Engineering Applications of Artificial Intelligence 172 (2026) 114291

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Research paper

Trustworthy Alzheimer's diagnosis: Integrating robustness, fairness, and explainability in neuroimaging based deep ensemble framework

Maria Bashir [a], Nasir Rahim [b], Shaker El-Sappagh [c,d,e], Omar Amin El-Serafy [f], Tamer Abuhmed [c,*]

[a] Department of Electrical & Computer Engineering, College of Information and Communication Engineering, Sungkyunkwan University, Suwon, 16419, South Korea
[b] School of Computing, Gachon University, Seongnam, 13120, South Korea
[c] College of Computing and Informatics, Sungkyunkwan University, Suwon, 16419, South Korea
[d] Faculty of Computer Science and Engineering, Galala University, Suez, 435611, Egypt
[e] Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Benha, 13518, Egypt
[f] Neurology Department, Faculty of Medicine, Cairo University, 11562, Egypt

ARTICLE INFO

Keywords:
Alzheimer's Disease
Trustworthy Artificial Intelligence
Explainability
Adversarial Robustness
Data Fairness
Medical Imaging

ABSTRACT

Mild Cognitive Impairment represents a clinically significant transitional condition in the progression toward Alzheimer's Disease, underscoring the need for diagnostic tools that are both accurate and clinically reliable. This study presents a deep learning based diagnostic framework designed to support trustworthy medical decision making by explicitly addressing three pillars of trustworthy artificial intelligence, including adversarial robustness, fairness across gender groups and visual explainability. The proposed approach analyzes gray matter, white matter, and cerebrospinal fluid through dedicated tissue segmentation, followed by an ensemble of deep learning classifiers optimized using Bayesian optimization. The framework was evaluated on the Alzheimer's Disease Neuroimaging Initiative cohort and demonstrated strong and consistent diagnostic performance, outperforming representative baseline methods in distinguishing progressive and stable forms of Mild Cognitive Impairment. The generalization capability was further assessed using an independent external dataset. To ensure clinical relevance, visual explainability of disease related brain regions and model robustness were jointly evaluated by a medical domain expert, supporting a clinician in the loop workflow and demonstrating the practical utility of the system in real clinical settings. The results indicate that the proposed framework provides an interpretable, robust, and fair diagnostic solution that aligns with emerging expectations for trustworthiness in medical imaging and offers meaningful support for clinical decision making.

1. Introduction

Alzheimer's Disease (AD) is among the most prevalent neurodegenerative disorders worldwide and represents an escalating clinical and socioeconomic burden for patients, caregivers, and healthcare systems (Erdogmus and Kabakus, 2023). Mild Cognitive Impairment (MCI) lies along the transitional continuum between healthy aging and AD, where measurable cognitive decline is present while daily functional abilities are largely retained. Importantly, individuals diagnosed with progressive MCI (pMCI) exhibit a substantially higher risk of conversion to AD compared to those with stable MCI (sMCI). Early identification of this high-risk subgroup is therefore critical for enabling timely intervention and personalized disease management. However, distinguishing pMCI from sMCI remains inherently challenging, as neuroanatomical changes at this stage are often subtle, heterogeneous, and partially overlapping across individuals. Magnetic Resonance Imaging (MRI) plays a central role in this setting due to its capacity to capture fine-grained structural alterations associated with neurodegeneration (Valizadeh et al., 2025).

In recent years, Deep Learning (DL) has emerged as a dominant paradigm for MRI-based classification of MCI and AD (Valizadeh et al., 2025). A large body of existing work focuses on manually defined regions of interest, most notably the hippocampus, motivated by its strong association with memory impairment (Poloni et al., 2022). While region-based strategies can effectively capture localized atrophy, they may fail to account for distributed structural changes that extend beyond a single anatomical site. In contrast, whole-slice or whole-brain approaches offer broader spatial coverage but frequently sacrifice clinical interpretability and may entangle heterogeneous tissue-specific

* Corresponding author.
E-mail addresses: mariabashir@g.skku.edu (M. Bashir), nrahim3797@ieee.org (N. Rahim), shaker@skku.edu (S. El-Sappagh), omarserafy@kasralainy.edu.eg (O.A. El-Serafy), tamer@skku.edu (T. Abuhmed).

https://doi.org/10.1016/j.engappai.2026.114291
Received 4 October 2025; Received in revised form 16 February 2026; Accepted 20 February 2026
Available online 28 February 2026

patterns in ways that obscure disease-relevant mechanisms (Al-bakri et al., 2025).

Accumulating neuroimaging evidence suggests that disease progression in MCI is not confined to isolated brain regions but is instead reflected through coordinated alterations across Gray Matter (GM), White Matter (WM), and Cerebrospinal Fluid (CSF). These tissue-specific changes encode complementary and physiologically distinct information that can support improved discrimination between sMCI and pMCI (Shi et al., 2019). Despite this, many DL-based frameworks process all tissue information jointly within a single model, which can limit transparency, blur anatomically meaningful signatures, and reduce confidence in model-driven clinical decisions (Valizadeh et al., 2025). Beyond architectural considerations, increasing attention has been directed toward the trustworthiness of Artificial Intelligence (AI) systems deployed in healthcare settings. Although DL models often demonstrate strong predictive accuracy, they may remain sensitive to minor input perturbations, provide limited justification for their predictions, or exhibit uneven performance across demographic subgroups. In the context of neurodegenerative disease prognosis, such shortcomings carry serious clinical implications, as diagnostic errors may influence long-term treatment planning and patient counseling. As a result, diagnostic frameworks intended for real-world clinical use must extend beyond accuracy alone and explicitly address interpretability, robustness, and fairness as foundational design principles.

Motivated by these challenges, we propose a comprehensive diagnostic framework that integrates tissue-specific modeling with the core dimensions of trustworthy AI. The framework begins by segmenting MRI scans into GM, WM, and CSF, allowing each tissue type to be independently learned using Bayesian-optimized Convolutional Neural Networks (CNNs). The resulting representations are then combined through an ensemble fusion strategy to capture complementary structural information. Rather than introducing a single novel algorithmic component, MRI-DeepTrust advances the state of the art through a unified system-level design that systematically integrates performance, transparency, robustness, fairness and explainability within a single clinically motivated pipeline. The proposed framework adopts a 2D tissue-based modeling strategy, which offers practical advantages in terms of computational efficiency, data availability, and training stability, particularly in multi-site neuroimaging cohorts. While this design choice does not explicitly model inter-slice spatial continuity, it enables scalable learning across heterogeneous datasets and facilitates tissue-aware interpretability. Extensions toward 3D volumetric or multi-view architectures are therefore regarded as a natural direction for future work rather than a prerequisite for establishing trustworthy diagnostic behavior at the current stage.

In this study, we propose MRI-DeepTrust, a unified and clinically grounded DL framework designed for reliable classification of pMCI vs. sMCI using MRI. The framework integrates tissue-specific CNNs for GM, WM, and CSF with Bayesian-optimized ensemble fusion to capture complementary neurodegenerative patterns. Unlike conventional MRI based DL approaches that typically focus on predictive accuracy alone or address interpretability and robustness separately, the proposed architecture is explicitly designed to incorporate multiple pillars of trustworthy AI including adversarial robustness, gender based fairness and visual explainability within a single diagnostic pipeline. To enhance clinical relevance, the development and evaluation of MRI-DeepTrust were conducted with continuous involvement of medical domain experts. The visual explainability of disease related brain regions, as well as the stability of model outputs under adversarial perturbations, were qualitatively reviewed by a clinician throughout the study. This clinician in the loop workflow enables expert guided validation of salient brain regions and supports the translational feasibility of the framework in real clinical settings. The explainability component leverages Grad-CAM to generate voxel level activation maps across tissue-specific models, facilitating time consistent and anatomically meaningful interpretation of predictions. Extensive experiments conducted on the ADNI cohort demonstrate that MRI-DeepTrust achieves improved predictive performance while maintaining stable behavior under robustness and fairness evaluations. In addition to standard performance metrics, the framework systematically evaluates adversarial resilience using PGD-based perturbation testing and examines gender based fairness through established statistical measures. Although individual components such as Grad-CAM visualization, PGD-based robustness analysis, and fairness metrics are well established in the literature, their coordinated integration within a tissue-aware ensemble architecture is novel for the challenging task of differentiating pMCI vs. sMCI classes. To the best of our knowledge, no prior study combines GM, WM, and CSF specific CNNs optimized via Bayesian search with unified assessments of adversarial robustness, demographic fairness, and clinically guided visual interpretability in a single end to end diagnostic framework. To further examine generalizability, external validation was performed using the NACC cohort. The model trained on ADNI was directly evaluated on NACC without additional fine tuning, alongside a 3D Vision Transformer (ViT) baseline, to assess cross cohort robustness. The results indicate that the proposed framework maintains competitive performance under distributional shifts, reinforcing its translational potential.

Despite substantial progress in MRI-based DL for AD, existing approaches remain fragmented. Many studies prioritize predictive accuracy or focus on a single dimension such as regional modeling, explainability, or robustness in isolation. However, the absence of a unified framework that simultaneously integrates tissue-specific modeling, ensemble fusion, adversarial robustness, fairness evaluation, and clinically guided explainability limits the translational reliability of these systems. This gap is particularly critical for distinguishing pMCI and sMCI, where subtle anatomical variations require both high discriminative power and trustworthy decision behavior. The proposed MRI-DeepTrust framework directly addresses this limitation by embedding the core pillars of trustworthy AI into a cohesive tissue-aware ensemble architecture, thereby advancing beyond performance-centric methods toward clinically dependable neuroimaging diagnostics.

An overview of the proposed pipeline is illustrated in Fig. 1, and the main contributions of this study are summarized as follows.

- We propose a tissue-aware trustworthy MRI diagnostic framework for classification of progressive and stable MCI, in which GM, WM, and CSF are modeled independently using tissue-specific CNNs.
- We design and evaluate homogeneous and heterogeneous deep ensemble learning framework that fuse tissue-specific representations to enhance robustness and generalization, systematically analyzing their complementary contributions to diagnostic performance.
- We conduct a tissue contribution and ablation analysis to explicitly quantify the individual and combined roles of GM, WM, and CSF in the disease progression prediction, improving anatomical transparency and interpretability.
- We assess the adversarial robustness of the proposed framework under PGD-based attack, evaluating its stability against malicious perturbations in clinically realistic threat scenarios.
- We further strengthen model resilience by incorporating PGD-based adversarial training, demonstrating improved robustness without sacrificing diagnostic precision.
- We perform a comprehensive gender-based fairness evaluation using demographic parity, equal opportunity, and group sufficiency metrics, ensuring equitable performance across male and female subgroups.
- We provide tissue-level explainability analysis using visual attribution techniques to highlight disease-related brain regions across the GM, WM, and CSF, supporting transparent and clinically interpretable decision making.

