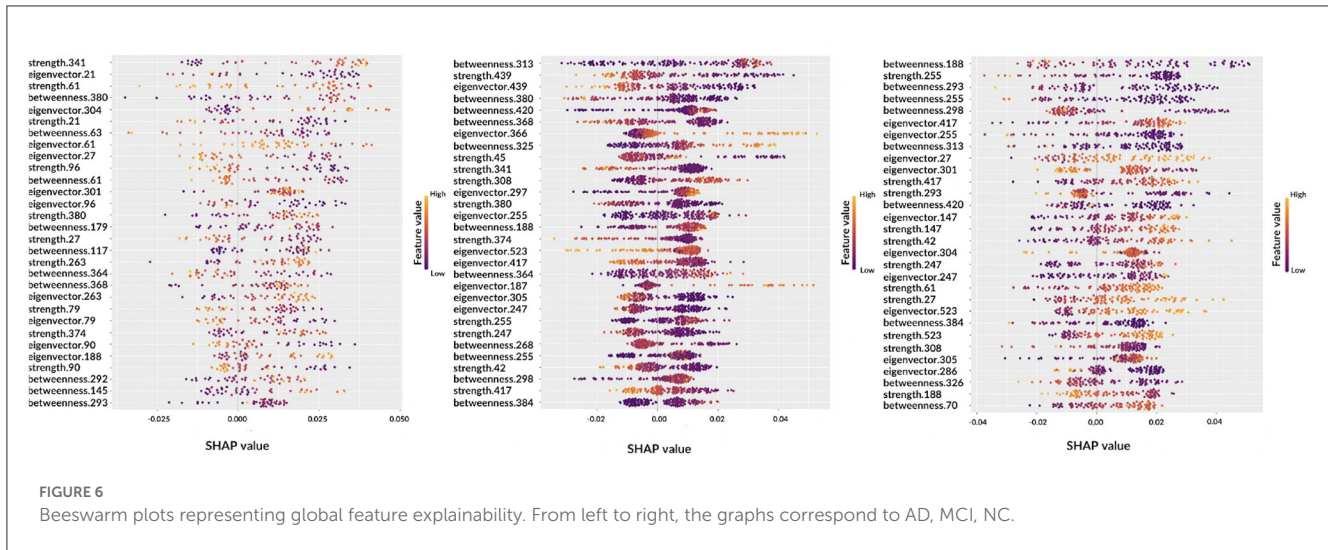
Feature importance between classes	Spearman's ρ	p -value
AD vs. MCI	-0.23	0.04
AD vs. NC	-0.22	0.04
MCI vs. NC	0.08	0.46

Spearman's correlation coefficient ρ . The results are presented in [Table 2](#).

While MCI and NC rankings were significantly correlated to the overall ranking, the AD ranking showed a very weak correlation. These findings allow us to highlight how much brain connectivity is influenced by brain atrophy;

particularly in AD patients, in which the high heterogeneity is more impactful than in subjects of the other two clinical classes.

Lastly, we analyzed the correlation among the three feature importance rankings restricted to only correctly classified subjects (see [Table 3](#)).



The degree of correlation between these rankings allows us to understand to what extent the correct classification of subjects of different classes may depend on the order of importance of the features. Using Spearman’s correlation, it is possible to determine the strength and direction of the monotonic relationship between two ranks. Although the observed correlations are definitely lower than the previous ones, at 5% significance anti-correlation between AD vs. MCI and AD vs. NC is detected. These negative Spearman correlation coefficients highlight a decreasing monotonic trend among the ranks of the features that better correctly classify the AD class compared to those of the other two clinical classes. These correlations, along with the previous ones, suggest the presence of coherent and distinguishable patterns for NC and MCI, whereas a more heterogeneous and elusive condition characterizes the AD class.

3.3. Explaining the observed patterns

SHAP were also used to examine the effects of features on predictions: how and how much each feature affected the prediction score.

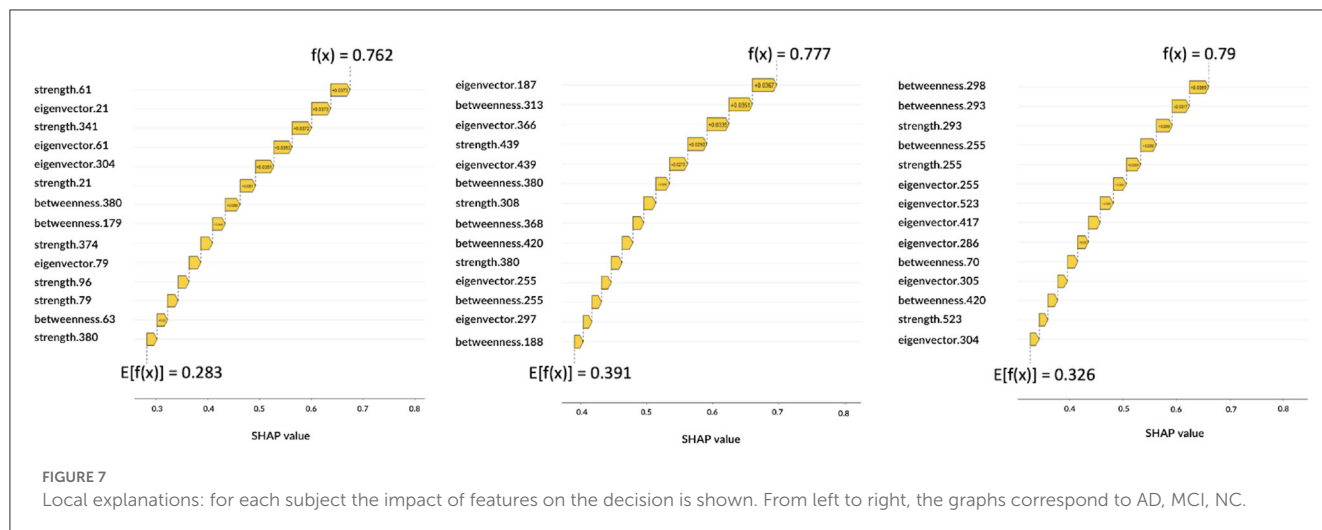
This analysis provided additional information compared to the simple rankings showed in the previous section: the features are ranked from top to bottom, from the most important to the least important, and they are distributed based on both their impact on the prediction as well as their value, for high values, colors tend toward yellow, and for low values, colors tend toward purple. A positive Shapley value represents a positive impact on the correct predicted outcome, and a negative value implies a negative impact (see Figure 6).

