Journal of Neuroscience Methods 413 (2025) 110318

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

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

Unveiling the decision making process in Alzheimer's disease diagnosis: A case-based counterfactual methodology for explainable deep learning

Adarsh Valoor [a], G.R. Gangadharan [b,*]

[a] Department of Electronics and Computer Science, University of Southampton, United Kingdom
[b] National Institute of Technology, Tiruchirappalli, India

ARTICLE INFO

Keywords:
Explainable Artificial Intelligence
Case Based Reasoning
Counterfactual Analysis
Alzheimer's Disease
Deep Learning

ABSTRACT

Background: The field of Alzheimer's disease (AD) diagnosis is undergoing significant transformation due to the application of deep learning (DL) models. While DL surpasses traditional machine learning in disease prediction from structural magnetic resonance imaging (sMRI), the lack of explainability limits clinical adoption. Counterfactual inference offers a way to integrate causal explanations into these models, enhancing their robustness and transparency.

New method: This study develops a novel methodology combining U-Net and generative adversarial network (GAN) models to create comprehensive counterfactual diagnostic maps for AD. The proposed methodology uses case-based counterfactual reasoning for robust decision classification for counterfactual maps to understand how changes in specific features affect the model's predictions.

Comparison with existing methods: The proposed methodology is compared with state-of-the-art visual explanation techniques across the ADNI dataset. The proposed methodology is also benchmarked against other gradient-based and generative models for its ability to generate comprehensive counterfactual maps.

Results: The results demonstrate that the proposed methodology significantly outperforms existing methods in accuracy, sensitivity, and specificity while providing detailed counterfactual maps that visualize how slight changes in brain morphology could lead to different diagnostic outcomes. The proposed methodology achieves an accuracy of 95 % and an AUC of 0.97, illustrating its superiority in capturing subtle yet crucial anatomical features.

Conclusions: By generating intuitive visual explanations, the proposed methodology improves the interpretability and robustness of AD diagnostic models, making them more reliable and accountable. The use of counterfactual inference enhances clinicians' understanding of disease progression and the impact of different interventions, fostering precision medicine in AD care.

1. Introduction

Alzheimer's disease (AD) is a chronic neurological disorder that significantly alters cognitive functions, notably affecting memory, thought processes, and behavior. The disease is characterized by the accumulation of beta-amyloid plaques and tau protein tangles in the brain, which disrupt neuronal signaling. This disruption precipitates a cascade of detrimental effects, including neuronal death, cognitive decline, and loss of functional abilities. AD typically manifests as mild memory loss and confusion, progressively advancing to severe dementia and a loss of basic bodily functions, ultimately leading to death. The condition presents not only a medical challenge but also represents a complex, yet not fully understood, system (Gao et al., 2016; Selkoe, Hardy, 2016; Lindner et al., 2008).

In the rapidly evolving field of Alzheimer's Disease diagnosis, numerous artificial intelligence (AI) techniques including machine learning and deep learning methodologies have emerged, each offering distinct strengths and facing particular limitations. The use of AI in

Abbreviations: AD, Alzheimer's Disease; DL, Deep Learning; GAN, Generative Adversarial Network; SMRI, Structural Magnetic Resonance Imaging; XAI, Explainable Artificial Intelligence; MCI, Mild Cognitive Impairment; CN, Cognitively Normal; CBR, Case Based Reasoning; LEAR, Learn-Explain-Reinforce; KNN, K-Nearest Neighbors; CNN, Convolutional Neural Network; MI, Mutual Information; MDS, Multidimensional Scaling; PMCI, Progressive MCI; SMCI, Stable MCI; AUC, Area Under Curve.

* Corresponding author.
E-mail addresses: adarsh.valoor@soton.ac.uk (A. Valoor), ganga@nitt.edu (G.R. Gangadharan).

https://doi.org/10.1016/j.jneumeth.2024.110318
Received 16 July 2024; Received in revised form 3 October 2024; Accepted 3 November 2024
Available online 9 November 2024

medicine often involves complex algorithms that are opaque or 'black boxes,' making it difficult for users to understand how decisions are made. Explainable Artificial Intelligence (XAI) demystifies these processes by making AI decisions more transparent and understandable (Ali et al., 2023). This transparency is critical in healthcare, where decisions can significantly impact patient outcomes. By providing insights into how AI reaches its conclusions, XAI helps build confidence among clinicians and patients in the technology's recommendations, which is essential for its adoption and effective utilization. XAI contributes significantly to clinical decision-making by providing clinicians with comprehensible explanations of AI-generated recommendations.

The use of explainable AI techniques has increased the transparency and comprehensibility of machine learning and deep learning models, which is crucial for building trust and encouraging the adoption of therapies for Alzheimer's disease. Essemlali et al. (2020) used a modified BrainNet convolutional neural network on diffusion-weighted Magnetic Resonance Imaging (MRI) to examine the structural connectomics of AD and Mild Cognitive Impairment (MCI). El-Sappagh et al. (2021) developed an interpretable two-layer design with a Random Forest classifier optimized using key biological and clinical markers for AD diagnosis and progression detection, enabling both binary categorization and detailed layer-by-layer data analysis. Yu et al. (2022) proposed an explainable framework combining attention mechanisms and multi-scale features for the visual classification of AD. Lombardi et al. (2022) developed an explainable Random Forest based classification for diagnosing MCI and AD categories. Shojaei et al. (2023) employed an explainable 3D convolutional neural network with a genetic algorithm model for AD diagnosis. Adarsh et al. (2024) presented explainable convolutional neural networks with discriminative dictionary learning for precise classification of individuals into AD, MCI, and Cognitively Normal (CN) categories while identifying subtle phases within the MCI spectrum. These said studies highlight the critical role of XAI in making complex models more interpretable, essential for clinical acceptance and the effective application of AI in AD diagnosis and treatment. However, the said studies are not based on the concept of counterfactual approach to explainable AI, providing insights into alternative outcomes.

The counterfactual approach to explainable AI aims to enhance understanding of AI systems by offering counterfactual explanations (Ser et al., 2024; Baron, 2023; Wachter et al., 2017). Counterfactual reasoning and map generation represent a cutting-edge approach in the domain of medical imaging, particularly in enhancing the interpretability and diagnostic accuracy of deep learning models for diseases such as Alzheimer's disease. This methodology is based on the "what if" scenarios, offering insights into alternative outcomes (Chou et al., 2022). By applying this concept, we can generate counterfactual maps that illuminate how changes in specific brain regions might influence the model's diagnosis, providing a deeper understanding of the disease's pathology and the model's decision-making process. However, the current limitation of counterfactual explanation techniques is their inclination to consider only a limited number of hypothetical scenarios for counterfactual reasoning (Bass et al., 2023; Sauer and Geiger, 2021; Looveren and Klaise, 2021). In case of disease progression of AD where numerous diagnostic categories such as CN, MCI, and AD are of interest, the process becomes complex. Under these conditions, it is necessary to construct different and independent models for various combinations of clinical labels.

In this context, this paper develops a case-based comprehensive counterfactual methodology for various stages of Alzheimer's disease, ensuring high interpretability and accountability in the diagnostic process. The role of counterfactual map estimation using the concept of mutual information (Song et al., 2012) in our paper is to select and assign weights to the features used in counterfactual reasoning. This approach focuses on local similarities in the case space rather than global similarities, ensuring that the explanations are relevant to the specific query case. The case-based counterfactual reasoning approach, inspired by case based reasoning (CBR) (Bichindaritz and Montani, 2011), makes our proposed methodology novel by integrating case-based reasoning with counterfactual reasoning to validate the model and generate robust explanations. This dual approach allows for a deeper understanding of the decision-making process of the model and ensures high interpretability and accountability in diagnosing Alzheimer's Disease. By paralleling current cases with past instances, the proposed methodology identifies essential similarities and differences, thus reinforcing the classification process and making the decisions more robust and reliable. The salient contributions of this paper are as follows:

1. Develops a novel counterfactual methodology using U-Net with Generative Adversarial Network (GAN) to create comprehensive, interpretable diagnostic maps for Alzheimer's disease
2. Incorporates case-based counterfactual reasoning for robust decision classification for counterfactual maps to understand how changes in specific features affect the predictions of the model.