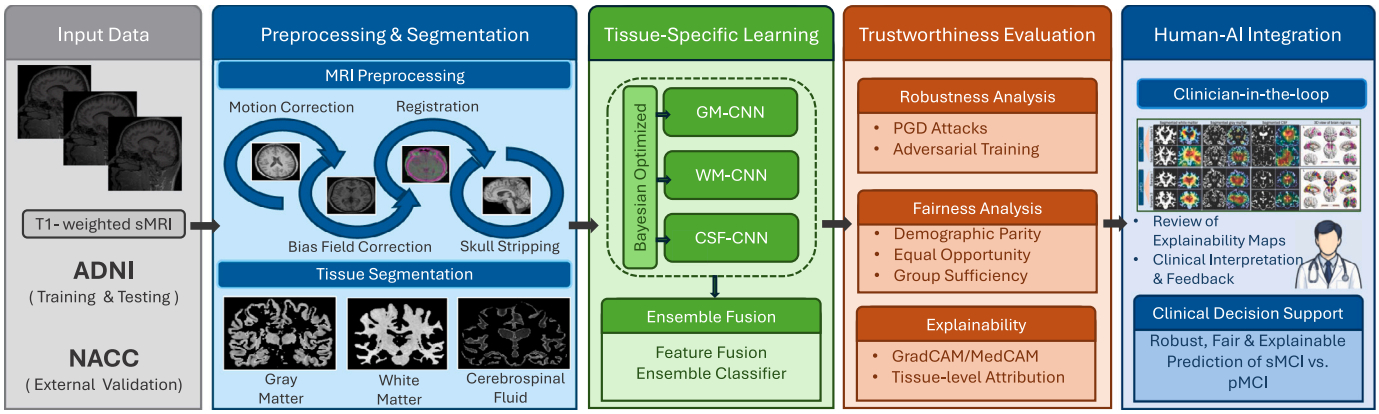

Fig. 1. Block diagram of the proposed MRI-DeepTrust framework for trustworthy classification of sMCI vs. pMCI. The architecture consists of MRI preprocessing and tissue segmentation, followed by Bayesian-optimized tissue-specific CNN modules for gray matter, white matter, and cerebrospinal fluid regions. Learned representations are fused through an ensemble classifier and evaluated under robustness, fairness, and explainability components. The framework further incorporates adversarial analysis, gender-based fairness assessment, tissue-level attribution, and clinician-in-the-loop validation, with external validation on the NACC cohort to demonstrate generalizability and clinical reliability.

- We demonstrate the generalizability of the proposed framework through an external validation using the NACC cohort, showing robust cross-cohort performance and strong translational reliability beyond the training distribution.
- Collectively, these components form a unified trustworthy diagnostic pipeline that integrates tissue-specific CNN optimization, ensemble fusion, adversarial robustness evaluation, gender-based fairness assessment and visual explainability analysis for classification of sMCI vs. pMCI within a single system-level framework.
- The proposed framework is evaluated in close collaboration with medical domain expert, where visual explainability outputs and model stability are qualitatively reviewed within a clinician-in-the-loop workflow to ensure consistency with clinical reasoning and real-world diagnostic practice.

The remainder of this manuscript is organized as follows. Section 2 reviews the body of literature relevant to this study. Section 3 introduces the fundamental concepts related to CNNs, model explainability, adversarial robustness, and fairness considerations. The architecture and design of the proposed framework are detailed in Section 4. Section 5 describes the experimental setup. The experimental results and their discussion are presented in Section 6. Section 7 outlines the limitations of the current work and highlights potential directions for future research. Finally, Section 8 concludes the paper.

2. Related work

The application of MRI for AD diagnosis has advanced considerably in recent years, largely driven by the adoption of Machine Learning (ML) and DL-based techniques capable of detecting subtle neuroanatomical changes associated with disease progression. A substantial body of work has focused on segmenting the brain into major anatomical components, including GM, WM, and CSF, to capture region-specific patterns of atrophy. More recent studies further subdivide these tissues into finer subregions to improve diagnostic sensitivity. In this section, we review prior work on MRI-based diagnosis of MCI, with particular emphasis on whole-brain modeling strategies and segmentation-driven approaches aimed at improving diagnostic accuracy and model reliability. We further discuss recent advances in trustworthy AI for medical imaging, highlighting approaches that address robustness, fairness, and explainability.

2.1. Diagnosis of MCI patients using whole brain MRI

Diagnosing MCI and predicting its progression remain challenging due to the reliance of traditional clinical assessments on subjective cognitive evaluations and long-term monitoring (Porsteinsson et al., 2021). Whole-brain MRI offers an objective alternative by capturing global structural and functional alterations associated with neurodegeneration. Several DL approaches have leveraged whole-brain MRI to improve diagnostic performance. (Orouskhani et al., 2022) introduced a Siamese network with a modified contrastive loss to address data scarcity in Cognitively Normal (CN)–AD classification, though model interpretability was limited. (Rahim et al., 2023b) proposed a 3D CNN–RNN framework using longitudinal MRI to model spatiotemporal disease progression, but clinical robustness was not fully established. Recent work has focused on improving model interpretability. (Leonardsen et al., 2024) developed an explainable CNN-based framework with LRP maps highlighting regions such as the amygdala and parahippocampal gyrus. Similarly, (Khan et al., 2024) presented a Fractional-Order CNN with attention mechanisms and LIME-based explanations on the ADNI dataset. Multimodal extensions further enhance prediction and generalization. (Rahim et al., 2024) combined longitudinal MRI and cognitive assessments using an ensemble framework validated on ADNI and NACC, while (Zhang et al., 2025) integrated MRI, FDG-PET, and clinical variables to generate interpretable probability maps linked to disease-relevant brain regions.

2.2. Diagnosis of MCI patients using segmented brain regions in MRI

While MRI is widely used to study structural brain changes associated with cognitive decline, its high dimensionality can limit the efficiency of DL models. To address this, many studies have adopted ROI-based strategies that focus on regions strongly linked to AD, including the hippocampus, EC, and medial temporal areas, thereby reducing computational complexity while retaining clinically meaningful information (Aderghal et al., 2018; Choi et al., 2020). Early ROI-based work concentrated on hippocampal segmentation. (Aderghal et al., 2018) reported 66% accuracy using atlas-based hippocampal analysis, which improved to 80% after incorporating mean diffusivity measures, MNI-space alignment, and AAL segmentation. Similarly, (Choi et al., 2020) applied CNNs to hippocampal ROIs extracted using 3D Slicer, achieving 78.1% accuracy, while EC-based texture features with logistic regression yielded 71% accuracy for classification of MCI from CN. Later extensions introduced age-adjusted, patch-based CNNs with FreeSurfer-derived descriptors, achieving 79.9% accuracy and an AUC of 86.1%.

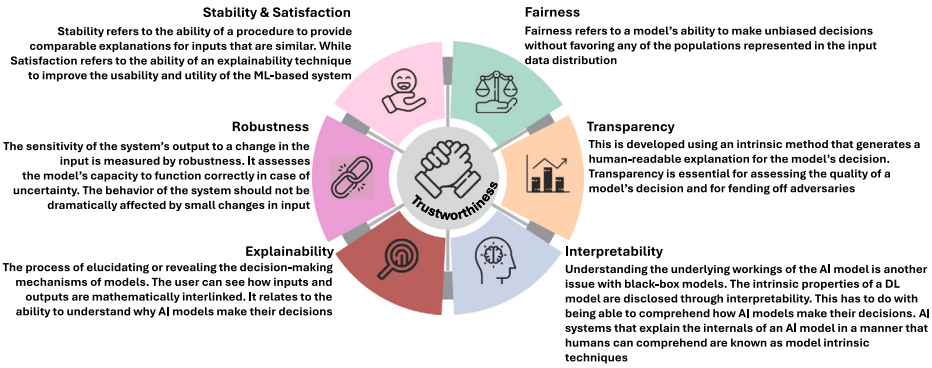

Fig. 2. This representation illustrates the relationship among core Trustworthy AI concepts, highlighting how robustness, fairness, and explainability are interconnected. Together, these dimensions guide method selection, analysis depth, and the development of reliable and ethically grounded systems.

More recently, segmentation based pipelines have expanded beyond accuracy driven objectives. (Tong et al., 2025) proposed FACIMS, which extracted volumetric features from 80 cortical and subcortical ROIs within ADNI and incorporated fairness-aware constraints to reduce demographic bias related to age, sex, and race, while maintaining competitive performance. Collectively, these studies reflect the evolution of ROI-based MCI diagnosis from localized hippocampal analysis toward more robust, interpretable, and equitable frameworks.

2.3. Trustworthy deep models for MCI

Trustworthy AI refers to a set of requirements that span the full lifecycle of predictive models, from data preparation to deployment and monitoring (Wang et al., 2021; El-Sappagh et al., 2023; Albahri et al., 2023). Often framed as a chain of trust (Choung et al., 2023), this concept emphasizes accountability, fairness, robustness, transparency, reproducibility, safety, and privacy. In medical AI, these principles are particularly critical, as models must behave reliably under uncertainty, mitigate bias, protect sensitive data, and provide interpretable outputs that clinicians can understand and act upon (Javed et al., 2024). Despite growing awareness, translating these principles into routine clinical practice remains challenging, largely due to the persistent gap between technical solutions and real-world healthcare constraints.

Recent work on MCI and AD diagnosis illustrates how trustworthiness can be embedded into DL pipelines. (Rahim et al., 2023a) proposed a CNN–BiLSTM framework using longitudinal MRI and cognitive scores from ADNI, combining temporal rank pooling with guided Grad-CAM to visualize disease progression in regions such as the hippocampus and temporal gyrus. Similarly, (Mahmud et al., 2024) introduced an explainable transfer learning approach using pretrained CNNs and EfficientNet variants with saliency maps and Grad-CAM on the OASIS dataset. Extending this direction, (Rahim et al., 2025) developed a multi-view ensemble analyzing axial, coronal, and sagittal MRI slices, where Grad-CAM based explanations enhanced interpretability. Together, these studies reflect a broader shift from purely performance-centric models toward trustworthy frameworks that integrate explainability and robustness, supporting more reliable and clinically acceptable MCI diagnosis.

2.3.1. Trustworthy AI for medical imaging

While earlier studies have examined individual aspects of trustworthiness such as robustness, fairness or explainability across different imaging modalities, the literature remains fragmented, with no single work jointly addressing all three pillars. We therefore review prior efforts that relate to these dimensions in medical imaging to highlight existing progress and the gap that motivates our integrated approach. (Lee, 2025) demonstrated that chest X-ray classifiers can be gradually corrupted through poisoned training data, and used SHAP visualizations to show how model attention shifts toward hidden triggers during the attack process. Their work highlights the broader need for safety aware and interpretable AI systems in clinical imaging, reinforcing the motivation for robustness and explainability within trustworthy MRI-based diagnostics. (Tayebi Arasteh et al., 2024) evaluated differential privacy on large-scale chest X-ray and abdominal CT datasets and showed that privacy constraints can reduce accuracy but generally do not introduce substantial demographic bias across age or sex groups. These findings highlight the growing interest in understanding how trustworthy AI principles, particularly fairness and robustness, behave in real clinical imaging settings and provide useful context for our own trustworthiness evaluations. (Abbas et al., 2025) introduced a transfer learning model for skin disease detection using visual explainability to reveal which image regions guided the network's decisions, helping reduce the opacity of the DL predictions.

Although these studies rely on different imaging modalities, they collectively show how explainable AI enhances transparency in clinical decision systems and underscore the need to unify all three pillars of trustworthiness in a single framework, as pursued in our proposed method.

Taken together, while prior studies have made important contributions in segmentation based modeling, ensemble learning, adversarial robustness assessment, fairness aware evaluation and visual explainability, these components are most often explored independently. Despite these advances, no existing framework systematically integrates tissue-specific CNN optimization, ensemble fusion, adversarial robustness evaluation, gender based fairness analysis, and visual explainability within a unified diagnostic pipeline for distinguishing sMCI vs. pMCI. This methodological fragmentation directly provides the central motivation for the integrated, system-level architecture proposed in this study.

3. Preliminaries

This section introduces the core components supporting the development of the proposed trustworthy MRI-based diagnostic framework. The study uses tissue-segmented 2D MRI slices derived from GM, WM, and CSF, each providing complementary structural characteristics associated with AD progression. Deep feature extraction is performed using established CNN backbones, including VGG-16, DenseNet-121, InceptionV3, and EfficientNetV2, whose outputs are later combined within an ensemble learning strategy. Bayesian optimization is employed to tune the hyperparameters of each backbone to ensure stable and efficient learning. To assess generalization and cross-cohort robustness, the framework is additionally evaluated on the NACC dataset, with a ViT included as an additional baseline comparator. Both homogeneous and heterogeneous ensemble configurations are examined. We evaluate our system across all three pillars of trustworthy AI, including adversarial robustness, gender-based fairness, and visual explainability to ensure reliable and clinically meaningful AD progression analysis.