Observing the beeswarm plots, the features in common between the three cases are of particular interest. In particular, for AD and MCI the betweenness of patches 364, 368, and 380 were found along with the strength of patches 374 and 380. Analogously, for AD and NC, strength of patches 27 and 61, eigenvector centrality of patches 27, 301, and 304 and betweenness of patch

293 were highlighted. Finally, MCI and NC showed in common: betweenness of patches 188, 255, 298, 313, 384, and 420; eigenvector centrality of patches 247, 255, 305, 417, and 523; strength of patches 42, 247, 255, 308, and 417. Notably, in some cases, the same patch was detected from different metrics.

Let us consider some specific cases. For AD and MCI classes, it is possible to observe the opposite behaviors of betweenness of patches 364 and 380. Thus, these features are important for the correct classification of AD subjects if their values are high; conversely, low values of these features are significant for the correct classification of MCI subjects. Similar findings can be obtained for other classes, for example, NC subjects are characterized by low values of eigenvector centrality of patch 27 while AD subjects tend to show are higher values. Interestingly, the importance of a feature for classification can differ from one class to another, for example the betweenness of patch 293 ranks third for NC subjects, but it is ranked last for AD patients. Understanding this findings without the help of anatomical interpretation can be extremely complicated, this is why in the following section the XAI visualization is presented. Another aspect which deserves to be outlined is the possibility to carry out personalized XAI analyses, which explain for each patient how classification was determined (see Figure 7).

The Figure 7 shows three correctly classified subjects belonging to AD, MCI, and NC, respectively, which correspond to three example cases to show possible differences with global XAI analysis. These graphs are waterfall plots and provide an analysis tool to visualize how each feature impacts on the average of the model output over the training data (baseline value $E[f(x)]$) for a specific subject. The bars represent the Shapley values of the features that most influentially drive the individual prediction. As can be seen from the three plots, positive values indicate a greater impact on the prediction, increasing the baseline predicted score. Longer bars indicate a greater influence of the feature and adding them together we obtain the final prediction $f(x)$; the Figure 7 shows how this value has increased the initial baseline value. It is apparent from these graphs that some features extremely important from a global perspective (e.g., eigenvector centrality of patch 27 or betweenness



of patch 61) and that are essential to the global analysis of classes, do not appear in the class-specific analyses. While global XAI explains on average which features are important for classification, subject-level XAI is of fundamental importance to clinically validate what happens for a specific patient; in anticipation of future developments toward personalized diagnostics.

3.4. From network metrics to brain regions

Finally, to clinically validate the methodology we associated an anatomical district to each important feature (see Figure 8).

The goal was to show the model consistency by demonstrating that the outlined regions are clinically related to AD. For what concerns the overall ranking, the brain regions found were putamen, middle and superior temporal gyrus, anterior cingulate, precentral and postcentral gyrus, insula, sub-gyral, thalamus, culmen, lingual gyrus, cuneus, middle occipital gyrus, and brain stem. Analogously, we found the most relevant brain regions for the class-specific rankings.

There are a number of brain regions that were included in the best 30 for AD, including the amygdala, parahippocampal gyrus, fusiform gyrus, precuneus, posterior cingulate, as well as other regions previously ranked in global rankings (insula and lingual gyrus) that represent additional features nodes. For the MCI class, the following brain regions have been included in the top 30: posterior cingulate, precuneus and others already found in the global ranking, such as middle temporal gyrus, sub-gyral, and insula. The following regions are ranked in the top 30 for the NC class: uncus, parahippocampal gyrus, lateral ventricle, caudate, precuneus, posterior cingulate, and other regions already found in the global ranking (culmen, thalamus).