The remainder of this paper is organized as follows. Section 2 provides a concise review of the related works on counterfactual analysis. Section 3 presents our methodology for developing comprehensive interpretable diagnostic maps for Alzheimer's disease using counterfactual analysis and case-based reasoning. Section 4 illustrates the results analysis of the proposed methodology followed by concluding remarks in Section 5.

2. Literature review

Recently, counterfactual reasoning has gained interest as a method for providing visual explanations in current research. One stage of providing a counterfactual explanation involves assessing the model's output performance in hypothetical scenarios. Zhai et al. (2017) presented a visual feature attribution employing Wasserstein GANs (VAGAN), which uses the modified version of the Generative Adversarial Network to create a counterfactual representation. This representation can change a given input instance into a classification corresponding to a different label. However, the VAGAN framework requires knowledge about the true label assigned to the input sample, a condition that might not always be feasible.

Pawlowski et al. (2020) suggested a modular framework which integrates structural causal models and deep learning components to learn about the functional causal relationships among relevant variables. They applied the technique of amortized inference to make counterfactual inference feasible for high-dimensional problems. They then put this methodology into practice on magnetic resonance scans of healthy brain tissue. However, the framework prevents unobserved confounding and entails all variables to be observed during a counterfactual training. Looveren and Klaise (2021) developed a model-agnostic approach for generating interpretable counterfactual explanations of classifier outcome predictions by utilizing class prototypes. As the approach proposes using class prototypes to guide perturbations towards more interpretable counterfactuals, this may add significant complexity to the optimization process.

Bass et al. (2023) devised the ICAM methodology, which aims to provide bidirectional counterfactual explanations. ICAM is a well-regarded model for generating counterfactual explanations and feature attributions in medical imaging. The bidirectional counterfactual explanations provided by ICAM allow to explain classification outcomes for two distinct classes. However, the ICAM approach has a limitation when dealing with tasks involving more than two distinct categories of interest, and its explanation capability is limited to a bidirectional manner. Unlike ICAM, our proposed model does not limit itself to binary classes but could produce explanations across all stages of cognitive decline, including more subtle intermediate stages.

Oh et al. (2023) proposed a learn-explain-reinforce (LEAR) framework, a first-of-its-kind unified model for generating multiway

counterfactual maps, which integrates diagnostic model learning, visual explanation generation, and the reinforcement of the trained diagnostic model based on the visual explanations to diagnose Alzheimer's disease. The generic framework of LEAR may not always focus on the most relevant or localized brain features needed for precise AD diagnosis. While the LEAR framework emphasizes a single model that combines diagnostic learning, visual explanation, and model reinforcement, our approach incorporates mutual information for local feature selection and case-based reasoning to enhance decision classification robustness.

Mulyadi et al. (2023) proposes a novel deep-learning approach through eXplainable AD Likelihood Map Estimation (XADLiME) for AD progression modeling. Using a hypothetical AD progression manifold portrayed with a set of topological-aware prototypes in an unsupervised manner, this method eliminates the necessity to predetermine the number of class-specific prototypes. However, the proposed model relies on complex computations, including 3D brain image processing, variational autoencoders, and self-organizing maps.

Dhinagar et al. (2024) proposes a novel methodology integrating conditional latent diffusion models and denoising diffusion probabilistic models to provide insights into Alzheimer's disease effects on the brain's anatomy at the individual level. The methodology uses implicit classifier-free guidance to alter the conditioning of an encoded individual scan to its counterfactual while preserving subject-specific image details. Although this provides significant advantage for interpretability of AD effects, the process of creating these counterfactuals with preserving subject-specific details remains complex and might require further refinement to make it more intuitive for clinical applications. GCAN (Shen et al., 2024) proposes a generative counterfactual attention-directed network for identifying predictable functional connectivity features and related brain regions. These regions are then used as counterfactual attention maps onto functional connectivity matrices to increase the prediction performance. An Atlas-aware Bidirectional Transformer is developed to reconstruct functional connectivity within the GCAN framework. As GCAN involves multi-step processes, this may add significant complexity to the real world implementation and clinical interpretability.

Cao and Yamins (2024) explored concerns regarding the intelligibility of neural network models and their relevance to brain function. They argued that understanding behavior requires considering top-down factors, such as evolutionary constraints. Although optimization techniques capture essential aspects, they may oversimplify the complexity of biological systems, reducing their explanatory power. Huang et al. (2024) hypothesized that eCFT-induced memory modification is associated with changes in neural activity, particularly within the default mode network, and that this relationship is influenced by anxiety levels. Their study, conducted with 39 participants exhibiting varying levels of trait anxiety, provided evidence supporting this link. However, the small sample size limits the generalizability of the results, highlighting the need for further research with larger sample sizes. Tang et al. (2024) proposed a QSM-based arising-from-chair assessment using a causal graph-neural-network framework, integrating counterfactual and debiasing strategies to capture causal features. The counterfactual strategy suppresses irrelevant features, while debiasing addresses sampling bias. However, the approach's effectiveness may be limited by its reliance on specific data sampling conditions. Goodwin et al. (2024) developed an open-source Simple Behavioral Analysis (SimBA) platform to improve the model explainability in behavioral neuroscience. SimBA enhances transparency and comparability of behavioral classifiers. However, its effectiveness may be limited by the complexity of behaviors and variability between species and datasets.

To the best of our knowledge, most of the existing works on counterfactual explanations contemplate either one or two hypothetical scenarios (Bass et al., 2020; Sauer and Geiger, 2021; Looveren and Klaise, 2021). Current methodologies such as VAGAN (Zhai et al., 2017) and ICAM (Bass et al., 2023) concentrate on generating images that are classified as a specific target category, but they fall short in delivering thorough explanations or justifications for the decisions made by the classifier. The novelty of this work lies in developing a case-based counterfactual methodology to enhance the explainability of deep learning models in Alzheimer's disease diagnosis, offering clearer insights into the decision-making process of AI systems for neuroscience.

3. Proposed methodology

Given the structural MRI (sMRI) images, we classify the preprocessed sMRI images using the K-nearest neighbors algorithm, into three categories: cognitively normal, mild cognitive impairment, and Alzheimer's Disease. The extracted features, which offer a strong foundation for differentiating between various phases of cognitive decline, are the basis for this classification. We apply a U-Net architecture (Ronneberger et al., 2015) integrated with a Generative Adversarial Network (Goodfellow et al., 2020) to generate the counterfactual map. These maps visually represent how hypothetical changes in specific brain regions could influence the diagnosis. We apply mutual information to select and assign weights to locally significant features. This local approach addresses a critical gap in existing methodologies that typically consider only a limited number of hypothetical scenarios. To further enhance the reliability and robustness of the classification model, we incorporate case-based reasoning which involves comparing the current case with past similar cases to identify essential similarities and differences. Then, we present the counterfactual maps and classification results using a visual interface that aids in interpreting the counterfactual scenarios and supports clinical decision-making. The proposed methodology is presented in Fig. 1.

3.1. Preprocessing

We start by pre-processing the sMRI images to remove artifacts and noise. This vital step allowed us to remove artifacts and noise that could mask the underlying neurological patterns, thereby impeding the discovery of significant features. The first step in our preprocessing pipeline is bias field correction, which addresses intensity non-uniformities commonly found in MRI scans due to scanner imperfections. These inconsistencies can lead to misleading intensity variations, potentially misguiding the model. We used the N4ITK algorithm (Tustison et al., 2010), a widely recognized method for bias field correction, to ensure that the intensity values across the brain tissue were uniform, thus improving the reliability of feature extraction for critical regions like the hippocampus and the cortex.