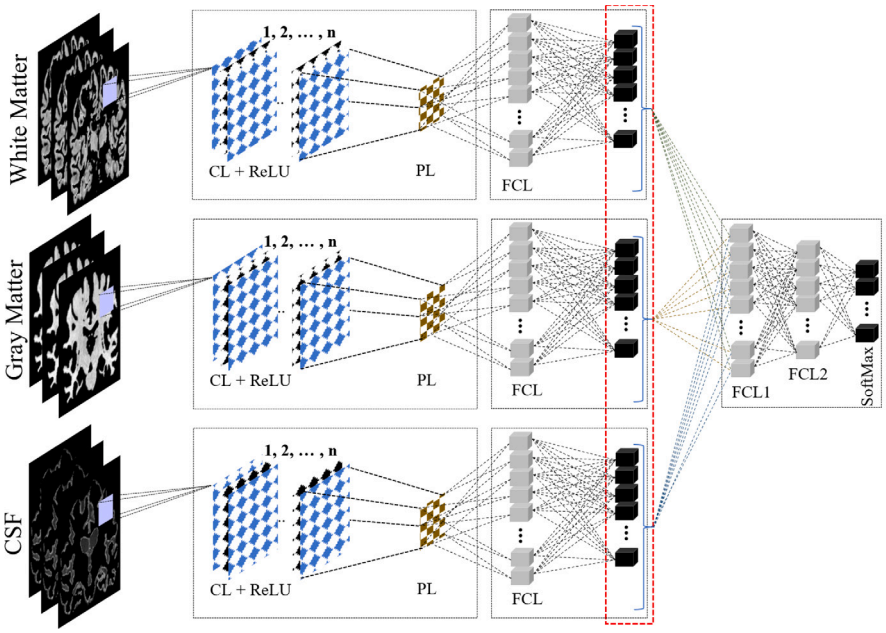

Fig. 3. Proposed ensemble framework for classifying sMCI vs. pMCI using segmented brain regions and an optimized CNN backbone.

3.1. Convolutional neural networks

CNNs are widely adopted in medical image analysis due to their ability to learn hierarchical spatial representations through local receptive fields and weight sharing, making them well suited for high-dimensional MRI data. Compared with fully connected models such as Multilayer Perceptrons (MLPs), CNNs support more efficient feature learning and improved stability when combined with normalization and regularization techniques, although excessive depth may negatively affect generalization (Valizadeh et al., 2025). In this work, VGG, DenseNet, InceptionV3, and EfficientNetV2 are employed as baseline CNN architectures. While all follow the same convolution-based learning paradigm, they differ in connectivity, feature propagation, and computational efficiency, with DenseNet enabling feature reuse through dense connections and VGG relying on a sequential design (Valizadeh et al., 2025). These models provide diverse and well-established reference baselines for evaluating tissue-specific MRI feature extraction.

3.2. Tissue-specific CNN design and optimization strategy

CNNs have demonstrated strong capability in extracting discriminative representations from high-dimensional data across 1D, 2D, and 3D domains. In medical imaging applications, 2D CNNs are commonly adopted when working with slice-based inputs due to their computational efficiency and stable training behavior. Since the present study focuses on AD progression using 2D MRI slices, a 2D CNN architecture was selected as the foundational modeling unit. Following tissue segmentation, separate CNN models were designed for each anatomical component, namely GM, WM, and CSF. This tissue-aware strategy enables each network to focus on learning modality-specific structural characteristics rather than entangling heterogeneous patterns within a single model. Each CNN follows a standard hierarchical structure composed of convolutional layers, pooling layers, and fully connected layers, culminating in a Softmax based classifier. To avoid manual architecture tuning and to ensure fair and systematic model design across tissues, Bayesian optimization was adopted as the core hyperparameter selection mechanism. Rather than enforcing a single shared architecture, the optimizer independently identified optimal configurations for GM, WM, and CSF, allowing each tissue-specific CNN to adapt its depth, capacity, and regularization strength according to its anatomical properties. This process resulted in three complementary yet structurally distinct models that form the basis of the proposed framework. After optimizing the tissue-specific CNNs, the final Softmax layers were removed and the networks were frozen to preserve the learned feature representations. These features were subsequently fused through additional fully connected layers, forming an ensemble classifier that integrates complementary information from GM, WM, and CSF. The overall integration and fusion workflow is illustrated in Fig. 3.

3.3. Ensemble learning

Ensemble learning (EL) combines multiple models to improve predictive performance and generalization on unseen data. Common ensemble strategies include bagging, boosting, and stacking, which aim to reduce variance, bias, or both by leveraging model diversity (Dong et al., 2020). Ensembles may be constructed in homogeneous settings, where the same algorithm is trained on different data subsets, or in heterogeneous settings, where different algorithms are applied to the same data (Chunxia and Jiangshe, 2011). Model outputs are typically aggregated using averaging or voting schemes, with majority voting commonly employed for classification tasks. The choice of ensemble configuration and fusion strategy depends on task characteristics and the desired balance between robustness, performance stability, and interpretability.

3.4. Trustworthy AI

Trustworthy AI refers to the development of ML systems that operate reliably, transparently, and in a manner aligned with ethical and societal expectations. In medical imaging, a key challenge arises from the black-box nature of many DL models, which limits insight into how predictions are generated and raises concerns about robustness, fairness and interpretability. To address these issues, prior work has focused on improving model resilience to adversarial or environmental perturbations, reducing biased outcomes across population subgroups and enhancing explainability to support human understanding of decisions. Together, these aspects form the foundation of trustworthy AI systems suitable for real-world deployment as illustrated in Fig. 2. In this study, we focus specifically on three core dimensions, including, adversarial robustness, gender fairness and explainability, which are defined operationally and evaluated in the subsequent subsections.

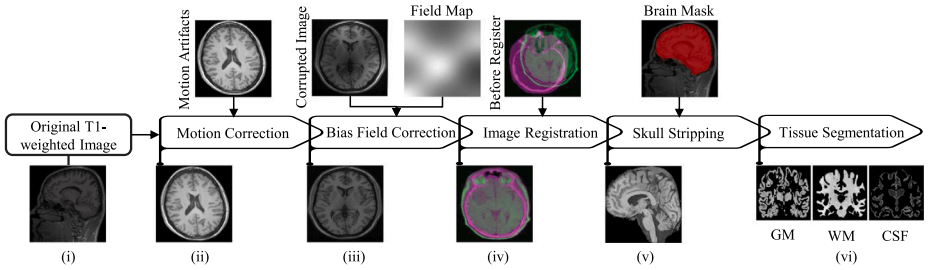

Fig. 4. Preprocessing pipeline of MRI: (i) original T1-weighted MRI, (ii) remove the motion artifacts, (iii) intensity inhomogeneity normalization, (iv) co-registration to MNI152 template, (v) cleaning non-brain matter, (vi) GM, WM, and CSF segments in MRI image.

3.4.1. Adversarial robustness

Adversarial robustness is a fundamental requirement for deploying ML models in safety-critical medical domains, as it ensures consistent performance even when the input data are corrupted by noise, variability, or malicious perturbations. In general, robustness refers to the stability of the predictive function when its input undergoes slight deviations due to acquisition artifacts, demographic heterogeneity, or adversarial manipulations (Minh et al., 2022). Formally, for a given input sample x and perturbation budget $\epsilon > 0$, a neighborhood can be defined as:

$$B_\epsilon(x) = \{x' \in \mathcal{X} : \|x' - x\| \le \epsilon\}, \tag{1}$$

where $\|\cdot\|$ denotes a norm such as ℓ_2 or ℓ_∞. A model f_w, parameterized by w, is regarded as robust if for all perturbed inputs $x' \in B_\epsilon(x)$ the model output does not deviate beyond a small tolerance δ:

$$\sup_{x' \in B_\epsilon(x)} \|f_w(x') - f_w(x)\| \le \delta. \tag{2}$$

For classification problems, robustness can also be formalized through the *adversarial risk*:

$$R_{\text{adv}}(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{x' \in B_\epsilon(x)} \mathbf{1}\{f_w(x') \neq y\}\right], \tag{3}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. In addition, certified robustness can be analyzed using Lipschitz continuity assumptions. Specifically, if f_w is L-Lipschitz continuous, then for all $x' \in B_\epsilon(x)$ it holds that

$$\|f_w(x') - f_w(x)\| \le L\epsilon \tag{4}$$

Adversarial robustness is particularly concerned with perturbations intentionally constructed to deceive a model. An adversarial example is an input x' such that $\|x' - x\| \le \epsilon$ but yields a misclassifications $f_w(x') \neq f_w(x)$. Among defense techniques, adversarial training has been extensively studied. This strategy incorporates adversarially perturbed samples into the training pipeline, which can be expressed as a minimax optimization problem:

$$\min_w \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\|x'-x\|\le\epsilon} \ell(f_w(x'), y)\right], \tag{5}$$

where $\ell(\cdot)$ denotes the task-specific loss function (e.g., cross-entropy). Although adversarial training significantly enhances robustness, it comes with practical limitations. In particular, it substantially increases computational complexity and may compromise the accuracy on clean, unperturbed data. This trade-off is highly consequential in medical applications, where both robustness and high baseline accuracy are indispensable. Consequently, adversarial robustness remains a key open challenge in the development of trustworthy medical AI systems.

3.4.2. Gender fairness

Fairness in ML aims to ensure that model predictions do not systematically disadvantage individuals or groups defined by sensitive attributes. In this study, we focus on gender-based fairness and evaluate whether diagnostic decisions are made equitably across male and female subpopulations using a group-level fairness perspective. Let $Y \in \{0, 1\}$ denote the ground-truth class label, where $Y = 1$ indicates the positive class, and let $\hat{Y} \in \{0, 1\}$ represent the model prediction. The protected attribute A corresponds to gender, with A = male or A = female.

Demographic Parity (DP) requires that the probability of a positive prediction is independent of gender (Tarzanagh et al., 2023):

$$P(\hat{Y} = 1 \mid A = \text{male}) = P(\hat{Y} = 1 \mid A = \text{female}). \tag{6}$$

Equal Opportunity (EO) evaluates sensitivity fairness by enforcing equal true positive rates across genders (Hardt et al., 2016):

$$P(\hat{Y} = 1 \mid A = \text{male}, Y = 1) = P(\hat{Y} = 1 \mid A = \text{female}, Y = 1). \tag{7}$$

Group Sufficiency (GS) assesses calibration fairness by requiring that the reliability of positive predictions is consistent across genders (Shui et al., 2022):

$$P(Y = 1 \mid \hat{Y} = 1, A = \text{male}) = P(Y = 1 \mid \hat{Y} = 1, A = \text{female}). \tag{8}$$

These fairness metrics provide a concise and clinically meaningful assessment of distributional balance, diagnostic sensitivity, and predictive reliability across gender groups.