4. Discussion

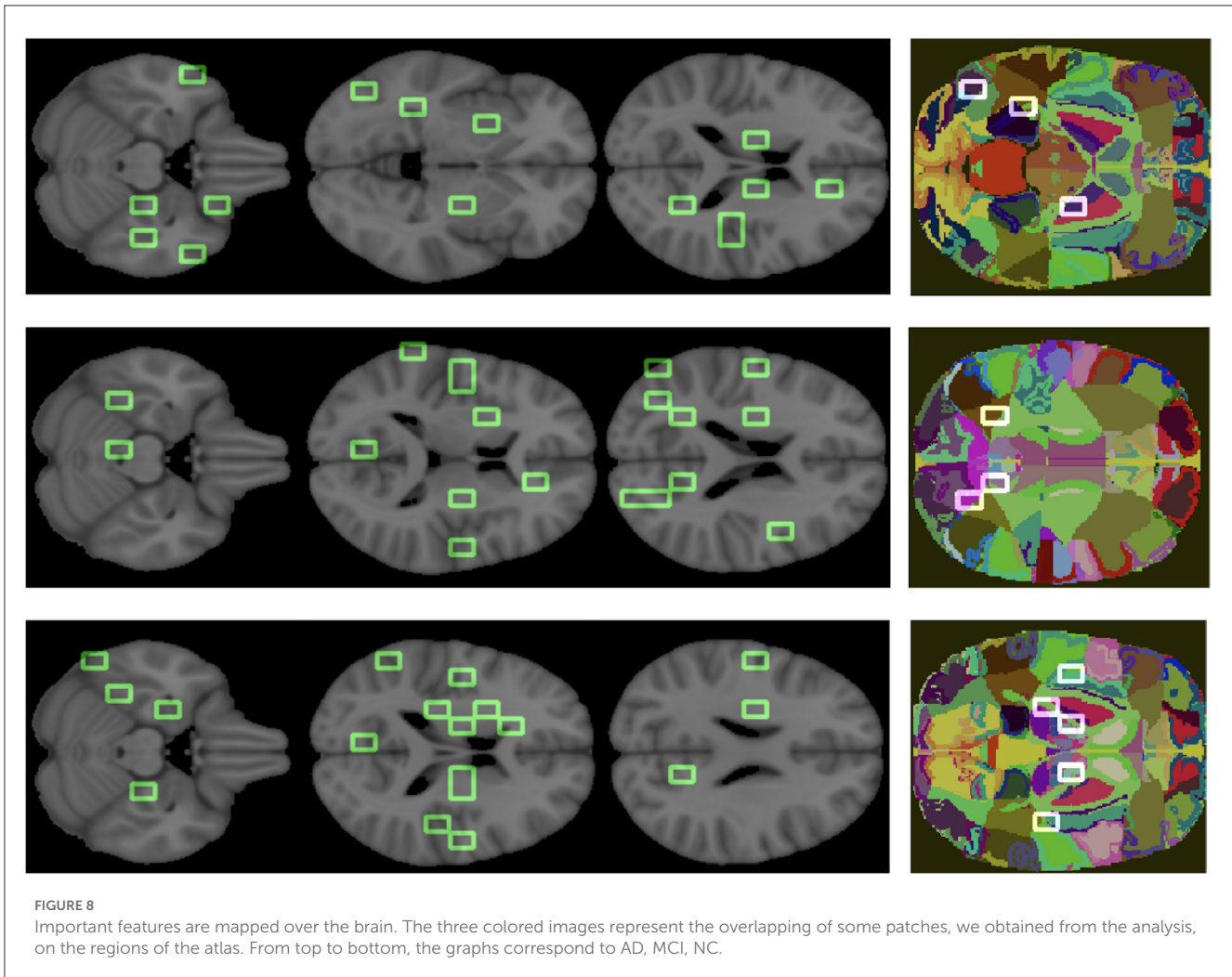
The proposed approach aimed at evaluating brain connectivity as a tool for Alzheimer’s disease staging and its explainability. We set up a three-label classification problem (NC, MCI, AD) in order to estimate whether the brain connectivity model, based

on T1 brain MRI, was able to accurately distinguish these clinical conditions and whether the SHAP could offer support to explain the decision-making of a classifier. Our findings perfectly match performances presented in international challenges (Bron et al., 2015; Amoroso et al., 2018a; Dimitriadis et al., 2018; Sørensen et al., 2018; Lin et al., 2021); the proposed connectivity model can suitably characterize the disease onset and the anticipating cognitive impairment.

We compared several classification models to ensure that the method accuracy depended on the brain connectivity model proposed more than learning algorithms. In fact, we found accuracy levels algorithm independent. Among the different models, the most reliable classifier was RF which was able to achieve a three-class median accuracy of 66% followed by SVM with practically indistinguishable performance. Further, we verified that these models were also in agreement from the prediction point of view, in fact the classification scores achieved a Pearson’s correlation > 0.70 for all three classes.

A large number of studies have achieved impressive results related to AD/NC binary classification in recent years. Using machine learning techniques, the accuracy value could vary between 80 and 90% (Amoroso et al., 2018b; Gupta et al., 2019; Sheng et al., 2021; Zhao et al., 2021). Although excellent, these binary results do not take into consideration the most interesting clinical class, MCI subjects. As a result of the heterogeneity of this intermediate stage, studying MCI patients is quite challenging, but also complex. Three-label classification makes the study much more complex, which explains why accuracy performance is not excessively high compared to binary classifications (Cabral and Silveira, 2013; Sørensen et al., 2014; Cárdenas-Peña et al., 2016; Lama et al., 2017; Lee et al., 2019; Jimenez-Mesa et al., 2020).

Among the five classifiers considered here, the worst results were obtained by NB and LR classifiers. This result could be due to high-dimensional feature space; RF and SVM are more flexible to handle dataset with a high number of features (Badillo et al., 2020; Myszczyńska et al., 2020; Spooner et al., 2020). A further factor that could have affected these results is the small sample size, which could certainly have represented a limitation to the performance of the classifiers. The difficulty of finding many valid clinical data, especially in the field of neuroimaging, makes it necessary to use



high dimensional data with small number of samples and improve the models on these same small datasets (Vabalas et al., 2019). Combining this limit with the need to use many features can make the training sample affected by dataset blind spots, producing model performances that are highly variable compared to the real ones (Berisha et al., 2021). This result suggests the importance of considering algorithms like RF or SVM and, therefore, carrying out explainability analyses to make these models more interpretable for both patients and clinical practitioners. Thus, we investigated the influence of different features on classification using the XAI approach based on SHAP. By using this methodology, we can gain a better understanding of how algorithms work and how artificial intelligence methods can be used in clinical practice. As a means of evaluating how features affect RF classifier predictions, we conducted a global analysis that included all subjects in the dataset and compared it with the three class-specific analyses that included only patients correctly classified. We found a significant agreement between global feature importance and the class-specific one of MCI and NC classes, while the association tends to be weaker for the AD ranking. This result confirms how AD brain atrophy affects connectivity in highly heterogeneous ways (Zhang et al., 2016; John et al., 2017; Poulakis et al., 2018; Sui et al., 2018; Badhwar et al.,

2020). Moreover, by comparing the three class-specific rankings we found no significant correlation; these findings emphasize again the disease heterogeneity and show how brain connectivity within each class follows substantially different patterns (John et al., 2017; Khazaei et al., 2017; Yu et al., 2017; Sheng et al., 2019).