Then, we applied spatial normalization, aligning each MRI scan to a standardized template. This step ensures that the anatomical structures are consistently positioned across all subjects, allowing the model to focus on actual morphological differences rather than variations in image orientation or size. We utilized affine registration, followed by non-linear transformations to account for both global and local anatomical variations. This approach is particularly important in AD diagnosis, where subtle structural differences in brain regions are key to differentiating between CN, MCI, and AD cases.

We also implemented denoising techniques to remove high-frequency noise without sacrificing the important structural details. A combination of Gaussian smoothing and non-local means filtering was used to preserve edges while reducing noise, crucial for maintaining the integrity of small but significant brain regions. These preprocessing steps collectively ensure that the input data is clean, consistent, and ready for effective classification, thereby improving the reproducibility and robustness of our methodology.

After pre-processing, we isolated the following key features; Gray matter volume, cortical thickness, and white matter integrity. These features are universally acknowledged as vital indicators of Alzheimer's disease progression.

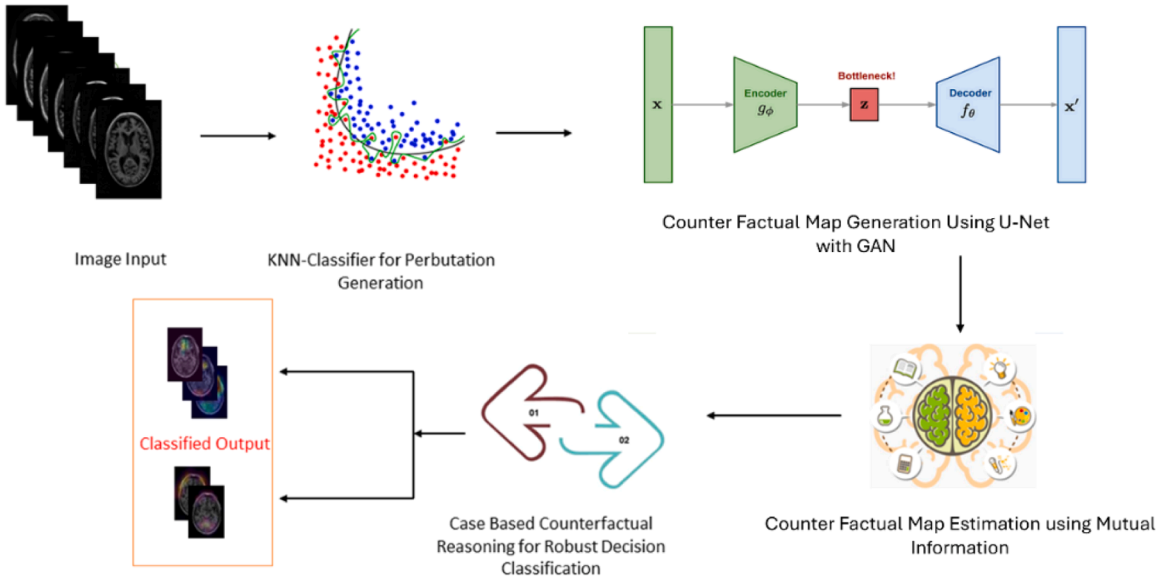

Fig. 1. Proposed methodology.

3.2. Feature extraction and classification

Our strategy involved utilizing the K-nearest neighbors (KNN) algorithm with the aim of building a dependable and efficient classification model. The methodology involves using the K-nearest neighbors algorithm for classifying sMRI images into categories: CN, MCI, and AD. The model focuses on key features like gray matter volume and cortical thickness integral to AD progression.

The choice of using K-Nearest Neighbors (KNN) for classification in our methodology, despite its simplicity compared to more sophisticated classifiers like Random Forest or Support Vector Machines (SVM), is based on several crucial factors especially in like Alzheimer's Disease (AD) diagnosis. KNN is a highly interpretable model, which is a key requirement in healthcare applications, where the ability to explain decisions to clinicians and patients is paramount. Unlike Random Forest or SVM, which often act as black-box models, KNN provides transparent decision-making by classifying a new MRI scan based on its similarity to already classified scans, known as its "nearest neighbors." This simple yet effective mechanism aligns with the goals of Explainable AI (XAI) in medical imaging, where clinicians need to trust and understand the decision-making process behind a diagnosis. Another reason for selecting KNN is its non-parametric nature, meaning it makes no assumptions about the underlying data distribution, which is particularly useful for complex, multi-modal datasets like those used in AD diagnosis, where data points from different classes (e.g., Cognitively Normal, Mild Cognitive Impairment, and Alzheimer's Disease) may overlap significantly. Random Forest, while powerful, often requires more tuning and offers less interpretability, and SVM, while strong in handling high-dimensional data, is more complex and harder to explain in clinical settings, especially when using non-linear kernels. Moreover, KNN is well-suited for small to medium-sized datasets, which are common in medical research due to data privacy constraints and the difficulty of acquiring large-scale medical imaging datasets. Its simplicity also means that KNN can work efficiently with the ADNI dataset without the need for complex hyperparameter tuning, which is often required in models like SVM.

In our methodology, the integration of KNN with Mutual Information for feature selection further enhances its performance. Mutual Information helps identify the most relevant features, reducing the dimensionality of the data and allowing KNN to focus on the critical features that influence AD diagnosis, such as subtle changes in brain regions like the hippocampus. This localized feature selection is especially important in AD diagnosis, where small anatomical changes can significantly impact disease classification. Additionally, KNN is robust to class imbalances, which is common in AD datasets where certain classes, such as CN or MCI, might be overrepresented. KNN addresses this by considering the local majority class within a given neighborhood, thus avoiding the bias toward larger classes. While models like Random Forest or SVM offer more complex decision boundaries, they also increase the difficulty of interpretation and may not always provide the local, instance-specific explanations that KNN can offer. Overall, the simplicity, interpretability, non-parametric nature, and efficiency in handling high-dimensional and imbalanced data, especially when enhanced by Mutual Information for local feature selection makes KNN is an ideal choice for AD diagnosis in this context.

3.3. Counterfactual map generation using U-Net with GAN

We develop a U-Net architecture integrated with a Generative Adversarial Network, capturing complex patterns in the sMRI images and generates counterfactual maps. The U-Net with GAN architecture consists of a fully convolutional neural network (CNN) having a contracting path (encoder) to capture context and a symmetric expanding path (decoder) that enables precise localization. The model also incorporates skip connections between the encoder blocks and the decoder blocks to help the network learn fine-grained details more effectively (see Table 1). These maps visually represent how hypothetical changes in specific brain regions could influence the diagnosis.

This deep feature extraction by the encoder is crucial as it forms the foundational base from which nuanced modifications in the counterfactual maps are drawn. The decoder, on the other hand, reconstructs these features into full image formats but with alterations guided by target conditions defined by the counterfactual scenarios, such as hypothetical changes in specific brain regions, to simulate different stages of Alzheimer's disease.

Integrated within a Generative Adversarial Network (GAN), the U-Net serves as the generator, creating counterfactual images that are assessed by a discriminator model for realism and accuracy against actual sMRI data. This adversarial training ensures the U-Net learns to produce highly realistic and medically plausible images, which is critical for exploring how slight changes in brain morphology could lead to different diagnostic outcomes. The model's ability to focus on local features while retaining contextual information through skip connections enables precise modifications without losing the overall brain

Table 1
Various components of the U-Net with GAN architecture.