3.4.3. Explainability

Explainability focuses on methods that make the internal reasoning of ML models more accessible and interpretable. Rather than treating the model as a black box, these approaches aim to clarify why specific predictions are made. Visualization-based techniques are particularly valuable in this regard, as they provide intuitive insights by showing which parts of the input most strongly influence the model's decision. Gradient-weighted Class Activation Mapping (Grad-CAM) is one of the most widely adopted visualization methods for explaining CNN predictions. Unlike the original Class Activation Mapping (CAM), which requires modifications such as global average PLs and retraining, Grad-CAM can be applied to standard CNN architectures without structural changes (Chattopadhay et al., 2018). By leveraging the gradients of target class scores with respect to the final CL, Grad-CAM highlights discriminative image regions that influence the prediction. This capability makes it valuable for tasks such as weakly supervised localization and segmentation. For a given class c, the classification score Y^c can be expressed as a weighted combination of the activation maps A^k:

$$Y^c = \sum_k w_k^c \sum_i \sum_j A_{ij}^k \tag{9}$$

where A_{ij}^k denotes the activation at location (i, j) of the kth feature map. The corresponding class-discriminative localization map L_{ij}^c is then obtained as:

$$L_{ij}^c = \sum_k w_k^c A_{ij}^k \tag{10}$$

with w_k^c representing the importance of feature map A^k for class c. These weights are determined by backpropagating gradients:

$$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \tag{11}$$

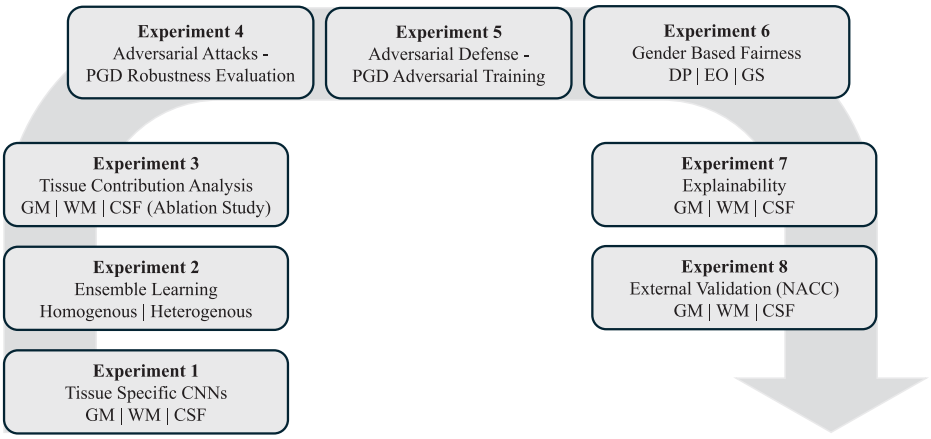

Fig. 5. Overview of the experimental design and evaluation pipeline of MRI-DeepTrust.

where Z is the total number of pixels in the activation map. By combining these weighted activations, Grad-CAM produces heatmaps that visually emphasize the regions most relevant to the model's decision, thereby improving interpretability and enhancing user trust.

4. Proposed trustworthy framework

This section presents the proposed trustworthy framework and describes its core building blocks in a unified and coherent manner. As illustrated in Fig. 1, the pipeline begins with MRI data acquisition followed by standardized preprocessing and tissue segmentation to derive structurally meaningful inputs. The processed data are then partitioned into training and testing subsets and fed into a Bayesian optimization driven convolutional neural network, where model architecture and learning parameters are systematically tuned to ensure reliable and discriminative feature extraction. Beyond predictive performance, the framework explicitly incorporates trustworthiness assessment modules. Model robustness is evaluated under both natural data variations and adversarial perturbations to examine stability against distributional shifts and malicious attacks. In parallel, individual level fairness analysis is conducted to quantify and mitigate biased behavior across subjects. Finally, explainability mechanisms are applied to the optimized model to generate interpretable outputs, enabling transparent inspection of learned patterns and supporting trustworthy clinical decision making.

4.1. MRI data

This study relies on sMRI data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a widely used and openly accessible research resource.[1] MRI scans were collected in March 2019 across the ADNI-1, ADNI-GO, and ADNI-2 phases. To ensure a well defined and clinically consistent cohort, only baseline MRI scans from subjects diagnosed with sMCI and pMCI were included. Since AD is considered an irreversible condition, subjects whose diagnostic labels reverted over time, for example from AD to MCI, were regarded as potentially misdiagnosed and were therefore excluded (Abuhmed et al., 2021). Similarly, individuals who progressed directly from CN to AD were removed to maintain diagnostic continuity. The resulting dataset comprises 592 baseline MRI volumes. All scans correspond to 3T T1 weighted anatomical images acquired using the volumetric 3D MPRAGE protocol, with an isotropic voxel resolution of $1 \times 1 \times 1$ mm. Prior to model development, the MRI volumes were subjected to a standardized preprocessing pipeline designed to ensure data quality and consistency, which is described in the following section.

[1] https://adni.loni.usc.edu/

4.2. MRI preprocessing

Following data acquisition, the MRI scans undergo a sequence of preprocessing steps designed to support accurate brain tissue segmentation and ensure consistency across subjects. An overview of the complete MRI preprocessing workflow is provided in Fig. 4, which illustrates the standard pipeline adopted for MRI preparation and tissue segmentation. The preprocessing stages include motion correction, bias field correction, spatial registration, skull stripping, and tissue segmentation. These operations enhance image quality, reduce the influence of artifacts, and produce reliable inputs for subsequent analyses.

4.2.1. Motion correction and reorientation:

An essential initial step in the preprocessing pipeline involves correcting for subject motion and enforcing a consistent image orientation across all scans. This step standardizes voxel alignment and supports reliable processing across different software environments. Prior to applying subsequent preprocessing procedures, all scans were visually inspected to identify potential orientation errors or imaging artifacts. Open-source visualization tools such as FSLEyes (McCarthy, 2020) and FreeView (Jahn, 2020) are commonly used for this purpose. In this study, FSLEyes was employed for scan inspection, during which several images were found to be rotated by 180 degrees. These orientation inconsistencies were corrected using the `fslreorient2std` utility from the FSL package, thereby ensuring standardized orientation before proceeding to later preprocessing stages.

4.2.2. Bias field correction

MRI data are often affected by intensity inhomogeneities arising from variations in patient positioning, scanner hardware, or software configurations. These artifacts typically manifest as smooth, low-frequency intensity variations across the image and can degrade image quality. If left uncorrected, such inhomogeneities may adversely affect downstream processes, including skull stripping and tissue segmentation. To mitigate this issue, N4 bias field correction from Advanced Normalization Tools (ANTs) was applied. N4 is an improved extension of the earlier N3 algorithm and employs an enhanced B-spline fitting strategy with multi-resolution optimization, enabling more accurate and robust correction of intensity non-uniformities.

4.2.3. MNI152 standard template registration

To establish spatial correspondence across subjects, individual brain images were aligned to a common reference space. In this work, affine registration to the MNI152 standard template was performed. Affine transformations involve basic geometric operations, including scaling, rotation, translation, and shearing, and preserve the overall anatomical structure without introducing nonlinear deformations. Registration was

Table 1
Layer-wise configuration of each module trained on GM, WM, and CSF. The input size is the same for all the CNN modules, i.e., $224 \times 224 \times 1$. Each layer configures for a number of kernels, their size, stride, dropout, and output size.

ID	Layer	Number of Krn.			Krn./ Str./ Drp.			Output Size		
		GM	WM	CSF	GM	WM	CSF	GM	WM	CSF
	Input	224 × 224 × 1								
1	Conv1	96	96	128	5 × 5/2	5 × 5/2	3 × 3/1	110 × 110 × 96	110 × 110 × 96	222 × 222 × 128
2	Conv2	96	128	128	5 × 5/1	3 × 3/1	3 × 3/1	108 × 108 × 96	106 × 106 × 128	220 × 220 × 128
3	Batch norm.	–	–	–	–	–	–	108 × 108 × 96	106 × 106 × 128	220 × 220 × 128
4	Max pooling	–	–	–	2 × 2/2	2 × 2/2	2 × 2/2	54 × 54 × 96	53 × 53 × 128	110 × 110 × 128
5	Conv3	128	128	192	3 × 3/1	3 × 3/1	3 × 3/1	52 × 52 × 128	51 × 51 × 128	108 × 108 × 192
6	Conv4	160	192	192	3 × 3/1	3 × 3/1	3 × 3/1	50 × 50 × 160	49 × 49 × 192	106 × 106 × 192
7	Batch norm.	–	–	–	–	–	–	50 × 50 × 160	49 × 49 × 192	106 × 106 × 192
8	Max pooling	–	–	–	2 × 2/2	2 × 2/2	2 × 2/2	25 × 25 × 160	24 × 24 × 192	53 × 53 × 192
9	Conv5	192	192	192	3 × 3/1	3 × 3/1	3 × 3/1	23 × 23 × 192	22 × 22 × 192	51 × 51 × 192
10	Conv6	192	192	256	3 × 3/1	3 × 3/1	3 × 3/1	21 × 21 × 192	20 × 20 × 192	49 × 49 × 256
11	Batch norm.	–	–	–	–	–	–	21 × 21 × 192	20 × 20 × 192	49 × 49 × 256
12	Max pooling	–	–	–	2 × 2/2	2 × 2/2	2 × 2/2	10 × 10 × 192	10 × 10 × 192	24 × 24 × 256
13	Conv7	224	–	256	3 × 3/1	–	3 × 3/1	8 × 8 × 224	–	22 × 22 × 256
14	Conv8	256	–	256	3 × 3/1	–	3 × 3/1	6 × 6 × 256	–	20 × 20 × 256
15	Batch norm.	–	–	–	–	–	–	6 × 6 × 256	–	20 × 20 × 320
16	Max pooling	–	–	–	2 × 2/2	–	2 × 2/2	3 × 3 × 256	–	10 × 10 × 320
17	Flatten	–	–	–	–	–	–	2304	19 200	32 000
18	Fully connected	256	256	256	–	–	–	256	256	256
19	Dropout	–	–	–	0.4	0.3	0.1	256	256	256
20	Fully connected	256	128	128	–	–	–	256	128	128
21	Dropout	–	–	–	0.5	0.2	0.0	256	128	128
22	Output layer	2	2	2	–	–	–	2	2	2

Krn:Kernel; Str:Stride; Drp:Dropout.

conducted using the FLIRT utility from the FSL software suite, with the correlation ratio selected as the similarity metric to achieve accurate alignment.

4.2.4. Skull stripping

Skull stripping aims to remove non-brain tissues from MRI scans in order to isolate relevant brain structures for subsequent analysis. Accurate brain extraction allows extraneous regions, such as residual neck voxels, to be eliminated, thereby improving data quality and reducing unnecessary variability. Retaining non-brain tissues increases data dimensionality and introduces noise, which can negatively affect classification performance. In this study, initial skull stripping was performed using the Brain Extraction Tool (BET2) (Jenkinson et al., 2005) from the FSL software suite. However, in several cases, the presence of extensive neck voxels limited the effectiveness of BET2 in achieving precise brain extraction. To address this limitation, an additional strategy incorporating segmentation based refinement was employed, resulting in more reliable skull stripping outcomes.

4.3. Tissue segmentation

As the final step of the MRI preprocessing pipeline illustrated in Fig. 4, brain tissue segmentation was performed to enable region specific analysis within the proposed framework. Using FSL 5.0, each MRI volume was segmented into three primary tissue classes, namely GM, WM, and CSF, resulting in separate binary masks for each tissue type. These tissue-specific masks form a core component of the proposed framework, as they allow the learning process to focus on anatomically meaningful regions rather than the whole brain volume. To ensure the reliability of the segmentation results, all generated masks were visually inspected to verify anatomical accuracy and consistency across subjects. This quality control step ensured that the segmented GM, WM, and CSF volumes provided robust and dependable inputs for the subsequent classification tasks.

4.4. Bayesian optimization module

Within the proposed framework, Bayesian optimization is employed as a dedicated module to guide the design of tissue-specific CNN models for GM, WM, and CSF. Rather than relying on manually selected architectures, this module systematically searches the hyperparameter space to identify configurations that best capture modality specific structural characteristics. The optimization process jointly considers architectural and training related factors, enabling each tissue-specific CNN to adapt its depth, feature capacity, and regularization strength according to the underlying anatomical properties. Although a common search space is defined across all tissues, Bayesian optimization yields distinct optimized architectures for GM, WM, and CSF, resulting in complementary feature representations that are later integrated through ensemble fusion. By embedding Bayesian optimization directly into the framework, the proposed system ensures consistent, data-driven model design and avoids ad hoc architectural choices, thereby strengthening both performance and reproducibility.