Since the SHAP method provides information for analyzing personalized predictions about individual patients, we also presented some personalized analyses as an example. This is particularly important to envisage future developments toward personalized diagnostics and treatment (Fellous et al., 2019; Van der Velden et al., 2022; Vrahatis et al., 2023). Moreover, this is also important as our findings confirm that class-specific rankings can significantly differ from global importance: it is not possible to conceive a subject-specific model of AD without considering the subject's peculiarity. Finally, explainability analysis also provides a way clinically validate the proposed brain connectivity model in that it allows to directly relate the features driving classification to specific brain regions. In particular, our findings demonstrate that connectivity metrics reveal brain regions, such as parahippocampal gyrus, amygdala, uncus, fusiform gyrus and lateral ventricle, whose relation with AD is established (Pearson et al., 1985; Schill et al., 2002; Zhang and Wang, 2015; Amoroso et al., 2018b).

5. Conclusion

In this work, a structural brain connectivity model was proposed to study Alzheimer's disease and mild cognitive impairment; to this aim an accurate three-label classification model was designed and, based on its decision scores, an explainable SHAP approach was implemented. To the best of our knowledge, this is the first attempt to equip a novel patch-based connectivity model with an XAI framework. Moreover, to ease interpretability, SHAP were mapped onto a brain atlas; the main advantage of this method, in fact, is that it allows to directly relate mathematical graph entities to anatomical districts. The accuracy and the robustness of the network model were assessed by comparing several classifiers. XAI analyses provided additional information about both global and patient-level explanations; in particular our findings confirmed that despite the presence of well determined regions related to AD, each patient deserves specific attention in that the disease heterogeneity makes its patterns extremely varying. Future studies could investigate how this variability reflects the specific brain resilience to the disease and, therefore, the possibility to design stage-specific therapies. Moreover, while here the wide MCI class was considered, further studies could also investigate differences within this class. The proposed approach is general and it could be hopefully applied to shed light over several pathologies; besides, its interpretability could ease its adoption in a domain where learning algorithms are mistrustfully seen as black boxes.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

Ethics statement

All experiments were performed with the informed consent of each participant or caregiver in line with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Local institutional Ethics Committees approved the study.

Author contributions

NA designed the experiments. SQ and ML carried out the analyses. NA and SQ wrote the original draft. AM and RB supervised the analyses. All authors revised and approved the manuscript.

Funding

Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4—Call for tender No. 3138 of 16 December 2021 of Italian Ministry of University and Research funded by the European Union—NextGenerationEU. Award Number:

Project code: CN00000013, Concession Decree No. 1031 of 17 February 2022 adopted by the Italian Ministry of University and Research, CUP H93C22000450007, Project title: National Centre for HPC, Big Data and Quantum Computing.

Acknowledgments

Data used in the preparation of this article was obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a 60 million, 5 year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is M. W. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55–90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2, and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information (see www.adni-info.org). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI was funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N V; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support

ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Alzheimer's Association (2020). 2020 Alzheimer's disease facts and figures. *Alzheimer Dement.* 16, 391–460. doi: 10.1002/alz.12068
- Alzheimer's Association (2022). 2022 Alzheimer's disease facts and figures. *Alzheimer Dement.* 18, 700–789. doi: 10.1002/alz.12638
- Amoroso, N., Bellantuono, L., Monaco, A., De Nicolò, F., Somma, E., and Bellotti, R. (2021). Economic interplay forecasting business success. *Complexity* 2021, 1–12. doi: 10.1155/2021/8861267
- Amoroso, N., Diacono, D., Fanizzi, A., La Rocca, M., Monaco, A., Lombardi, A., et al. (2018a). Deep learning reveals Alzheimer's disease onset in mci subjects: results from an international challenge. *J. Neurosci. Methods* 302, 3–9. doi: 10.1016/j.jneumeth.2017.12.011
- Amoroso, N., Errico, R., Bruno, S., Chincarini, A., Garuccio, E., Sensi, F., et al. (2015). Hippocampal unified multi-atlas network (human): protocol and scale validation of a novel segmentation tool. *Phys. Med. Biol.* 60, 8851. doi: 10.1088/0031-9155/60/22/8851
- Amoroso, N., La Rocca, M., Bellantuono, L., Diacono, D., Fanizzi, A., Lella, E., et al. (2019). Deep learning and multiplex networks for accurate modeling of brain age. *Front. Aging Neurosci.* 11, 115. doi: 10.3389/fnagi.2019.00115
- Amoroso, N., La Rocca, M., Bruno, S., Maggipinto, T., Monaco, A., Bellotti, R., et al. (2018b). Multiplex networks for early diagnosis of Alzheimer's disease. *Front. Aging Neurosci.* 10, 365. doi: 10.3389/fnagi.2018.00365
- Amoroso, N., La Rocca, M., Monaco, A., Bellotti, R., and Tangaro, S. (2018c). Complex networks reveal early mri markers of Parkinson's disease. *Med. Image Anal.* 48, 12–24. doi: 10.1016/j.media.2018.05.004
- Anjomshoae, S., and Pudas, S. (2022). Explaining graph convolutional network predictions for clinicians-an explainable ai approach to Alzheimer's disease classification. *Electronic J. doi: 10.2139/ssrn.4194675*
- Badhwar, A., McFall, G. P., Sapkota, S., Black, S. E., Chertkow, H., Duchesne, S., et al. (2020). A multiomics approach to heterogeneity in Alzheimer's disease: focused review and roadmap. *Brain* 143, 1315–1331. doi: 10.1093/brain/awz384
- Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., et al. (2020). An introduction to machine learning. *Clin. Pharmacol. Ther.* 107, 871–885. doi: 10.1002/cpt.1796
- Beebe-Wang, N., Okeson, A., Althoff, T., and Lee, S.-I. (2021). Efficient and explainable risk assessments for imminent dementia in an aging cohort study. *IEEE J. Biomed. Health Inf.* 25, 2409–2420. doi: 10.1109/JBHI.2021.3059563
- Bellantuono, L., Marzano, L., La Rocca, M., Duncan, D., Lombardi, A., Maggipinto, T., et al. (2021). Predicting brain age with complex networks: from adolescence to adulthood. *Neuroimage* 225, 117458. doi: 10.1016/j.neuroimage.2020.117458
- Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., et al. (2021). Digital medicine and the curse of dimensionality. *NPJ Dig. Med.* 4, 153. doi: 10.1038/s41746-021-00521-5
- Biecek, P. (2018). Dalex: explainers for complex predictive models in R. *J. Mach. Learn. Res.* 19, 3245–3249. Available online at: <http://jmlr.org/papers/v19/18-416.html>
- Biecek, P., and Burzykowski, T. (2020). *Explanatory Model Analysis*. Available online at: <https://ema.drwhy.ai/shapley.html#SHAPMethod> (accessed April 28, 2023).
- Brejijeh, Z., and Karaman, R. (2020). Comprehensive review on Alzheimer's disease: causes and treatment. *Molecules* 25, 5789. doi: 10.3390/molecules25245789
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324
- Bron, E. E., Smits, M., Van Der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., et al. (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: the caddementia challenge. *Neuroimage* 111, 562–579. doi: 10.1016/j.neuroimage.2015.01.048
- Bullmore, E., and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10, 186–198. doi: 10.1038/nrn2575
- Byun, M. S., Kim, S. E., Park, J., Yi, D., Choe, Y. M., Sohn, B. K., et al. (2015). Heterogeneity of regional brain atrophy patterns associated with distinct progression rates in Alzheimer's disease. *PLoS ONE* 10, e0142756. doi: 10.1371/journal.pone.0142756
- Cabral, C., Morgado, P. M., Costa, D. C., Silveira, M., s Disease Neuroimaging Initiative, A., et al. (2015). Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages. *Comput. Biol. Med.* 58, 101–109. doi: 10.1016/j.compbiomed.2015.01.003
- Cabral, C., and Silveira, M. (2013). "Classification of Alzheimer's disease from FDG-PET images using favourite class ensembles," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Osaka: IEEE), 2477–2480.
- Cárdenas-Peña, D., Collazos-Huertas, D., and Castellanos-Dominguez, G. (2016). Centered kernel alignment enhancing neural network pretraining for mri-based dementia diagnosis. *Comput. Math. Methods Med.* 2016. doi: 10.1155/2016/9523849
- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015). *Xgboost: Extreme Gradient Boosting. R Package Version 0.4-2 1*, 1–4. Available online at: https://scholar.google.com/scholar?hl=it&as_sdt=0%2C5&q=Xgboost%3A+extreme+gradient+boosting+2015&btnG=
- Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018
- Daianu, M., Jahanshad, N., Nir, T. M., Jack C. R. Jr., Weiner, M. W., Bernstein, M. A., et al. (2015). Rich club analysis in the Alzheimer's disease connectome reveals a relatively undisturbed structural core network. *Hum. Brain Mapp.* 36, 3087–3103. doi: 10.1002/hbm.22830
- de Haan, W., van der Flier, W. M., Koene, T., Smits, L. L., Scheltens, P., and Stam, C. J. (2012). Disrupted modular brain dynamics reflect cognitive dysfunction in Alzheimer's disease. *Neuroimage* 59, 3085–3093. doi: 10.1016/j.neuroimage.2011.11.055
- Devanand, D., Pradhaban, G., Liu, X., Khandji, A., De Santi, S., Segal, S., et al. (2007). Hippocampal and entorhinal atrophy in mild cognitive impairment: prediction of Alzheimer disease. *Neurology* 68, 828–836. doi: 10.1212/01.wnl.0000256697.20968.d7
- Dimitriadis, S. I., Liparas, D., Tsolaki, M. N., Initiative, A. D. N., et al. (2018). Random forest feature selection, fusion and ensemble strategy: combining multiple morphological MRI measures to discriminate among healthy elderly, MCI, cMCI and Alzheimer's disease patients: from the Alzheimer's disease neuroimaging initiative (ADNI) database. *J. Neurosci. Methods* 302, 14–23. doi: 10.1016/j.jneumeth.2017.12.010

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnagi.2023.1238065/full#supplementary-material>

- Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Nat. Acad. Sci. U. S. A.* 113, 7900–7905. doi: 10.1073/pnas.1602413113
- El-Sappagh, S., Alonso, J. M., Islam, S., Sultan, A. M., and Kwak, K. S. (2021). A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Sci. Rep.* 11, 1–26. doi: 10.1038/s41598-021-82098-3
- Emre, M., Aarsland, D., Brown, R., Burn, D. J., Duyckaerts, C., Mizuno, Y., et al. (2007). Clinical diagnostic criteria for dementia associated with Parkinson's disease. *Mov. Disord.* 22, 1689–1707. doi: 10.1002/mds.21507
- Farahani, F. V., Fiok, K., Lahijanian, B., Karwowski, W., and Douglas, P. K. (2022). Explainable AI: a review of applications to neuroimaging data. *Front. Neurosci.* 16, 906290. doi: 10.3389/fnins.2022.906290
- Fellous, J.-M., Sapiro, G., Rossi, A., Mayberg, H., and Ferrante, M. (2019). Explainable artificial intelligence for neuroscience: behavioral neurostimulation. *Front. Neurosci.* 13, 1346. doi: 10.3389/fnins.2019.01346
- Frisoni, G. B., Fox, N. C., Jack Jr, C. R., Scheltens, P., and Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nat. Rev. Neurol.* 6, 67–77. doi: 10.1038/nrneuro.2009.215
- Galazzo, I. B., Cruciani, F., Brusini, L., Salih, A., Radeva, P., Storti, S. F., et al. (2022). Explainable artificial intelligence for magnetic resonance imaging aging brainprints: grounds and challenges. *IEEE Signal Process. Mag.* 39, 99–116. doi: 10.1109/MSP.2021.3126573
- Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., et al. (2006). Mild cognitive impairment. *Lancet* 367, 1262–1270. doi: 10.1016/S0140-6736(06)68542-5
- Goenka, N., and Tiwari, S. (2022). Alzynet: A volumetric convolutional neural network for multiclass classification of Alzheimer's disease through multiple neuroimaging computational approaches. *Biomed. Signal Process. Control* 74, 103500. doi: 10.1016/j.bspc.2022.103500
- Górriz, J.-M., Jiménez-Mesa, C., Segovia, F., Ramirez, J., Group, S., and Suckling, J. (2021). A connection between pattern classification by machine learning and statistical inference with the general linear model. *IEEE J. Biomed. Health Inf.* 26, 5332–5343. doi: 10.1109/JBHI.2021.3101662
- Gupta, Y., Lee, K. H., Choi, K. Y., Lee, J. J., Kim, B. C., Kwon, G. R., et al. (2019). Early diagnosis of Alzheimer's disease using combined features from voxel-based morphometry and cortical, subcortical, and hippocampus regions of MRI T1 brain images. *PLoS ONE* 14, e0222446. doi: 10.1371/journal.pone.0222446
- Huang, L.-K., Chao, S.-P., and Hu, C.-J. (2020). Clinical trials of new drugs for Alzheimer disease. *J. Biomed. Sci.* 27, 1–13. doi: 10.1186/s12929-019-0609-7
- Huckans, M., Hutson, L., Twamley, E., Jak, A., Kaye, J., and Storzbach, D. (2013). Efficacy of cognitive rehabilitation therapies for mild cognitive impairment (MCI) in older adults: working toward a theoretical model and evidence-based interventions. *Neuropsychol. Rev.* 23, 63–80. doi: 10.1007/s11065-013-9230-9
- Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17, 825–841. doi: 10.1006/nimg.2002.1132
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). FSL. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015
- Jimenez-Mesa, C., Arco, J. E., Valenti-Soler, M., Frades-Payo, B., Zea-Sevilla, M. A., Ortiz, A., et al. (2023). Using explainable artificial intelligence in the clock drawing test to reveal the cognitive impairment pattern. *Int. J. Neural Syst.* 33, 2350015. doi: 10.1142/S0129065723500156
- Jimenez-Mesa, C., Illán, I. A., Martin-Martin, A., Castillo-Barnes, D., Martinez-Murcia, F. J., Ramirez, J., et al. (2020). Optimized one vs. one approach in multiclass classification for early Alzheimer's disease and mild cognitive impairment diagnosis. *IEEE Access* 8, 96981–96993. doi: 10.1109/ACCESS.2020.2997736
- John, M., Ikuta, T., and Ferbinteanu, J. (2017). Graph analysis of structural brain networks in Alzheimer's disease: beyond small world properties. *Brain Struct. Funct.* 222, 923–942. doi: 10.1007/s00429-016-1255-4
- Julkunen, V., Niskanen, E., Koikkalainen, J., Herukka, S.-K., Pihlajamäki, M., Hallikainen, M., et al. (2010). Differences in cortical thickness in healthy controls, subjects with mild cognitive impairment, and Alzheimer's disease patients: a longitudinal study. *J. Alzheimers Dis.* 21, 1141–1151. doi: 10.3233/JAD-2010-100114
- Khazaei, A., Ebrahimzadeh, A., Babajani-Feremi, A., Initiative, A. D. N., et al. (2017). Classification of patients with MCI and AD from healthy controls using directed graph measures of resting-state fMRI. *Behav. Brain Res.* 322, 339–350. doi: 10.1016/j.bbr.2016.06.043
- Kim, H., Irimia, A., Hobel, S. M., Pogoyan, M., Tang, H., Petrosyan, P., et al. (2019). The LONI QC system: a semi-automated, web-based and freely-available environment for the comprehensive quality control of neuroimaging data. *Front. Neuroinf.* 13, 60. doi: 10.3389/fninf.2019.00060
- Kim, J., Lee, M., Lee, M. K., Wang, S.-M., Kim, N.-Y., Kang, D. W., et al. (2021). Development of random forest algorithm based prediction model of Alzheimer's disease using neurodegeneration pattern. *Psychiatry Investig.* 18, 69. doi: 10.30773/pi.2020.0304