Components	Description
Input Layer	The model that accepts input images of shape (256, 256, 1)
Encoder (Contracting Path)	The encoder consists of repeated applications of two 3×3 convolutions (each followed by a rectified linear unit (ReLU) and batch normalization) and a 2×2 max pooling operation with stride 2 for down-sampling. At each down-sampling step, the number of feature channels is doubled:
Encoder Block 1:	Two 3×3 convolutions with 32 filters, ReLU activation, and batch normalization. Followed by a 2×2 max pooling operation. (Dropout of 0.3)
Encoder Block 2:	Two 3×3 convolutions with 64 filters, ReLU activation, and batch normalization. Followed by a 2×2 max pooling operation. (Dropout of 0.3)
Encoder Block 3:	Two 3×3 convolutions with 128 filters, ReLU activation, and batch normalization. Followed by a 2×2 max pooling operation. (Dropout of 0.3)
Encoder Block 4:	Two 3×3 convolutions with 256 filters, ReLU activation, and batch normalization. Followed by a 2×2 max pooling operation. (Dropout of 0.3)
Bottleneck	This part of the network bridges the encoder and decoder modules. It consists of two 3×3 convolutions with 512 filters, ReLU activation, and batch normalization.
Decoder (Expansive Path)	The decoder or expansive path includes a series of up-sampling and concatenation followed by two 3×3 convolutions (each followed by ReLU and batch normalization). The concatenations from the corresponding encoder block are added to provide high-resolution features for the up-sampling operations:
Decoder Block 1:	Up-sampling of the feature map followed by a 2×2 convolution ("up-convolution") that halves the number of feature channels. A concatenation with the corresponding cropped feature map from the encoder. Two 3×3 convolutions with 256 filters, ReLU activation, and batch normalization. (Dropout of 0.3)
Decoder Block 2:	Similar structure with 128 filters.
Decoder Block 3:	Similar structure with 64 filters.
Decoder Block 4:	Similar structure with 32 filters.
Output Layer	A 1×1 convolution that maps each 32-component feature vector to the desired number of classes. In this case, it is a single class output with a sigmoid activation function.

structure context.

Algorithm 1. Counterfactual Map Generation using U-Net with GAN

Input: sMRI images (pre-processed), target condition for counterfactual scenario

Output: Counterfactual map

1. Initialize the U-Net model with GAN architecture.

2. Apply U-Net encoder to extract deep features from input sMRI images.

3. Use GAN's discriminator to evaluate the generated images for realism.

4. For each input image:

Generate the corresponding counterfactual map by modifying target brain regions as per the hypothetical condition

End for

5. Reconstruct the image with the decoder based on modified features.

6. Return the counterfactual map showing changes in brain morphology for the counterfactual condition.

Once trained, the model generates counterfactual maps for input sMRI, highlighting necessary changes for shifting a diagnosis from one category to another (See Algorithm 1). This functionality is for understanding disease progression and potential intervention points. The produced counterfactual maps serve as a form of visual explanation to support clinical decision-making, helping clinicians understand the model's reasoning and gain insights into which features or regions are considered most indicative of Alzheimer's or its progression.

3.4. Mutual information for local feature selection

Our study uses Mutual Information (MI) (Song et al., 2012) for selecting and assigning weights to locally significant features. We applied MI locally, at the query case neighborhood, as opposed to

globally. The idea of local similarity in the literature originally meant being "local in feature space," indicating a similarity confined to a single attribute. Recent studies have shifted their focus toward explaining "local similarity in case space", a similarity restricted to a specific subset of the case base. Our study examined the community close to the subject of inquiry, found via a universal similarity metric, and identified regional similarities using mutual information. While these similarities are relevant near the specific query case, they cannot be generalized globally. By focusing on local similarities rather than global ones, we ensure that the counterfactual maps are highly relevant to the specific query case. This local approach addresses a critical gap in existing methodologies that typically consider only a limited number of hypothetical scenarios.

Algorithm 2. Mutual Information for Local Feature Selection

Input: Set of features from sMRI images, query case, neighbouring cases

Output: Weights for local features

1. Calculate mutual information (MI) for each feature between the query case and neighbouring cases.

2. For each feature:

Compute local similarity in the feature space.

Calculate MI values between query cases and neighbourhood cases.

End for

3. Assign weights based on the MI value, prioritizing higher weights for features with more significant local similarity.

4. Return feature weights for counterfactual map generation.

Mutual information quantifies the dependency between two variables—in this case, specific brain features and the target classification (Cognitively Normal, Mild Cognitive Impairment, or Alzheimer's Disease). In our approach, mutual information is applied locally, meaning it focuses on selecting features that are most informative for a particular query case, rather than across the entire dataset. The weights were assigned based on the mutual information score of each feature, where higher scores indicated a stronger relationship between the feature and the diagnosis. These weights were used to prioritize features such as hippocampal volume, cortical thickness, and white matter integrity, which are critical in understanding the progression of Alzheimer's Disease. The mutual information scores were computed for each feature in the training set, and features with the highest scores were selected for classification using K-Nearest Neighbors (KNN). The pseudo code for this process is given in Algorithm 2.

In terms of alternative methods, we explored other feature selection techniques such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). However, PCA reduces dimensionality by transforming features into a new set of orthogonal components, which can obscure the interpretability of individual features—something critical in medical imaging. Similarly, RFE, while effective in identifying the most influential features, is computationally expensive and less intuitive for local feature selection, where the focus is on the immediate relevance of specific brain regions to a particular diagnosis. Mutual information was chosen because it maintains both interpretability and efficiency, allowing us to focus on features that provide the most diagnostic value without eliminating clarity in how those features impact the classification.

3.5. Case-based reasoning

To strengthen the reliability of our classification model, we integrated two creative and complementary methods: the production of permuted values and the application of a case-based reasoning (CBR) approach. By producing permuted values around the predicted regions, our objective was to examine the stability of the predictions and evaluate the discrepancies between these values and the ground truth. A counterfactual generator facilitated this process, assisting us in analyzing the sensitivity of the model to changes in input features and

providing critical insights into its overall performance.

Algorithm 3. Case-Based Reasoning

Input: Query case q, historical cases in the database X

Output: Explanation based on past cases

1. For query case q, identify the most similar (n-1) cases in the database.

2. Calculate the dissimilarity score for each case using Euclidean distance for numerical dimensions and categorical matching for nominal dimensions.

3. Select the closest similar cases and analyze similarities/differences.

4. Use identified past cases to generate counterfactual explanations.

5. Visualize and interpret the counterfactual scenarios to support decision-making.

Case-Based Reasoning (Bichindaritz and Montani, 2011) represents a distinctive approach within the field of artificial intelligence and cognitive science, emphasizing the role of past experiences in solving new problems. Unlike traditional problem-solving methodologies that rely on generic rules or algorithms applicable across various scenarios (Kolodner, 1992; Guerrero et al., 2022), CBR is grounded in the principle that similar problems have similar solutions. This methodological approach mirrors human reasoning more closely, as it reflects our natural tendency to tackle new challenges by drawing upon our memories of how we addressed similar situations in the past (Aamodt and Plaza, 1994; Leake et al., 2021). The CBR approach served to reinforce the classification model by leveraging the extensive knowledge embedded in previous similar cases (Barman et al., 2020). By paralleling the current case with these past instances, we identified essential similarities and differences (see Algorithm 3). This information helps to the classification process, reinforcing its reliability and robustness.

3.6. Visualization and interpretation

We developed the counterfactual maps and classification results using a visual interface. This includes scatter plots and rainbow boxes that help clinicians understand the reasoning behind the model and gain insights into which brain regions are most indicative of Alzheimer's Disease progression. We designed a visual interface to determine the category of a particular query case, q. The problem involves a database, X, with (p+1) dimensions and query case q. Each dimension's value can be real numbers, integers, Booleans, or a set of nominals. The solution set, Y, consists of a finite number of classes, {y1, y2, …}, with the cardinality of Y satisfying the condition.