4.5. Robustness

Robustness is treated as a central design objective of the proposed framework, driven by the well-known sensitivity of DL models to noise, acquisition variability, and adversarial perturbations in medical imaging. Rather than viewing robustness only as a post hoc property to be measured after training, the framework is intentionally structured to support stable and reliable predictions under both clean and perturbed input conditions. To achieve this, robustness is addressed at two complementary levels. First, the framework makes use of adversarial perturbations generated through Projected Gradient Descent (PGD) to examine the stability of the learned representations. By evaluating tissue-specific CNN models on adversarially perturbed inputs, the framework investigates whether diagnostic predictions remain consistent when exposed to bounded yet challenging input variations. This approach provides a systematic way to analyze model sensitivity to perturbations that may stem from noise, scanner variability, or unintended disturbances in clinical settings. In addition, robustness is strengthened through ensemble based integration of tissue-specific models. Instead of relying on a single predictive pathway, the framework combines independently optimized CNNs trained on GM, WM, and CSF. This design introduces both diversity and redundancy into the decision process, which helps limit performance degradation when individual subnetworks are influenced by perturbations. In practice, errors arising

Table 2
Performance of tissue-specific CNN models trained independently on GM, WM, and CSF. The best-performing result for each metric is highlighted in bold.

Model	BR	Precision	Recall	F1-score	AUC	Accuracy
VGG-16	GM	74 ± 0.05	72 ± 0.04	73 ± 0.06	72 ± 0.05	69 ± 0.05
	WM	71 ± 0.04	72 ± 0.07	70 ± 0.03	72 ± 0.04	68 ± 0.03
	CSF	**83 ± 0.05**	**81 ± 0.03**	**82 ± 0.06**	**83 ± 0.04**	**79 ± 0.05**
DenseNet-121	GM	75 ± 0.07	77 ± 0.03	76 ± 0.06	74 ± 0.04	71 ± 0.05
	WM	**81 ± 0.05**	**79 ± 0.04**	**80 ± 0.07**	**81 ± 0.02**	**78 ± 0.06**
	CSF	76 ± 0.02	74 ± 0.02	75 ± 0.02	74 ± 0.03	70 ± 0.02
InceptionV3	GM	**83 ± 0.03**	**81 ± 0.02**	**83 ± 0.03**	**82 ± 0.03**	**80 ± 0.02**
	WM	73 ± 0.05	74 ± 0.03	72 ± 0.07	74 ± 0.05	69 ± 0.04
	CSF	73 ± 0.04	71 ± 0.05	72 ± 0.06	73 ± 0.03	70 ± 0.05
EfficientNetV2	GM	80 ± 0.05	81 ± 0.07	81 ± 0.03	80 ± 0.04	78 ± 0.04
	WM	**82 ± 0.02**	**81 ± 0.02**	**81 ± 0.03**	**80 ± 0.02**	**78 ± 0.03**
	CSF	82 ± 0.04	80 ± 0.06	81 ± 0.04	81 ± 0.05	79 ± 0.03
Proposed	GM	82 ± 0.04	80 ± 0.03	81 ± 0.05	82 ± 0.03	82 ± 0.04
	WM	**84 ± 0.05**	**81 ± 0.04**	**82 ± 0.05**	**83 ± 0.04**	**79 ± 0.03**
	CSF	**83 ± 0.03**	**83 ± 0.02**	**82 ± 0.03**	**83 ± 0.05**	**80 ± 0.04**

BR: Brain region corresponding to the segmented tissue (GM, WM, or CSF).

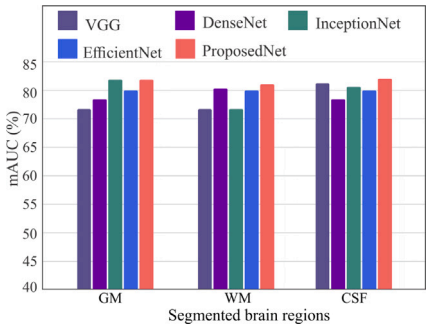

Fig. 6. Comparison of mAUC scores for sMCI vs. pMCI classification using tissue-specific CNN models trained on GM, WM, and CSF. Results are shown for individual backbone architectures (VGG, DenseNet, EfficientNet, and InceptionNet) and the proposed MRI-DeepTrust model, illustrating the impact of tissue segmentation on classification performance.

in one tissue-specific pathway can be compensated by complementary information from the others, resulting in more stable subject-level predictions. By jointly examining resilience to adversarial perturbations and leveraging the stabilizing effect of ensemble diversity, the proposed framework is designed to deliver consistent diagnostic behavior across challenging input conditions. This integrated view of robustness supports the broader goal of developing trustworthy AI systems for healthcare, where dependable performance under realistic and potentially adverse conditions is essential.

4.6. Fairness

Fairness is treated as a fundamental component of the proposed trustworthy framework, motivated by the growing evidence that DL models in medical imaging may exhibit systematic performance disparities across demographic groups. In the context of neurodegenerative disease diagnosis, such disparities are particularly concerning, as biased predictions can influence clinical decision making and exacerbate existing healthcare inequalities. To address this challenge, the framework explicitly incorporates fairness considerations throughout the model evaluation pipeline, with a focus on gender as a protected attribute. Rather than assuming demographic neutrality, the proposed framework is designed to examine whether diagnostic performance remains consistent across gender defined subgroups. This is achieved by evaluating tissue-specific CNNs and the final ensemble model under multiple controlled gender based data splits.

By analyzing model behavior across these settings, the framework aims to identify potential performance imbalances that may arise due to unequal representation, anatomical differences, or implicit bias learned during training. Fairness assessment is applied consistently across all components of the framework, including individual GM, WM, and CSF models as well as their ensemble integration. This unified evaluation strategy ensures that fairness is not treated as an isolated post-processing step but is instead examined alongside predictive accuracy, robustness, and interpretability. In doing so, the framework enables a comprehensive understanding of how demographic factors interact with tissue-specific representations and ensemble decision making. By embedding fairness-aware evaluation into the overall design, the proposed framework supports the development of diagnostic models that are not only accurate but also equitable. This perspective aligns with the broader objectives of trustworthy AI in healthcare, where reliable performance across diverse patient populations is essential for safe and responsible clinical deployment.

4.7. Explainability

In the clinical context of AD and MCI, reliable diagnostic systems must extend beyond predictive accuracy to provide insights that clinicians can meaningfully interpret. To address this need, explainability is embedded as a core component of the proposed framework, with the aim of bridging the gap between DL predictions and neuroanatomical understanding. The framework adopts a dual level explainability strategy that combines global structural visualization with localized attribution analysis. At the global level, three-dimensional surface rendering is employed to reconstruct and visualize brain tissues derived from MRI scans. These renderings allow clinicians to examine structural variations across GM, WM, and CSF, offering anatomically meaningful views of cortical and subcortical regions commonly affected during disease progression. Such visualizations support intuitive assessment of tissue-level changes that underlie model predictions.

At the local level, fine-grained interpretability is achieved through voxel-wise attribution using two-dimensional attention maps. These maps highlight regions within individual MRI slices that contribute most strongly to the model's predictions, thereby revealing localized discriminative patterns associated with disease status. The CNN-based modules incorporate an attention mechanism through Med-Cam (Gotkowski et al., 2020), which overlays saliency information directly onto the original image slices. This representation provides clinicians with an intuitive view of the features driving the distinction between stable and progressive MCI. Together, the combination of three-dimensional structural visualizations and two-dimensional attention maps provides complementary perspectives on model behavior.

Table 3
Performance comparison of homogeneous and heterogeneous ensemble models constructed from tissue-specific CNNs trained on GM, WM, and CSF. Results illustrate the effect of ensemble strategy on classification performance for sMCI vs. pMCI.

Model	Precision	Recall	F1-score	AUC	Accuracy
Proposed EL	**92 ± 0.02**	**89 ± 0.02**	**90 ± 0.02**	**89 ± 0.03**	**88 ± 0.03**
Ensemble VGG-16*	84 ± 0.04	83 ± 0.06	81 ± 0.03	82 ± 0.05	78 ± 0.03
Ensemble DenseNet-121*	85 ± 0.04	87 ± 0.03	86 ± 0.03	85 ± 0.02	86 ± 0.04
Ensemble InceptionNet*	79 ± 0.02	80 ± 0.04	78 ± 0.05	79 ± 0.03	76 ± 0.02
Ensemble EfficientNet*	84 ± 0.05	85 ± 0.03	84 ± 0.04	84 ± 0.02	81 ± 0.03
InceptionNet–EfficientNet–DenseNet	89 ± 0.04	88 ± 0.03	88 ± 0.02	87 ± 0.03	89 ± 0.02

* Homogeneous ensemble.

While global renderings capture broader anatomical context, localized saliency maps expose focal disease related patterns at the slice level. This multi-scale explainability framework enhances clinical interpretability and supports informed decision making. By integrating explainability directly into the system design, the proposed framework reinforces its trustworthiness and aligns with emerging expectations for transparent and responsible use of AI in healthcare.

5. Experimental setup

This section describes the experimental protocols and implementation details used to train and evaluate the proposed framework. It outlines the subject-level data partitioning strategy, training and optimization procedures, and evaluation metrics. These components ensure a consistent and reproducible assessment across all experiments, including tissue-specific CNN models, homogeneous and heterogeneous ensembles, adversarial robustness analysis, gender-based fairness evaluation and explainability analysis.

5.1. Parameter details

For reproducibility, we summarize the key hyperparameter ranges and experimental settings used across all components of the proposed framework. The Bayesian optimizer searched over convolutional depth (4 to 10 layers), number of filters (96 to 320), kernel sizes (3×3 or 5×5), dropout rates (0.1 to 0.5), dense layer widths (128 to 512), and learning rates $1e-2, 1e-3, 1e-4$. The final selected configurations for GM, WM, and CSF are reported in Table 1. The ensemble classifier consisted of two fully connected layers with 256 and 128 units, followed by a Softmax output layer. For adversarial robustness evaluation, PGD-based attacks were generated using a perturbation budget of $\epsilon = 0.01$, a step size of 0.002, and 10 iterations. Fairness experiments used gender as the sensitive attribute and followed the same training configurations and subject-level cross-validation splits as the main classification experiments and fairness metrics were computed post hoc on the test folds.

5.2. Subject-level cross-validation

All experiments were conducted using a stratified 10-fold cross-validation strategy applied at the subject level. Subjects were partitioned into ten non-overlapping folds while preserving the proportion of stable and progressive MCI cases. For each fold, one subset was used for testing, one for validation, and the remaining eight for training. To prevent information leakage, all two-dimensional slices derived from the same subject were assigned to the same fold. This ensured that models were evaluated on entirely unseen subjects rather than unseen slices from subjects encountered during training. The same partitioning strategy was consistently applied across all experiments, including ensemble learning, robustness evaluation, and fairness analysis.

5.3. Training details

All CNNs and the final ensemble classifier were trained using the binary cross entropy loss. This loss function is commonly adopted for two-class medical imaging problems, as it encourages the network to assign higher confidence to the correct class and supports stable gradient based optimization. A step size of 32 was used for sampling values within the hyperparameter space. To mitigate overfitting, both dropout and Batch Normalization were incorporated, and early stopping was applied to terminate training once the validation accuracy plateaued.