- La Rocca, M., Amoroso, N., Bellotti, R., Diacono, D., Monaco, A., Monda, A., et al. (2017). "A multiplex network model to characterize brain atrophy in structural MRI," in *Emergent Complexity from Nonlinearity, in Physics, Engineering and the Life Sciences: Proceedings of the XXIII International Conference on Nonlinear Dynamics of Electronic Systems, Como, Italy, 7-11 September 2015* (Springer), 189–198.
- La Rocca, M., Amoroso, N., Monaco, A., Bellotti, R., Tangaro, S., Initiative, A. D. N., et al. (2018). A novel approach to brain connectivity reveals early structural changes in Alzheimer's disease. *Physiol. Meas.* 39, 074005. doi: 10.1088/1361-6579/aaef1f
- Lama, R. K., Gwak, J., Park, J.-S., and Lee, S.-W. (2017). Diagnosis of Alzheimer's disease based on structural MRI images using a regularized extreme learning machine and PCA features. *J. Healthc. Eng.* 2017, 5485080. doi: 10.1155/2017/5485080
- Lancaster, J., Rainey, L., Summerlin, J., Freitas, C., Fox, P., Evans, A., et al. (1997). Automated labeling of the human brain: a preliminary report on the development and evaluation of a forward-transform method. *Hum. Brain Mapp.* 5, 238–242. doi: 10.1002/(SICI)1097-0193(1997)5:4<238::AID-HBM6>3.0.CO;2-4
- Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., et al. (2000). Automated talairach atlas labels for functional brain mapping. *Hum. Brain Mapp.* 10, 120–131. doi: 10.1002/1097-0193(200007)10:3<120::AID-HBM30>3.0.CO;2-8
- Larrieu, S., Letenneur, L., Orgogozo, J., Fabrigoule, C., Amieva, H., Le Carret, N., et al. (2002). Incidence and outcome of mild cognitive impairment in a population-based prospective cohort. *Neurology* 59, 1594–1599. doi: 10.1212/01.WNL.0000034176.07159.F8
- Lebedeva, A. K., Westman, E., Borza, T., Beyer, M. K., Engedal, K., Aarsland, D., et al. (2017). MRI-based classification models in prediction of mild cognitive impairment and dementia in late-life depression. *Front. Aging Neurosci.* 9, 13. doi: 10.3389/fnagi.2017.00013
- Lee, E., Choi, J.-S., Kim, M., Suk, H.-I., Initiative, A. D. N., et al. (2019). Toward an interpretable Alzheimer's disease diagnostic model with regional abnormality representation via deep learning. *Neuroimage* 202, 116113. doi: 10.1016/j.neuroimage.2019.116113
- Leuch, J. P., Pruessner, J., Zijdenbos, A. P., Collins, D. L., Teipel, S. J., Hampel, H., et al. (2008). Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. *Neurobiol. Aging* 29, 23–30. doi: 10.1016/j.neurobiolaging.2006.09.013
- Lin, W., Gao, Q., Du, M., Chen, W., and Tong, T. (2021). Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data. *Comput. Biol. Med.* 134, 104478. doi: 10.1016/j.compbiomed.2021.104478
- Liu, J., Li, M., Pan, Y., Lan, W., Zheng, R., Wu, F.-X., et al. (2017). Complex brain network analysis and its applications to brain disorders: a survey. *Complexity* 2017, 8362741. doi: 10.1155/2017/8362741
- Liu, P.-P., Xie, Y., Meng, X.-Y., and Kang, J.-S. (2019). History and progress of hypotheses and clinical trials for Alzheimer's disease. *Signal Transd. Target. Therapy* 4, 29. doi: 10.1038/s41392-019-0063-8
- Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., and Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022). *Comp. Methods Progr. Biomed.* 226, 107161. doi: 10.1016/j.cmpb.2022.107161
- Lombardi, A., Diacono, D., Amoroso, N., Biecek, P., Monaco, A., Bellantuono, L., et al. (2022). A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of mild cognitive impairment and Alzheimer's disease. *Brain Informat.* 9, 1–17. doi: 10.1186/s40708-022-00165-5
- Lundberg, S. M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* 30, 4765–4774. Available online at: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
- Messalas, A., Kanellopoulos, Y., and Makris, C. (2019). "Model-agnostic interpretability with shapley values," in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (Patras: IEEE), 1–7.
- Michael, M. (2022). SHAP Visualizations. Available online at: <https://cran.r-project.org/web/packages/shapviz/shapviz.pdf> (accessed February 15, 2023).
- Morabito, F. C., Ieracitano, C., and Mammone, N. (2023). An explainable artificial intelligence approach to study MCI to AD conversion via HD-EEG processing. *Clin. EEG Neurosci.* 54, 51–60. doi: 10.1177/15500594211063662
- Myszczyńska, M. A., Ojames, P. N., Lacoste, A. M., Neil, D., Saffari, A., Mead, R., et al. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat. Rev. Neurol.* 16, 440–456. doi: 10.1038/s41582-020-0377-8
- National Institute on Aging (2002). *Alzheimer's Disease: Unraveling the Mystery, Vol. 1*. National Institutes of Health. Available online at: https://scholar.google.com/scholar?q=Alzheimer%27s+disease%3A+unraveling+the+mystery&hl=it&as_sdt=0%2C5&as_ylo=2000&as_yhi=2002
- Ng, A., and Jordan, M. (2001). On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. *Adv. Neural Inf. Process.*