Our visual interface comprises two distinct elements. A scatter plot is on the left, and a set of rainbow boxes is on the right. The scatter plot represents the distance matrix among the cases in X in two arbitrary dimensions. The center white point denotes the value of q. Dots shaded by category depict instances with similar attributes. Forms express the category for those with color vision deficiencies. Proximity between data points in the scatter plot indicates case similarity. The Multidimensional Scaling (MDS) technique eliminates the relevance of axes. A target with its center aligned to q is included in the scatter plot's background to improve distance comprehension. The scatter plot helps determine the number of similar instances in each category and discern the category with the highest number of comparable instances. The clear segregation of Chromatic data points in the scatter plot indicates the effectiveness of the CBR method in classifying instances. Intermingled colored points may suggest classification difficulties associated to the attributes of the database or the design or CBR. In such cases, interpreting CBR methodology outcomes requires caution. Rainbow-colored boxes on the interface's right side showcase each class's qualitative characteristics. These boxes help determine the dominant color and, subsequently, the class to which the variable q belongs.

During the retrieval phase of the CBR phase, the (n-1) cases most similar to the query case q are selected from the database. This is often accomplished using a dissimilarity measure to quantify the dissimilarity or distance between two given cases. A common measure is the Euclidean distance, which calculates the geometric distance between two points in the feature space. This distance metric works for numerical dimensions, with continuous values represented as points within the feature space. For nominal dimensions, which are categorical values without numerical interpretation, a simple method assigns a dissimilarity score of 0 to identical values and a score of 1 to different values. The methodology assumes that every nominal dimension has an equivalent level of importance in calculating dissimilarity and considers

Table 2
Parameter settings of the proposed model.

U-Net Parameters	
Input size	(256, 256, 1)
Encoder	4 blocks with 32, 64, 128, 256 filters respectively
Decoder	4 blocks with 256, 128, 64, 32 filters, respectively
Activation	ReLU for intermediate layers, Sigmoid for output
Pooling	2×2 max pooling
Dropout	0.3 for each encoder block
GAN Parameters	
Generator	U-Net as the generator
Discriminator	CNN with ReLU activation
Adversarial loss function	Binary cross-entropy
Optimizer	Adam with a learning rate of 0.0002

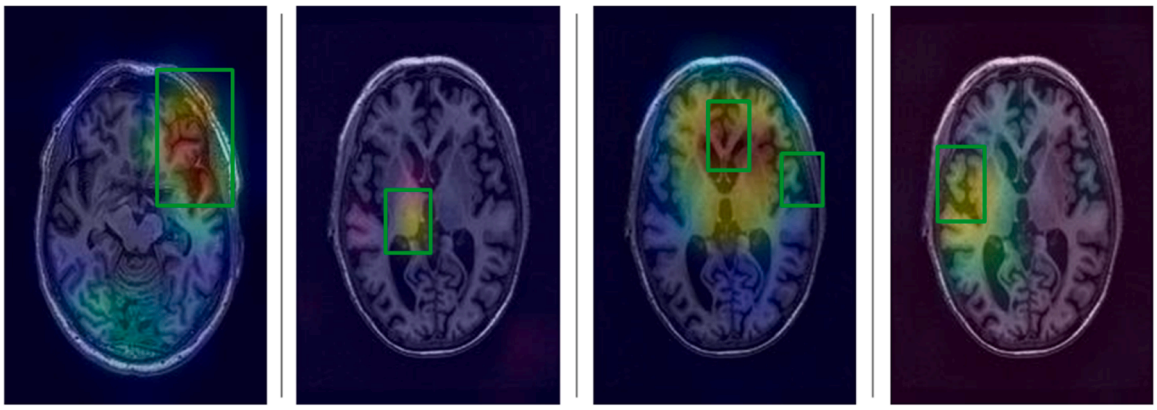

Fig. 2. Counter factual map generated highlighting the affected regions of the brain [The affected regions are highlighted and marked by a bounding box] 1: AD, 2: CN, 3 & 4: MCI.

Table 3
Inferences on the identified regions of the brain.

Region	CN Volume (mm^3)	MCI Volume (mm^3)	AD Volume (mm^3)	Volume Loss	Inference	Counterfactual Intervention
Hippocampus	4000	3800	3500	High	Significant shrinkage correlates with memory loss.	Simulate increased volume to match CN state and assess the impact on diagnosis.
Entorhinal Cortex	3500	3200	2800	High	Atrophy associated with early-stage AD.	Visualize volume restoration to a pre-atrophy state for early detection.
Ventricles	25,000	28,000	35,000	Moderate	Enlargement correlates with disease progression.	Model ventricular reduction to explore potential reversal of progression.
Cerebral Cortex	450,000	440,000	430,000	Low	General atrophy is observed in later stages.	Assess the impact of generalized cortical volume preservation on disease prognosis.
Amygdala	1800	1700	1500	Moderate	Reduction indicates limbic system involvement.	Evaluate the effects of limbic system volume maintenance on emotional and behavioral symptoms.

any variation in nominal value as equally divergent. The choice of a dissimilarity measure depends on the specific problem domain and the data attributes used in the CBR procedure. Selecting a suitable dissimilarity measure that matches the problem requirements and the data semantics is essential to ensure accurate and meaningful retrieval of similar features.

4. Experimentation and results

4.1. Dataset description

The study employs the ADNI (Alzheimer's Disease Neuroimaging Initiative, 2004) dataset, which is compiled by the Alzheimer's Disease Neuroimaging Initiative. The ADNI dataset is widely recognized as a complex and demanding resource due to its capacity to mirror authentic medical scenarios and encompasses a wide range of images featuring nuanced and heterogeneous morphological alterations. The dataset comprises 3D structural magnetic resonance imaging (sMRI) data obtained from diverse subject groups, including individuals with CN, MCI, and AD. The subjects with Mild Cognitive Impairment (MCI) are classified into two sub-groups, namely progressive MCI (pMCI) and stable MCI (sMCI). The former pertains to those MCI subjects who have progressed to AD within 36 months of screening, while the latter pertains to those who have remained in the MCI group within the same period. The experiments utilized a sample of 466 CN subjects, 427 sMCI subjects, 262 pMCI subjects, and 348 AD subjects drawn from the ADNI. The study exclusively utilized the baseline MRIs of participants despite the possibility of multiple MRIs being obtained throughout their lifetime. This approach yielded a dataset of 1540 images that were utilized in the experiments.

4.2. Experimental setup

The experimental methodology employs a five-fold cross-validation framework, with consistent indices utilized across all comparison techniques to prevent any potential data leakage during the training process. This evaluation was carried out through the incorporation of longitudinal samples. These samples consisted of a total of 1503 subjects who were selected from the ADNI. In an attempt to trace changes in the morphological features of sMRI, we undertook the creation of ground-truth maps. These changes were a response to alterations in clinical diagnosis. The main purpose behind generating these maps was to provide a clear and precise representation of any modifications that occurred.

To ensure that our results were not biased, we took strict measures against any potential data leakage. It was essential to prevent the training models from being exposed to data that they were not supposed to see yet. As a part of this prevention strategy, none of the images mentioned earlier were employed in any of the model training procedures, by ensuring the integrity of the training of models and, thus, the reliability of our findings.