The Bayesian optimizer was executed over 25 epochs with a batch size of 16 and a maximum of 75 trials. The upper trial limit was deliberately chosen to ensure adequate exploration of the search space and to increase the likelihood of identifying optimal hyperparameter sets. The final hyperparameter configurations determined for each tissue (GM, WM, CSF) are provided in Table 1. For ensemble experiments, each tissue-specific subnetwork was trained independently using identical optimization settings. Their learned feature representations were concatenated and passed to a fully connected classification head. The same training configuration was maintained across adversarial robustness and fairness experiments to ensure consistency.

5.4. Evaluation metrics

Model performance was evaluated using standard clinical classification metrics, including precision, recall, F1-score, area under the ROC curve (AUC), and accuracy. Predictions were computed at the subject level by aggregating slice-level probabilities through mean pooling. Fairness was evaluated by examining the consistency of model performance across gender-defined subgroups. For this purpose, the same classification metrics were computed separately for each subgroup, and performance differences in performance were analyzed to identify potential disparities in diagnostic behavior. For each experiment, reported results represent the mean and standard deviation across the ten cross-validation folds. This evaluation protocol was applied uniformly to individual tissue-based models, homogeneous and heterogeneous ensemble configurations, adversarial robustness experiments, and gender-based fairness analyses.

6. Results and discussion

This section presents the experimental results and key insights for distinguishing sMCI vs. pMCI. A structured evaluation was conducted, including tissue-specific CNN training, homogeneous and heterogeneous ensemble analysis, tissue ablation, adversarial robustness assessment, gender-based fairness evaluation, visual explainability analysis, and external validation to examine generalizability. The overall experimental workflow is summarized in Fig. 5.

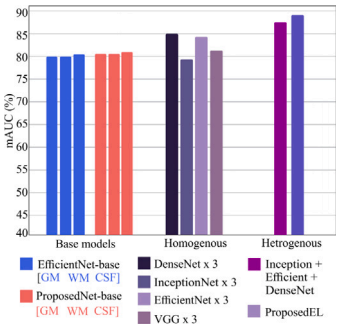

Fig. 7. Performance comparison of base CNN models and their ensemble variants using segmented brain regions. The plot contrasts individual models (left), homogeneous ensembles formed by repeating the same architecture three times (center), and heterogeneous ensembles combining different architectures (right). The results highlight how ensemble strategies influence mAUC values for sMCI vs. pMCI classification.

6.1. Experiment 1:Tissue-specific model performance

Experiment 1 evaluates the discriminative capability of GM, WM, and CSF for distinguishing sMCI vs. pMCI using multiple CNN backbones. Mean performance and corresponding standard deviations are reported in Table 2, with mAUC comparisons shown in Fig. 6. For GM, InceptionV3 achieved the highest overall performance, while the proposed model and EfficientNet followed closely. In WM, DenseNet, EfficientNet, and the proposed model demonstrated competitive results, with the proposed model showing consistent improvements across most metrics. For CSF, several models achieved performance comparable to GM and WM, indicating that CSF also contributes meaningful disease related information.

Despite satisfactory accuracy levels, relatively higher standard deviations suggest stability limitations in single tissue-specific models. Fig. 6 further highlights that EfficientNet and the proposed model maintain mAUC values above 80% across tissues, while other architectures exhibit tissue-dependent variability. These findings indicate complementary strengths across tissues and motivate the use of ensemble learning to improve robustness and consistency.

6.2. Experiment 2: Homogeneous and heterogeneous ensemble performance

Experiment 2 evaluates the impact of homogeneous and heterogeneous EL on MCI classification. Each ensemble combines tissue-specific subnetworks trained on GM, WM, and CSF, with features fused through a DNN classifier. Homogeneous ensembles employ identical backbone architectures across all tissues, whereas heterogeneous ensembles integrate diverse backbone models to capture complementary feature representations. The mean classification performance and corresponding standard deviations are reported in Table 3. For the proposed EL network, Bayesian-optimized weights presented in Table 1 were employed, while the Inception-Efficient-DenseNet ensemble was constructed by selecting the most stable models for each region based on their minimal standard deviations in Table 2. Among homogeneous ensembles, DenseNet based combinations achieved the strongest performance, while EfficientNet and VGG showed moderate improvements over their base counterparts. However, heterogeneous ensembles consistently outperformed homogeneous settings. The proposed ensemble achieved the best overall results (89% mAUC), surpassing both individual models and the Inception–EfficientNet–DenseNet ensemble.

Fig. 7 further illustrates that EL improves performance over single backbones, with heterogeneous configurations providing the largest gains. These findings confirm that architectural diversity reduces correlated errors and enhances robustness, supporting the integration of complementary tissue-specific representations for reliable classification of sMCI and pMCI.

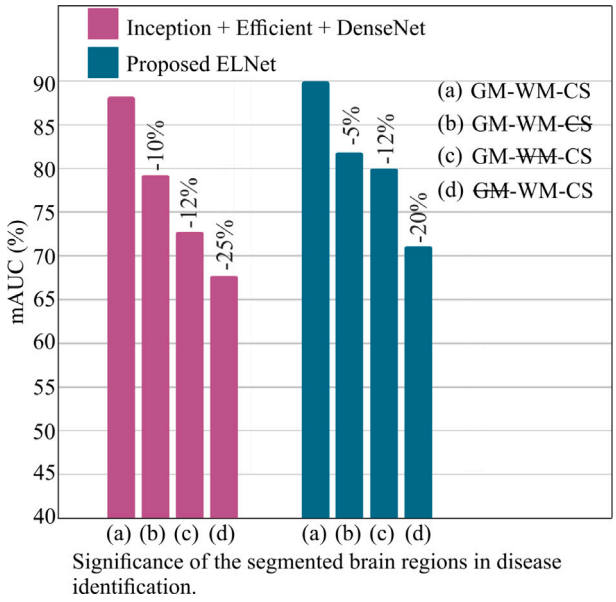

Fig. 8. Evaluation of mAUC performance changes when individual segmented brain tissues are excluded. The comparison between the heterogeneous ensemble and the proposed ELNet illustrates how different tissue omissions affect classification performance under partial-information conditions.

6.3. Experiment 3: Contribution of GM, WM, and CSF to MCI classification

Experiment 3 quantifies the contribution of each segmented tissue through an ablation study, where one region at a time was excluded from the heterogeneous ensembles and replaced with random noise. Results are summarized in Table 4, with mAUC degradation illustrated in Fig. 8. All three tissues contributed meaningfully to classification; however, GM emerged as the most critical. Excluding GM resulted in the largest performance drop, with mAUC reductions of approximately 20%–25% across ensembles. WM omission produced moderate degradation, while excluding CSF led to comparatively smaller mAUC declines between 5%–12%. These findings confirm the complementary roles of GM, WM, and CSF, while highlighting the dominant contribution of GM in reliably discriminating between sMCI and pMCI.

6.4. Experiment 4: Evaluation of model robustness to adversarial attacks

Experiment 4 evaluates robustness under PGD attacks applied to 10% of test samples across GM, WM, and CSF. Performance on clean and adversarial inputs is compared in Table 5, with degradation trends illustrated in Fig. 9. Adversarial perturbations caused consistent performance declines across all models. Base architectures exhibited mAUC reductions of approximately 16%–18%, with EfficientNet showing greater sensitivity and variance under attack. In contrast, the proposed model demonstrated comparatively stronger stability. Ensemble models further improved robustness, reducing mAUC degradation to approximately 13%, with the proposed ensemble achieving the most resilient performance. An additional ablation analysis under adversarial conditions reported in Table 6 revealed that excluding GM caused the largest performance drop (up to 19% accuracy loss), followed by WM, while CSF omission produced only modest degradation. These findings reinforce the dominant contribution of GM and WM to robust classification and demonstrate that ensemble diversity enhances resilience against adversarial perturbations.

Table 4
Analysis of the contribution of segmented brain regions (GM, WM, and CSF) to MCI classification performance.

Model	BR	Precision	Recall	F1-Score	AUC	Accuracy
InceptionNet–EfficientNet–DenseNet	GM–WM–CSF	89 ± 0.04	88 ± 0.03	88 ± 0.02	87 ± 0.03	89 ± 0.02
	~~GM~~–WM–~~CSF~~	78 ± 0.05	79 ± 0.07	77 ± 0.05	78 ± 0.03	78 ± 0.05
	GM–~~WM~~–CSF	75 ± 0.04	78 ± 0.08	77 ± 0.06	76 ± 0.04	77 ± 0.07
	~~GM~~–WM–CSF	66 ± 0.09	64 ± 0.05	64 ± 0.07	65 ± 0.04	64 ± 0.06
Proposed EL	GM–WM–CSF	92 ± 0.02	89 ± 0.02	90 ± 0.02	89 ± 0.03	88 ± 0.03
	GM–WM–~~CSF~~	81 ± 0.04	83 ± 0.03	83 ± 0.02	84 ± 0.04	84 ± 0.05
	GM–~~WM~~–CSF	78 ± 0.05	76 ± 0.03	77 ± 0.06	78 ± 0.04	78 ± 0.05
	~~GM~~–WM–CSF	71 ± 0.07	70 ± 0.05	70 ± 0.04	71 ± 0.03	72 ± 0.06

BR: Brain region corresponding to the segmented tissue (GM, WM, or CSF).

Table 5
Evaluation of model robustness under adversarial perturbations generated using PGD attacks.

Model	Image Type	BR	Precision	Recall	F1-Score	AUC	Accuracy
Proposed base*	Original	GM	85 ± 0.01	82 ± 0.02	81 ± 0.02	82 ± 0.03	82 ± 0.02
		WM	86 ± 0.02	81 ± 0.03	82 ± 0.03	83 ± 0.02	79 ± 0.01
		CSF	85 ± 0.02	83 ± 0.03	82 ± 0.02	85 ± 0.03	80 ± 0.01
	Adversarial	GM	71 ± 0.03	69 ± 0.04	68 ± 0.04	71 ± 0.04	72 ± 0.03
		WM	73 ± 0.02	70 ± 0.03	71 ± 0.03	69 ± 0.02	71 ± 0.03
		CSF	67 ± 0.04	71 ± 0.03	78 ± 0.04	78 ± 0.03	76 ± 0.04
EfficientNet**	Original	GM	80 ± 0.03	81 ± 0.03	81 ± 0.02	80 ± 0.03	78 ± 0.02
		WM	83 ± 0.02	84 ± 0.02	81 ± 0.04	80 ± 0.03	80 ± 0.03
		CSF	83 ± 0.04	82 ± 0.02	82 ± 0.03	81 ± 0.02	79 ± 0.03
	Adversarial	GM	66 ± 0.05	68 ± 0.04	67 ± 0.04	67 ± 0.04	66 ± 0.04
		WM	65 ± 0.03	64 ± 0.02	64 ± 0.04	65 ± 0.03	64 ± 0.04
		CSF	71 ± 0.04	67 ± 0.04	68 ± 0.03	66 ± 0.04	65 ± 0.03
InceptionNet–EfficientNet –DenseNet**	Original	GM WM CSF	89 ± 0.04	88 ± 0.03	88 ± 0.02	87 ± 0.03	86 ± 0.02
	Adversarial		77 ± 0.03	74 ± 0.04	76 ± 0.05	75 ± 0.04	73 ± 0.04
Proposed EL*	Original	GM WM CSF	92 ± 0.02	89 ± 0.02	90 ± 0.02	89 ± 0.03	88 ± 0.03
	Adversarial		80 ± 0.04	82 ± 0.03	80 ± 0.04	79 ± 0.04	80 ± 0.03

* First best model; ** Second best model; BR: Brain region corresponding to the segmented tissue (GM, WM, or CSF).

Table 6
Evaluating model's robustness using an adversarial attack by targeting a specific brain region.