- Syst. 14, 841–848. Available online at: <https://proceedings.neurips.cc/paper/2001/hash/7b7a53e239400a13bd6be6c91c4f6c4e-Abstract.html>
- Nichols, E., Steinmetz, J. D., Vollset, S. E., Fukutaki, K., Chalek, J., Abd-Allah, F., et al. (2022). Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019. *Lancet Public Health* 7, e105–e125. doi: 10.1016/S2468-2667(21)00249-8
- Pearson, R. C., Esiri, M., Hiorns, R., Wilcock, G., and Powell, T. (1985). Anatomical correlates of the distribution of the pathological changes in the neocortex in Alzheimer disease. *Proc. Nat. Acad. Sci. U. S. A.* 82, 4531–4534. doi: 10.1073/pnas.82.13.4531
- Petersen, R. C., Caracciolo, B., Brayne, C., Gauthier, S., Jelic, V., and Fratiglioni, L. (2014). Mild cognitive impairment: a concept in evolution. *J. Intern. Med.* 275, 214–228. doi: 10.1111/joim.12190
- Poulakis, K., Pereira, J. B., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., et al. (2018). Heterogeneous patterns of brain atrophy in Alzheimer's disease. *Neurobiol. Aging* 65, 98–108. doi: 10.1016/j.neurobiolaging.2018.01.009
- Qian, J., Li, H., Wang, J., and He, L. (2023). Recent advances in explainable artificial intelligence for magnetic resonance imaging. *Diagnostics* 13, 1571. doi: 10.3390/diagnostics13091571
- Rish, I. (2001). "An empirical study of the naive bayes classifier," in *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Vol. 3* (IBM New York), 41–46.
- Scahill, R. L., Schott, J. M., Stevens, J. M., Rossor, M. N., and Fox, N. C. (2002). Mapping the evolution of regional atrophy in Alzheimer's disease: unbiased analysis of fluid-registered serial MRI. *Proc. Nat. Acad. Sci. U. S. A.* 99, 4703–4707. doi: 10.1073/pnas.052587399
- Sheng, J., Wang, B., Zhang, Q., Liu, Q., Ma, Y., Liu, W., et al. (2019). A novel joint hcpmp method for automatically classifying Alzheimer's and different stage MCI patients. *Behav. Brain Res.* 365, 210–221. doi: 10.1016/j.bbr.2019.03.004
- Sheng, J., Wang, B., Zhang, Q., Zhou, R., Wang, L., and Xin, Y. (2021). Identifying and characterizing different stages toward Alzheimer's disease using ordered core features and machine learning. *Heliyon* 7, e07287. doi: 10.1016/j.heliyon.2021.e07287
- Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155. doi: 10.1002/hbm.10062
- Solé-Padullés, C., Bartrés-Faz, D., Junqué, C., Vendrell, P., Rami, L., Clemente, I. C., et al. (2009). Brain structure and function related to cognitive reserve variables in normal aging, mild cognitive impairment and Alzheimer's disease. *Neurobiol. Aging* 30, 1114–1124. doi: 10.1016/j.neurobiolaging.2007.10.008
- Song, M., Jung, H., Lee, S., Kim, D., and Ahn, M. (2021). Diagnostic classification and biomarker identification of Alzheimer's disease with random forest algorithm. *Brain Sci.* 11, 453. doi: 10.3390/brainsci11040453
- Sørensen, L., Nielsen, M., Alzheimer's Disease Neuroimaging Initiative. (2018). Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination. *J. Neurosci. Methods* 302, 66–74. doi: 10.1016/j.jneumeth.2018.01.003
- Sørensen, L., Pai, A., Anker, C., Balas, I., Lillholm, M., Igel, C., et al. (2014). "Dementia diagnosis using mri cortical thickness, shape, texture, and volumetry," in *Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data* (Boston), 111–118.
- Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., et al. (2011). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7, 280–292. doi: 10.1016/j.jalz.2011.03.003
- Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., et al. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Sci. Rep.* 10, 20410. doi: 10.1038/s41598-020-77220-w
- Strumbelj, E., and Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res.* 11, 1–18. Available online at: <http://jmlr.org/papers/v11/strumbelj10a.html>
- Sui, X., Rajapakse, J. C., Initiative, A. D. N., et al. (2018). Profiling heterogeneity of Alzheimer's disease using white-matter impairment factors. *NeuroImage Clin.* 20, 1222–1232. doi: 10.1016/j.nicl.2018.10.026
- Tábuas-Pereira, M., Baldeiras, I., Duro, D., Santiago, B., Ribeiro, M. H., Leitão, M. J., et al. (2016). Prognosis of early-onset vs. late-onset mild cognitive impairment: comparison of conversion rates and its predictors. *Geriatrics* 1, 11. doi: 10.3390/geriatrics1020011
- Talwar, P., Kushwaha, S., Chaturvedi, M., and Mahajan, V. (2021). Systematic review of different neuroimaging correlates in mild cognitive impairment and Alzheimer's disease. *Clin. Neuroradiol.* 31, 953–967. doi: 10.1007/s00062-021-01057-7
- Teipel, S., Grothe, M. J., Zhou, J., Sepulcre, J., Dyrba, M., Sorg, C., et al. (2016). Measuring cortical connectivity in Alzheimer's disease as a brain neural network pathology: toward clinical applications. *J. Int. Neuropsychol. Soc.* 22, 138–163. doi: 10.1017/S155617715000995
- Tijms, B. M., Wink, A. M., de Haan, W., van der Flier, W. M., Stam, C. J., Scheltens, P., et al. (2013). Alzheimer's disease: connecting findings from graph theoretical studies of brain networks. *Neurobiol. Aging* 34, 2023–2036. doi: 10.1016/j.neurobiolaging.2013.02.020
- Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE* 14, e0224365. doi: 10.1371/journal.pone.0224365
- Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., and Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.* 79, 102470. doi: 10.1016/j.media.2022.102470
- Vemuri, P., Gunter, J. L., Senjem, M. L., Whitwell, J. L., Kantarci, K., Knopman, D. S., et al. (2008). Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. *Neuroimage* 39, 1186–1197. doi: 10.1016/j.neuroimage.2007.09.073
- Vrahatis, A. G., Skolariki, K., Krokidis, M. G., Lazaros, K., Exarchos, T. P., and Vlamos, P. (2023). Revolutionizing the early detection of Alzheimer's disease through non-invasive biomarkers: the role of artificial intelligence and deep learning. *Sensors* 23, 4184. doi: 10.3390/s23094184
- Whitwell, J. L., Petersen, R. C., Negash, S., Weigand, S. D., Kantarci, K., Ivnik, R. J., et al. (2007). Patterns of atrophy differ among specific subtypes of mild cognitive impairment. *Arch. Neurol.* 64, 1130–1138. doi: 10.1001/archneur.64.8.1130
- Yu, M., Engels, M. M., Hillebrand, A., Van Straaten, E. C., Gouw, A. A., Teunissen, C., et al. (2017). Selective impairment of hippocampus and posterior hub areas in Alzheimer's disease: an MEG-based multiplex network study. *Brain* 140, 1466–1485. doi: 10.1093/brain/awx050
- Zeng, H.-M., Han, H.-B., Zhang, Q.-F., and Bai, H. (2021). Application of modern neuroimaging technology in the diagnosis and study of Alzheimer's disease. *Neural Regen. Res.* 16, 73. doi: 10.4103/1673-5374.286957
- Zhang, X., Mormino, E. C., Sun, N., Sperling, R. A., Sabuncu, M. R., Yeo, B. T., et al. (2016). Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. *Proc. Nat. Acad. Sci. U. S. A.* 113, E6535–E6544. doi: 10.1073/pnas.1611073113
- Zhang, Y., and Wang, S. (2015). Detection of Alzheimer's disease by displacement field and machine learning. *PeerJ* 3, e1251. doi: 10.7717/peerj.1251
- Zhao, X., Ang, C. K. E., Acharya, U. R., and Cheong, K. H. (2021). Application of artificial intelligence techniques for the detection of Alzheimer's disease using structural MRI images. *Biocybernet. Biomed. Eng.* 41, 456–473. doi: 10.1016/j.bbe.2021.02.006