This proposed model is implemented using Python language, leveraging Google Colab for development and execution. The primary libraries used include numpy for numerical operations, matplotlib for visualization, scikit-learn and Tensorflow for preprocessing and evaluation metrics. The parameter settings of the proposed model (including U-Net and GAN models) are described in Table 2. For the U-Net, the architecture consists of a contracting path with multiple convolutional layers, followed by an expansive path with upsampling layers. We used a dropout rate of 0.3 to prevent overfitting and batch normalization to ensure stable training. For the GAN model, we used an Adam optimizer with a learning rate of 0.0002 and a discriminator loss function based on binary cross-entropy. The reasoning behind these parameter choices is

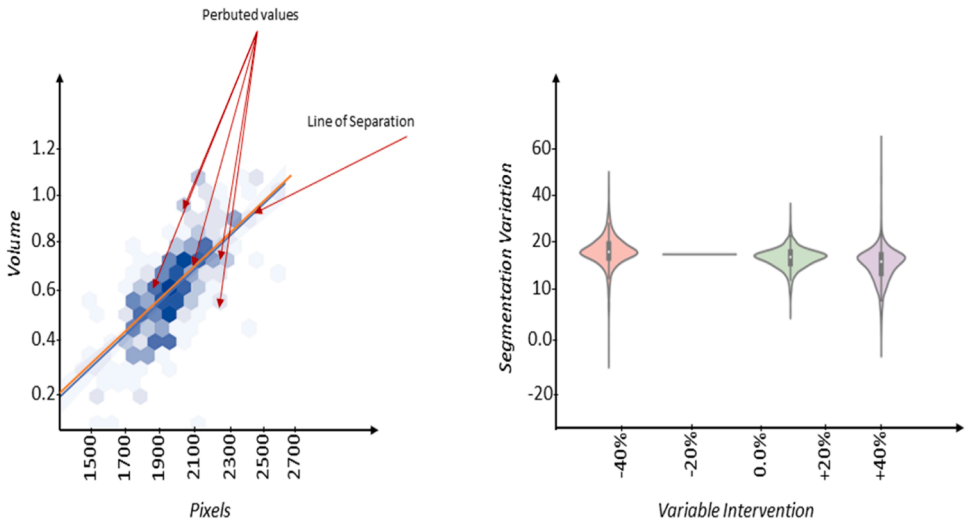

Fig. 3. Scatter plot and violin plots showing the perbuted values associated with the counterfactual map generation and the segmentation variation found out to make a classification.

based on their effectiveness in handling high-dimensional MRI data and generating realistic counterfactual maps.

4.3. Results analysis

The proposed methodology generates the counterfactual maps (See Fig. 2) which provide clinicians with an intuitive visual understanding of the model's decisions. Fig. 2 demonstrates the counterfactual maps that visually delineate changes in specific brain regions, which could potentially alter a patient's diagnostic category. These maps highlight areas such as the hippocampus and cerebral cortex, where modifications could shift a diagnosis from CN to MCI or from MCI to AD based on the progression of anatomical changes. For example, an increase in hippocampal volume represented in one counterfactual map might hypothetically move the diagnosis of a patient closer to a CN classification from MCI, suggesting a potential area of therapeutic intervention. Clinicians can use these insights to understand the critical anatomical features influencing Alzheimer's progression, thus aiding in both diagnosis and in strategizing treatment plans that target specific brain changes.

Table 3 provides a detailed comparative analysis of brain volumes across CN, MCI, and AD based on counterfactual map generation and analysis. It highlights regions such as the Hippocampus, Entorhinal Cortex, Ventricles, Cerebral Cortex, and Amygdala, with respective volume ranges indicating progressive changes associated with each condition. The significance of volume loss is rated from high to low, suggesting the degree to which volume changes can be indicative of disease progression. Inferences drawn from these volumetric changes align with clinical expectations: significant hippocampal shrinkage is linked to memory loss, entorhinal atrophy is an early AD indicator, ventricular enlargement denotes disease progression, cortical atrophy is seen in later stages, and amygdala reduction is connected to limbic system involvement. Furthermore, counterfactual interventions propose hypothetical adjustments to these volumes, aiming to assess the potential impact on diagnosis and understand the model's sensitivity to structural brain changes. This comprehensive approach not only enhances the interpretability of diagnostic models but also elucidates the intricate relationship between anatomical changes and cognitive decline in AD.

Unlike conventional gradient-based methods, which often provide sparse or regionally confined insights, the proposed methodology produced comprehensive visual explanations that spanned the entirety of the relevant regions of the brain. This detailed mapping enabled us to confirm the diagnostic decisions of the model against clinical ground truths, ensuring that our findings were both scientifically robust and practically relevant. Unlike alternative gradient-based methods such as LRP-Z (Bach et al., 2015) and DeepLIFT (Shrikumar et al., 2017), the proposed counterfactual maps showed clearer class-discriminative regions and did not overlook any relevant regions. While methods like DeepTaylor (Montavon et al., 2017) showed hints of counterfactual reasoning throughout the image, they were flawed by extraneous attributions at the brain's periphery or morphological boundaries. Even though Grad-CAM (Selvaraju et al., 2020) provides class-discriminative visualizations, it usually captures coarse regions. The proposed methodology has shown superior results over GAN based techniques including ICAM-Reg (Bass et al., 2023) and VAGAN (Zhai et al., 2017). The proposed method effectively identified all subtle regions showing increased cortical thickness, along with accurate detection of reductions

Table 4
Comparative studies with existing state of the art methods.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	AIC
Our Methodology	95	93	96	0.97	1823.21
LRP-Z (Bach et al., 2015)	88	85	90	0.89	2045.37
DeepLIFT (Shrikumar et al., 2017)	90	88	91	0.91	1982.45
LEAR (Oh et al., 2023)	92	89	94	0.96	1908.62
Deep Taylor (Montavon et al., 2017)	86	83	88	0.87	2102.84
Grad-CAM (Selvaraju et al., 2020)	84	81	86	0.85	2131.67
VAGAN (Zhai et al., 2017)	92	90	93	0.95	1887.73
ICAM-Reg (Bass et al., 2023)	92	90	93	0.95	1889.92
GCAM (Shen et al., 2024)	91	89	92	0.94	1967.23
LDM-DDPM (Dhinagar et al., 2024)	92	90	93	0.94	1905.11
XADLiME (Mulyadi et al., 2023)	91	90	90	0.89	1974.72

in ventricular volume and hypertrophy in the hippocampus.

These findings are consistent with previous clinical neuroscience research, particularly studies focusing on the morphological markers of Alzheimer's disease. For instance, the research by (DeTure and Dickson, 2019) highlighted the critical role of ventricular enlargement and cortical atrophy in diagnosing Alzheimer's disease, aligning closely with our observations of these features using our proposed methodology. Further, the progressive patterns of hippocampal atrophy correlated well with the clinical progression descriptions detailed by (Miao et al., 2022). The ability of the proposed model to accurately reflect these anatomical changes illustrates its potential as a reliable diagnostic tool, bridging cutting-edge AI capabilities with clinically verified markers of Alzheimer's disease progression.

In the context of Alzheimer's disease, Fig. 3 features a combination of scatter and violin plots, presenting the output variations of the proposed counterfactual map generation when subjected to different input perturbations. This visualization serves a dual purpose: it not only illustrates the variability in diagnostic categorization under slightly altered input conditions but also highlights the stability and reliability of the proposed predictions. The scatter plot component, by depicting the spread and clustering of diagnostic outcomes across CN, MCI, and AD, allows clinicians to visually assess the sensitivity of the model to minor changes in brain morphology. For instance, a tightly clustered group of points around the 'AD' label suggests that the proposed methodology consistently identifies AD-related features across varied inputs. This consistency is pivotal when clinicians are weighing the robustness of diagnostic tools, as it underscores the model's capacity to maintain diagnostic integrity despite normal variations in imaging data. The violin plot complements this by offering a density distribution of the predictions, thereby providing deeper insights into the probability density of each diagnostic category.

4.4. Comparative analysis

We have used the following evaluation parameters: Accuracy measures the proportion of correct predictions (both true positives and true negatives) out of all predictions made, calculated as follows. It gives an overall performance assessment of the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Table 5
Ablation studies.