Model	ATM	BR	Precision	Recall	F1-Score	AUC	Accuracy
InceptionNet–EfficientNet –DenseNet	No	GM-WM-CSF (All)	89 ± 0.04	88 ± 0.03	88 ± 0.02	87 ± 0.03	89 ± 0.02
	No	GM-WM-(AE-CSF)	73 ± 0.06	75 ± 0.05	74 ± 0.07	71 ± 0.06	73 ± 0.04
		GM-(AE-WM)-CSF	72 ± 0.06	74 ± 0.04	73 ± 0.06	73 ± 0.05	72 ± 0.07
		(AE-GM)-WM-CSF	62 ± 0.05	61 ± 0.05	62 ± 0.04	60 ± 0.06	61 ± 0.05
	Yes	GM-WM-(AS-CSF)	75 ± 0.04	77 ± 0.05	76 ± 0.05	76 ± 0.06	74 ± 0.09
		GM-(AE-WM)-CSF	76 ± 0.04	74 ± 0.05	75 ± 0.06	75 ± 0.07	75 ± 0.04
		(AE-GM)-WM-CSF	79 ± 0.05	75 ± 0.04	77 ± 0.04	74 ± 0.08	76 ± 0.05
Proposed EL	No	GM-WM-CSF (All)	92 ± 0.02	89 ± 0.02	90 ± 0.02	89 ± 0.03	88 ± 0.03
	No	GM-WM-(AE-CSF)	78 ± 0.04	79 ± 0.02	78 ± 0.04	75 ± 0.05	74 ± 0.03
		GM-(AE-WM)-CSF	76 ± 0.05	75 ± 0.03	75 ± 0.05	74 ± 0.06	76 ± 0.05
		(AE-GM)-WM-CSF	69 ± 0.06	68 ± 0.03	68 ± 0.05	69 ± 0.03	68 ± 0.04
	Yes	GM-WM-(AE-CSF)	81 ± 0.04	79 ± 0.03	80 ± 0.03	79 ± 0.04	80 ± 0.06
		GM-(AE-WM)-CSF	82 ± 0.06	81 ± 0.05	82 ± 0.05	82 ± 0.04	82 ± 0.03
		(AE-GM-WM-CSF)	85 ± 0.07	83 ± 0.03	83 ± 0.4	82 ± 0.03	81 ± 0.04

ATM: Adverserial Training Method; BR: Brain region corresponding to the segmented tissue (GM, WM, or CSF); AE: Adversarial Examples.

6.5. Experiment 5: Robustness under adversarial defense settings

Experiment. 5 evaluates adversarial training as a defense mechanism by incorporating perturbed samples into the training set at 20%, 35%, and 50% ratios. Models were then tested on adversarial inputs, with results reported in Table 7 and mAUC trends illustrated in Fig. 10. Adversarial training consistently improved robustness across configurations, though gains varied by model. At 20% augmentation, EfficientNet showed limited improvement of 70% mAUC, whereas the proposed base model and ensemble achieved higher resilience of 73%–74% and 81% mAUC respectively. Increasing the augmentation ratio further strengthened robustness. At 50%, the proposed ensemble reached 84% mAUC and 85% accuracy, outperforming all baselines. Fig. 10 demonstrates a stable and monotonic improvement for the proposed ensemble, while single-model baselines plateau earlier. These results indicate that adversarial training is most effective when combined with architectural diversity and multi-tissue fusion, reinforcing the robustness and scalability of the proposed framework for reliable clinical deployment.

6.6. Experiment 6: Gender-based fairness analysis

Experiment 6 evaluates gender-based fairness on the ADNI dataset, which exhibits a moderate yet realistic male–female imbalance reflecting real-world clinical conditions as shown in Table 8. Fairness was assessed using DP, EO, and GS, with lower values indicating reduced disparity. The quantitative results are presented in Table 9 and further illustrated in Fig. 11. Across GM, WM, and CSF, the proposed framework consistently achieves lower DP, EO, and GS gaps compared

Table 7
Evaluation of model performance under adversarial training settings, illustrating robustness behavior in the presence of adversarial perturbations.

Model	Ratio	BR	Precision	Recall	F1-Score	AUC	Accuracy
EfficientNet-base	20%	GM	69 ± 0.03	70 ± 0.03	69 ± 0.04	70 ± 0.03	68 ± 0.02
		WM	66 ± 0.04	65 ± 0.03	65 ± 0.04	66 ± 0.02	65 ± 0.04
		CSF	73 ± 0.05	72 ± 0.03	72 ± 0.04	70 ± 0.04	72 ± 0.03
Proposed base		GM	74 ± 0.04	71 ± 0.03	70 ± 0.03	73 ± 0.03	73 ± 0.04
		WM	74 ± 0.04	73 ± 0.04	73 ± 0.03	71 ± 0.04	73 ± 0.04
		CSF	72 ± 0.03	71 ± 0.03	72 ± 0.03	72 ± 0.04	72 ± 0.03
InceptionNet-EfficientNet-DenseNet		–	78 ± 0.03	76 ± 0.03	77 ± 0.03	77 ± 0.02	78 ± 0.03
Proposed EL		–	81 ± 0.04	80 ± 0.02	81 ± 0.03	81 ± 0.04	80 ± 0.03
EfficientNet-base	35%	GM	68 ± 0.05	71 ± 0.04	70 ± 0.04	72 ± 0.04	71 ± 0.04
		WM	67 ± 0.03	64 ± 0.02	66 ± 0.04	67 ± 0.03	65 ± 0.04
		CSF	73 ± 0.04	70 ± 0.04	72 ± 0.03	71 ± 0.04	71 ± 0.03
Proposed base		GM	75 ± 0.02	73 ± 0.03	74 ± 0.04	74 ± 0.04	73 ± 0.03
		WM	75 ± 0.05	74 ± 0.04	74 ± 0.04	74 ± 0.03	75 ± 0.02
		CSF	73 ± 0.03	70 ± 0.03	73 ± 0.03	73 ± 0.04	73 ± 0.03
InceptionNet-EfficientNet-DenseNet		–	79 ± 0.03	77 ± 0.03	78 ± 0.03	78 ± 0.02	79 ± 0.03
Proposed EL		–	83 ± 0.04	84 ± 0.02	82 ± 0.03	83 ± 0.04	84 ± 0.03
EfficientNet-base	50%	GM	74 ± 0.03	73 ± 0.04	74 ± 0.03	73 ± 0.04	75 ± 0.04
		WM	70 ± 0.03	71 ± 0.04	70 ± 0.03	72 ± 0.03	72 ± 0.03
		CSF	74 ± 0.04	71 ± 0.04	72 ± 0.03	72 ± 0.03	73 ± 0.03
Proposed base		GM	77 ± 0.02	75 ± 0.03	76 ± 0.04	77 ± 0.04	78 ± 0.03
		WM	76 ± 0.05	75 ± 0.04	75 ± 0.04	76 ± 0.03	76 ± 0.02
		CSF	74 ± 0.03	73 ± 0.03	74 ± 0.03	75 ± 0.04	75 ± 0.03
InceptionNet-EfficientNet-DenseNet		–	81 ± 0.03	79 ± 0.03	80 ± 0.04	82 ± 0.02	83 ± 0.03
Proposed EL		–	85 ± 0.04	83 ± 0.02	84 ± 0.03	84 ± 0.04	85 ± 0.03

BR: Brain region corresponding to the segmented tissue (GM, WM, or CSF).

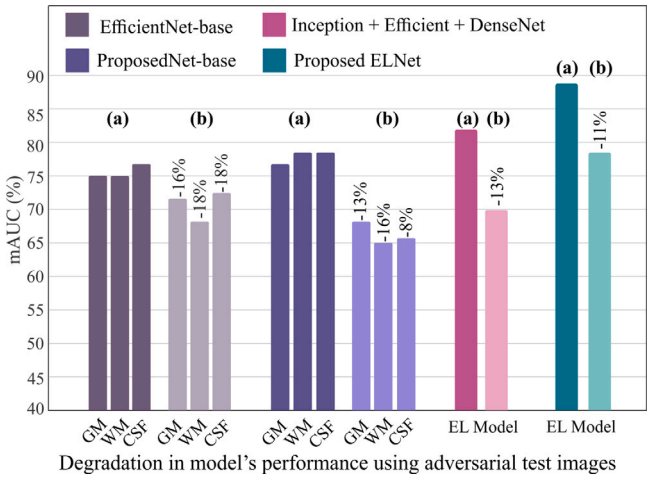

Fig. 9. mAUC performance changes observed under adversarial test conditions across base models, the heterogeneous ensemble, and the proposed EL. The figure compares model behavior when evaluated using adversarially perturbed GM, WM, and CSF inputs, highlighting differences in performance degradation under adversarial noise.

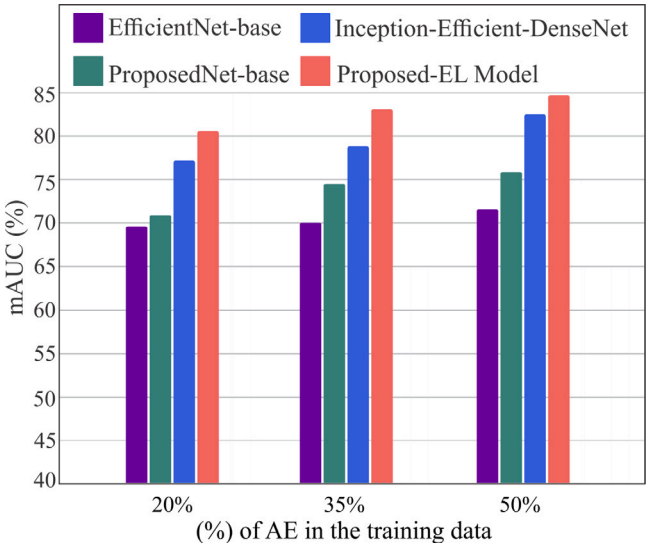

Fig. 10. Comparison of mAUC scores under varying proportions of adversarial examples incorporated during training and evaluation.

Table 8
Diagnostic class and gender distribution of the ADNI dataset used for fairness analysis.

Class	Count	Gender	Count
sMCI	330	Male	303
pMCI	262	Female	289

to baseline CNN architectures. The strongest fairness improvements are observed in GM and WM branches, while GS remains comparatively higher across all models due to its stricter calibration constraint. Nevertheless, the proposed model maintains the lowest GS disparities overall. Since all models were trained on identical data distributions, fairness differences stem from architectural design rather than dataset composition. These findings demonstrate that tissue-specific learning and ensemble fusion reduce gender related disparities without compromising diagnostic performance, reinforcing the framework's suitability for trustworthy clinical deployment.

6.7. Experiment 7: Explainability assessment of tissue-specific MRI representations

Experiment 7 evaluates the interpretability of the proposed framework using multi-scale visualization combining three-dimensional surface rendering and two-dimensional slice-level attention maps. Fig. 12 illustrates tissue-specific activations for GM, WM, and CSF in sMCI and

Table 9
Fairness evaluation of tissue-specific CNN models trained independently on GM, WM, and CSF. In addition to standard performance metrics, demographic parity (DP), equality of opportunity (EO), and group sufficiency (GS) are reported using gender as the sensitive attribute.