Model Configuration	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Baseline U-Net (used in the proposed method)	95	93	96	0.97
Without Skip Connections	89	86	91	0.90
Reduced Number of Filters	87	84	89	0.88
Increased Number of Filters	95	92	96	0.97
Shallower U-Net (Fewer Layers)	85	82	87	0.86
Deeper U-Net (More Layers)	95	92	96	0.97
Without Dropout Layers	88	85	90	0.89
With Batch Normalization	95	93	96	0.97
Without Batch Normalization	88	85	90	0.89
Alternative Activation Function (ReLU to Leaky ReLU)	94	92	95	0.96
Alternative Activation Function (ReLU to ELU)	95	93	96	0.97
Different Up-sampling Method (Transpose Convolution to Bilinear Up-sampling)	92	89	94	0.93

where *TP (True Positive)* indicates the number of instances where the model correctly predicted the positive class, *TN (True Negative)* indicates the number of instances where the model correctly predicted the negative class, *FP (False Positive)* indicates the number of instances where the model incorrectly predicted the positive class and FN (False Negative) indicates the number of instances where the model incorrectly predicted the negative class.

Sensitivity measures the proportion of actual positives that are correctly identified by the model, calculated as follows. It indicates how well the model can detect positive instances.

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity measures the proportion of actual negatives that are correctly identified by the model, calculated as follows. It shows how well the model can detect negative instances.

$$Specificity = \frac{TN}{TN + FP}$$

We have compared the proposed method with the existing state-of-the-art methods, including LRP-Z (Bach et al., 2015), DeepLIFT (Shrikumar et al., 2017), LEAR (Oh et al., 2023), Deep Taylor (Montavon et al., 2017), Grad-CAM (Selvaraju et al., 2020), VAGAN (Zhai et al., 2017), ICAM-Reg (Bass et al., 2023), GCAM (Shen et al., 2024), LDM-DDPM (Dhinagar et al., 2024), and XADLiME (Mulyadi et al., 2023). The comparative analysis highlights the performance of various deep learning models in identifying and classifying features relevant to Alzheimer's disease and CN cases (see Table 4). The proposed model exhibits superior performance across all metrics, achieving an accuracy of 95 %, sensitivity of 93 %, specificity of 96 %, and an Area Under Curve (AUC) of 0.97. This indicates its exceptional capability in precisely identifying and outlining critical features, making it highly effective in distinguishing between different classes. The performance metrics we reported, including accuracy, sensitivity, specificity, and AUC, were calculated using 5-fold cross-validation to ensure the robustness of the model.

The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data (Stoica and Selen, 2004). We have calculated the Akaike Information Criterion (AIC) for each of the models compared. The AIC accounts for both the goodness of fit and the complexity of the model, providing a more balanced comparison. Since models like Grad-CAM, DeepLIFT, and ICAM-Reg are non-parametric or involve complex neural networks, directly calculating the AIC using the traditional maximum likelihood method is not straightforward. So, we estimated AIC indirectly by using surrogate methods like comparing performance on validation dataset.

The model outputs predicted probabilities for each subject, reflecting the likelihood of being classified in one of these diagnostic categories. At higher probability thresholds, the model conservatively predicts a positive diagnosis (e.g., MCI or AD), resulting in a lower true positive rate (TPR) but also a lower false positive rate (FPR), as fewer misclassifications occur. As the threshold is reduced, the model begins to classify more subjects as positive, increasing the TPR by identifying more true cases of AD or MCI, but this also raises the FPR, indicating that more CN subjects are incorrectly diagnosed as having cognitive impairments. The Area Under the Curve (AUC) is 0.97, reflecting the exceptional ability of the proposed model to distinguish between the different stages of cognitive impairment. An AUC of 0.97 means that the model has a 97 % chance of correctly ranking a randomly chosen individual with AD or MCI higher than a randomly chosen CN individual. This high AUC demonstrates the robustness of the model in capturing subtle morphological changes in brain regions such as the hippocampus and entorhinal cortex, which are critical for early diagnosis of Alzheimer's disease. The AUC, combined with the ROC curve, illustrates the reliability of the proposed model in clinical decision-making, supporting its effectiveness in distinguishing between various stages of Alzheimer's

disease progression.

While the performance of our proposed methodology achieves an accuracy of 95 % and an AUC of 0.97, the proposed methodology fails particularly in distinguishing among MCI, CN and AD cases. Most of the errors arise in distinguishing between MCI and CN or MCI and AD, which is a common challenge in AD diagnosis due to the subtle and overlapping nature of brain morphology changes between these stages.

Further, the decision to use only baseline MRIs was primarily driven by data availability and the desire to maintain consistency across subjects. Baseline MRIs capture the initial snapshot of brain structure at the point of diagnosis, making them suitable for cross-sectional analysis. However, we acknowledge that incorporating longitudinal data where MRI scans are taken over time could enhance the robustness of the proposed model by providing insights into the temporal progression of brain changes.

4.5. Ablation studies

Ablation study is important in the field of artificial intelligence, which helps to investigate the performance of the method by removing certain components of the model and analyzing the importance of that component for the overall system. In our work, we focus on variations within the U-Net architecture to understand their impact on the performance of a deep learning model for Alzheimer's disease diagnosis using structural MRI data. Key components of the U-Net architecture, such as skip connections, the number of filters, network depth, dropout layers, and batch normalization, play crucial roles in determining the effectiveness of the proposed model (see Table 5). The results of the ablation study indicate that our proposed methodology performs better than the ablation models.

4.6. Clinical implications

Counterfactual reasoning allows clinicians to explore "what if" scenarios, showing how slight changes in patient data (e.g., brain morphology in MRI scans) might alter a diagnosis. While this enhances the interpretability of machine learning models and builds trust in AI-assisted diagnosis, their use in medical settings raises important considerations regarding patient safety and the reliability of interventions based on these models. While counterfactuals can aid in understanding disease dynamics, they are not at all replacements for traditional clinical assessments or expert judgment. There is a need for robust validation of these models in real clinical settings to ensure that the generated explanations align with established medical knowledge and do not introduce uncertainty or overconfidence in AI-driven decisions. Proper safeguards, such as integrating human oversight and ensuring data quality, are essential to mitigate these risks and fully realize the potential of counterfactual models in healthcare.

5. Conclusion

The proposed methodology in this paper combines a U-Net with GAN architecture and case-based counterfactual reasoning to enhance Alzheimer's disease diagnosis from structural MRI images. By generating comprehensive counterfactual maps, the model enables visual explanations that enhance the interpretability and robustness of deep learning models used in clinical settings. The results show that the methodology significantly outperforms existing models in terms of accuracy, sensitivity, and specificity, achieving an accuracy of 95 % and an AUC of 0.97. These findings provide clinicians with detailed insights into critical brain regions, such as the hippocampus and cerebral cortex, that influence diagnostic outcomes, fostering more informed decision-making. This approach not only aids in diagnosis but also supports precision medicine by allowing clinicians to explore "what-if" scenarios regarding disease progression and potential interventions.

Future work could explore the integration of longitudinal MRI data to improve the ability of the model to track and predict transitions between different stages of cognitive impairment over time. Incorporating additional imaging modalities, such as diffusion tensor imaging or positron emission tomography, could further enhance the diagnostic accuracy by providing more comprehensive information on brain structure and function. Another area for improvement is the development of more sophisticated counterfactual generation techniques that consider a wider range of hypothetical scenarios, potentially extending beyond binary classifications to better capture the complexities of AD progression. Finally, validating this model in real-world clinical settings and exploring its application across diverse populations would ensure its robustness and generalizability, further supporting its clinical adoption.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

CRediT authorship contribution statement

Adarsh Valoor: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **G.R. Gangadharan:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of Competing Interest

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

Acknowledgements

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this paper. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/howtoapply/ADNIAcknowledgementList.pdf. Adarsh Valoor would also like to thank RAI UK for funding the post-doctoral research at the University of Southmapton.

Data availability

The data for this study is publicly available on the Internet.

References

Aamodt, A., Plaza, E., 1994. Case-based reasoning: foundational issues, methodological variations, and system approaches. Artif. Intell. Commun. 7 (1), 39–52.

Adarsh, V., Gangadharan, G.R., Fiore, U., Zanetti, P., 2024. Multimodal classification of Alzheimer's disease and mild cognitive impairment using custom MKSCDDL kernel over CNN with transparent decision-making for explainable diagnosis. Nat. Sci. Rep. 14 (1), 1774.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F., 2023. Explainable Artificial Intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf. Fusion 99, 101805.

Alzheimer's Disease Neuroimaging Initiative, 2004. http://adni.loni.usc.edu/.
Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K.-R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10 (7). Art. no. e0130140.
Barman, R., Biswas, S.K., Sarkar, S., Purkayastha, B., Soni, B., 2020. Image processing using case-based reasoning: a survey. In: Lecture Notes in Electrical Engineering, 686. Springer, pp. 653–661.
Baron, S., 2023. Explainable AI and causal understanding: counterfactual approaches considered. Minds Mach. 33, 347–377.
Bass, C., da Silva, M., Sudre, C., Williams, L.Z.J., Sousa, H.S., Tudosiu, P.-D., Alfaro-Almagro, F., Fitzgibbon, S.P., Glasser, M.F., Smith, S.M., Robinson, E.C., 2023. ICAM-Reg: interpretable classification and regression with feature attribution for mapping neurological phenotypes in individual scans. IEEE Trans. Med. Imaging 42 (4), 959–970.
Bichindaritz, I., Montani, S., 2011. Advances in case-based reasoning in the health sciences. Artif. Intell. Med. 51 (2), 75–79.
Cao, R., Yamins, D., 2024. Explanatory models in neuroscience, part 2: functional intelligibility and the contravariance principle. Cogn. Syst. Res. 85, 101200.
Chou, Y.-L., Moreira, C., Bruza, P., Ouyang, C., Jorge, J., 2022. Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. Inf. Fusion 81, 59–83.
DeTure, M.A., Dickson, D.W., 2019. The neuropathological diagnosis of Alzheimer's disease. Mol. Neurodegener. 14, 32.
Dhinagar, N.J., Thomopoulos, S.I., Laltoo, E., Thompson, P.M., 2024. Counterfactual MRI Generation with Denoising Diffusion Models for Interpretable Alzheimer's Disease Effect Detection. https://www.biorxiv.org/content/10.1101/2024.02.05.578983v1.full.pdf.
El-Sappagh, S., Alonso, J.M., Islam, S.M.R.R., Sultan, A.M., Kwak, K.S., 2021. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Sci. Rep. 11 (1), 2660.
Essemlali, A., St-Onge, E., Descoteaux, M., Jodoin, P.-M., 2020. Understanding Alzheimer disease's structural connectivity through explainable AI. Proc. Mach. Learn. Res. 121, 217–229.
Gao, Y., Tan, M.-S., Wang, H.-F., Zhang, W., Wang, Z.-X., Jiang, T., Yu, J.-T., Tan, L., 2016. ZCWPW1 is associated with late-onset Alzheimer's disease in Han Chinese: a replication study and meta-analyses. Oncotarget 7 (15), 20305.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63 (11), 139–144.
Goodwin, N.L., Choong, J.J., Hwang, S., Pitts, K., Bloom, L., Islam, A., Zhang, Y.Y., Szelenyi, E.R., Tong, X., Newman, E.L., Miczek, K., 2024. Simple behavioral analysis (SimBA) as a platform for explainable machine learning in behavioral neuroscience. Nat. Neurosci. 1–14.
Guerrero, J.I., Miró-Amarante, G., Martín, A., 2022. Decision support system in health care building design based on case-based reasoning and reinforcement learning. Expert Syst. Appl. 187, 116037.
Huang, S., Faul, L., Parikh, N., LaBar, K.S., De Brigard, F., 2024. Counterfactual thinking induces different neural patterns of memory modification in anxious individuals. Sci. Rep. 14 (1), 10630.
Kolodner, J.L., 1992. An introduction to case-based reasoning. Artif. Intell. Rev. 6 (1), 3–34.
Leake, D., Ye, X., Crandall, D., 2021. Supporting case-based reasoning with neural networks: an illustration for case adaptation. In: Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021), AAAI, pp. 22-24.
Lindner, M.D., McArthur, R.A., Deadwyler, S.A., Hampson, R.E., Tariot, P.N., 2008. Development, optimization and use of preclinical behavioral models to maximize the productivity of drug discovery for Alzheimer's disease. Animal and Translational Models for CNS Drug Discovery. Academic Press, pp. 93–157.
Lombardi, A., Diacono, D., Amoroso, N., Biecek, P., Monaco, A., Bellantuono, L., Pantaleo, E., Logroscino, G., De Blasi, R., Tangaro, S., Bellotti, R., 2022. A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of mild cognitive impairment and Alzheimer's disease. Brain Inf. 9 (1), 17.
Looveren, A.V., Klaise, J., 2021. Interpretable counterfactual explanations guided by prototypes. In: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 650–665.
Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Muller, K.-R., 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit. 65, 211–222.
Mulyadi, A.W., Jung, W., Oh, K., Yoon, J.S., Lee, K.H., Suk, H., 2023. Estimating explainable Alzheimer's disease likelihood map via clinically-guided prototype learning. NeuroImage 273, 120073.
Oh, K., Yoon, J.S., Suk, H.-I., 2023. Learn-explain-reinforce: counterfactual reasoning and its guidance to reinforce an Alzheimer's disease diagnosis model. IEEE Trans. Pattern Anal. Mach. Intell. 45 (4), 4843–4857.
Pawlowski, N., Castro, D.C., Glocker, B., 2020. Deep structural causal models for tractable counterfactual inference. In: Proceedings of the Conference on Neural Information Processing Systems.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI.
Sauer, A., Geiger, A., 2021. Counterfactual generative networks. In: Proceedings of the International Conference on Learning Representations.
Selkoe, D.J., Hardy, J., 2016. The amyloid hypothesis of Alzheimer's disease at 25 years. EMBO Mol. Med. 8 (6), 595–608.
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. 128 (2), 336–359.
Ser, J.D., Barredo-Arrieta, A., Díaz-Rodríguez, N., Herrera, F., Saranti, A., Holzinger, A., 2024. On generating trustworthy counterfactual explanations. Inf. Sci. 655, 119898.
Shen, X., Song, Z., Zhang, Z., 2024. GCAN: Generative Counterfactual Attention-guided Network for Explainable Cognitive Decline Diagnostics based on fMRI Functional Connectivity. https://arxiv.org/pdf/2403.01758.
Shojaei, S., Saniee Abadeh, M., Momeni, Z., 2023. An evolutionary explainable deep learning approach for Alzheimer's MRI classification. Expert Syst. Appl. 220, 119709.
Shrikumar, A., Greenside, P., Kundaje, A., 2017. Learning important features through propagating activation differences. In: Proceedings of the International Conference on Machine Learning, pp. 3145–3153.
Song, L., Langfelder, P., Horvath, S., 2012. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinform. 13 (1), 328.
Stoica, P., Selen, Y., 2004. Model-order selection: a review of information criterion rules. IEEE Signal Process. Mag. 21 (4), 36–47.
Tang, X., Guo, R., Zhang, C., Qian, X., 2024. A causal counterfactual graph neural network for arising-from-chair abnormality detection in parkinsonians. Med. Image Anal. 97, 103266.
Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320.
Wachter, S., Mittelstadt, B., Russell, C., 2017. Counterfactual explanations without opening the black box: automated decisions and the GDPR. Cybersecurity 31, 841.
Yu, L., Xiang, W., Fang, J., Phoebe Chen, Y.P., Zhu, R., 2022. A novel explainable neural network for Alzheimer's disease diagnosis. Pattern Recognit. 131.
Zhai, S., Cheng, Y., Feris, R., Zhang, Z., 2017. Generative adversarial networks as variational training of energy based models. In: Proceedings of the International Conference on Learning Representations.