Model	BR	Precision	Recall	F1-score	AUC	Accuracy	DP	EO	GS
VGG-16	GM	74 ± 0.50	72 ± 0.09	73 ± 0.06	72 ± 0.43	69 ± 0.30	0.12 ± 0.06	0.17 ± 0.11	0.21 ± 0.03
	WM	71 ± 0.04	72 ± 0.17	70 ± 0.03	72 ± 0.04	68 ± 0.72	0.09 ± 0.03	0.15 ± 0.08	0.19 ± 0.75
	CSF	83 ± 0.10	81 ± 0.92	82 ± 0.06	83 ± 0.04	70 ± 0.15	0.13 ± 0.01	0.10 ± 0.05	0.19 ± 0.07
DenseNet-121	GM	75 ± 0.07	77 ± 0.31	76 ± 0.06	74 ± 0.04	71 ± 0.05	0.18 ± 0.10	0.20 ± 0.11	0.25 ± 0.04
	WM	81 ± 0.05	79 ± 0.40	80 ± 0.22	81 ± 0.34	78 ± 0.68	0.06 ± 0.09	0.18 ± 0.05	0.20 ± 0.02
	CSF	76 ± 0.02	74 ± 0.12	75 ± 0.02	74 ± 0.03	70 ± 0.11	0.13 ± 0.13	0.10 ± 0.20	0.19 ± 0.07
InceptionV3	GM	83 ± 0.03	81 ± 0.71	83 ± 0.71	82 ± 0.16	80 ± 0.95	0.08 ± 0.02	0.13 ± 0.09	0.20 ± 0.04
	WM	73 ± 0.42	74 ± 0.90	72 ± 0.07	74 ± 0.05	69 ± 0.24	0.12 ± 0.04	0.10 ± 0.03	0.19 ± 0.03
	CSF	73 ± 0.04	71 ± 0.43	72 ± 0.06	73 ± 0.82	70 ± 0.17	0.10 ± 0.07	0.18 ± 0.01	0.23 ± 0.09
EfficientNetV2	GM	80 ± 0.45	81 ± 0.73	81 ± 0.95	80 ± 0.04	78 ± 0.52	0.15 ± 0.07	0.17 ± 0.11	0.23 ± 0.01
	WM	82 ± 0.02	81 ± 0.62	81 ± 0.39	80 ± 0.16	78 ± 0.33	0.10 ± 0.04	0.12 ± 0.53	0.17 ± 0.09
	CSF	82 ± 0.11	80 ± 0.16	81 ± 0.04	81 ± 0.05	79 ± 0.11	0.12 ± 0.09	0.28 ± 0.05	0.22 ± 0.04
Proposed base	GM	82 ± 0.64	80 ± 0.25	81 ± 0.32	82 ± 0.43	82 ± 0.28	0.02 ± 0.26	0.04 ± 0.08	0.16 ± 0.03
	WM	84 ± 0.85	81 ± 0.32	82 ± 0.62	83 ± 0.12	79 ± 0.37	0.01 ± 0.08	0.03 ± 0.01	0.18 ± 0.05
	CSF	83 ± 0.56	83 ± 0.96	82 ± 0.91	83 ± 0.54	80 ± 0.04	0.04 ± 0.05	0.08 ± 0.05	0.10 ± 0.04
Proposed EL	–	92 ± 0.02	89 ± 0.02	90 ± 0.02	89 ± 0.03	88 ± 0.03	0.01 ± 0.06	0.02 ± 0.05	0.09 ± 0.03

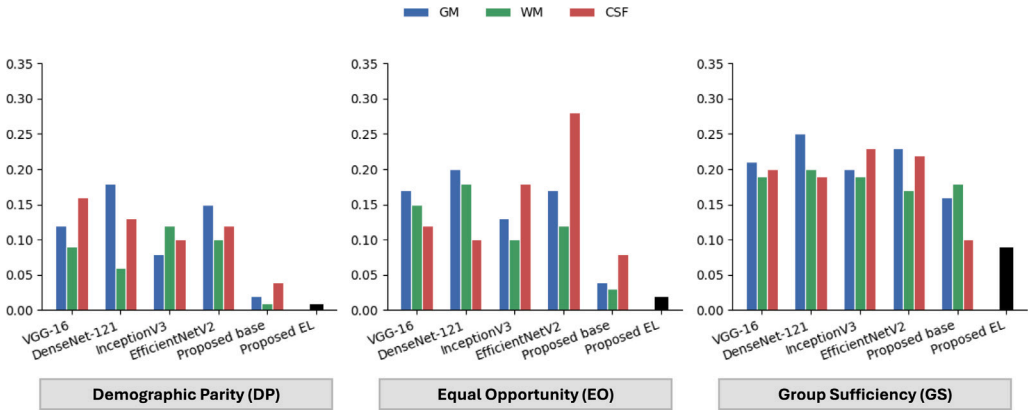

Fig. 11. Fairness evaluation using gender as the sensitive attribute. DP, EO, and GS are reported for tissue-specific models trained on GM, WM, and CSF. Lower values indicate improved fairness. The proposed ensemble consistently achieves reduced disparity across all fairness criteria.

pMCI subjects. In sMCI cases, attention patterns are relatively diffuse and fragmented, suggesting reliance on weak and spatially distributed cues. For instance, in the representative sMCI cases, highlighted brain regions include the rostral hippocampus (Greene et al., 2012), medial and lateral amygdala (Nelson et al., 2018), globus pallidus (Nelson et al., 2018), entorhinal cortex, and parahippocampal gyrus (van Hoesen et al., 2000), which remain relatively stable across follow-up scans. In contrast, pMCI cases show stronger and more spatially coherent activations across different regions, indicating more decisive feature attribution consistent with progression-related structural changes. The GM branch emphasizes broader cortical and associative regions, the WM branch highlights deep and periventricular structures, and the CSF branch captures ventricular-centric and sulcal patterns. The specific brain regions include, hippocampus (Greene et al., 2012), amygdala (Nelson et al., 2018), dorsolateral putamen (Reeves et al., 2010), fusiform gyrus (area 20) (van Hoesen et al., 2000), superior and middle temporal gyri (Castro et al., 2015; Ulloa et al., 2015), inferior temporal gyrus (Scheff et al., 2011), and parahippocampal regions, reflecting ongoing neurodegeneration. 3D surface-based renderings further reinforce these trends, revealing a more extensive and organized distribution of discriminative regions in pMCI relative to sMCI. These visualizations demonstrate that the framework captures complementary multi-tissue information and reflects the transition from diffuse attention in sMCI to anatomically structured focus in pMCI, providing clinically interpretable evidence to support disease progression assessment.

6.8. Experiment 8: External validation on the NACC dataset

Experiment 8 evaluates cross-cohort generalization of MRI-DeepTrust using the independent National Alzheimer's Coordinating Center (NACC) dataset (Beekly et al., 2007). Unlike ADNI, NACC includes multi-site acquisitions with greater variability in scanner characteristics, demographics, and diagnostic labeling, thereby providing a realistic assessment of robustness under domain shift. The evaluated subset consisted of 31 CN subjects, 29 CN individuals who later converted to AD, and 27 participants diagnosed with AD. The same preprocessing pipeline applied to ADNI, including MNI registration, skull stripping, normalization, and segmentation into GM, WM, and CSF, was used for NACC. Importantly, CNN configurations optimized on ADNI were kept fixed to ensure that performance reflects true external generalization rather than dataset specific retraining. Table 10 reports comparative results against baseline CNN architectures and ViT model (Dosovitskiy, 2020). Fig. 13 visualizes mAUC performance under both internal (ADNI→ ADNI) and external (ADNI→ NACC) evaluation settings. NACC experiments followed stratified 10-fold subject-level cross-validation to prevent slice-level leakage.

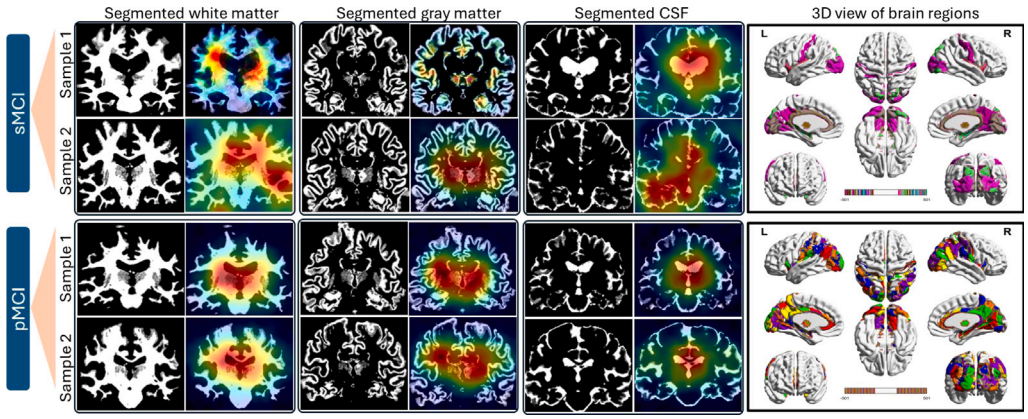

Fig. 12. Tissue-specific explainability maps for sMCI and pMCI samples. For each condition, representative GM, WM, and CSF regions are shown alongside their corresponding Grad-CAM activation. The upper block (sMCI) exhibits diffuse, spatially distributed, and comparatively unstable attention patterns, indicating weaker and less localized evidence. In contrast, the lower block (pMCI) demonstrates stronger, more focal, and anatomically structured activations, with increased involvement of medial temporal, lateral cortical, and periventricular regions. The rightmost panels provide surface-based visualization of group-level activation distributions, further highlighting the progressive shift from uncertain to more decisive network attention as cognitive impairment advances.

Table 10
Evaluating model generalizability across internal (ADNI→ADNI) and external (ADNI→NACC) datasets. Results illustrate the expected performance reduction under cross dataset testing and demonstrate that the proposed base models achieve consistently superior in distribution and out of distribution performance relative to competing architectures.

Model	BR	Train on ADNI, Test on NACC					Train on ADNI, Test on ADNI				
		Pre.	Rec.	F1	AUC	Acc.	Pre.	Rec.	F1	AUC	Acc.
VGG-16	GM	66 ± 2.8	64 ± 2.9	65 ± 2.9	66 ± 2.7	61 ± 3.0	74 ± 0.05	72 ± 0.04	73 ± 0.06	72 ± 0.05	69 ± 0.05
	WM	65 ± 3.0	63 ± 3.1	64 ± 3.1	66 ± 2.8	60 ± 3.2	71 ± 0.04	72 ± 0.07	70 ± 0.03	72 ± 0.04	68 ± 0.03
	CSF	69 ± 2.9	67 ± 2.8	68 ± 2.8	69 ± 2.6	64 ± 2.8	**83 ± 0.05**	**81 ± 0.03**	**82 ± 0.06**	**83 ± 0.04**	**79 ± 0.05**
DenseNet-121	GM	72 ± 2.8	70 ± 2.9	71 ± 2.8	72 ± 2.7	67 ± 2.7	75 ± 0.07	77 ± 0.03	76 ± 0.06	74 ± 0.04	71 ± 0.05
	WM	73 ± 2.6	71 ± 2.7	72 ± 2.3	73 ± 2.5	69 ± 2.6	**81 ± 0.05**	**79 ± 0.04**	**80 ± 0.07**	**81 ± 0.02**	**78 ± 0.06**
	CSF	69 ± 2.9	67 ± 3.0	68 ± 2.9	70 ± 2.7	64 ± 3.0	76 ± 0.02	74 ± 0.02	75 ± 0.02	74 ± 0.03	70 ± 0.02
InceptionV3	GM	76 ± 2.6	74 ± 2.7	75 ± 2.7	76 ± 2.5	71 ± 2.6	**83 ± 0.03**	**81 ± 0.02**	**83 ± 0.03**	**82 ± 0.03**	**80 ± 0.02**
	WM	73 ± 2.8	70 ± 2.9	71 ± 2.8	73 ± 2.6	68 ± 2.8	73 ± 0.05	74 ± 0.03	72 ± 0.07	74 ± 0.05	69 ± 0.04
	CSF	71 ± 2.9	69 ± 3.0	70 ± 3.0	72 ± 2.7	68 ± 2.9	73 ± 0.04	71 ± 0.05	72 ± 0.06	73 ± 0.03	70 ± 0.05
EfficientNetV2	GM	77 ± 2.5	75 ± 2.6	76 ± 2.6	77 ± 2.4	72 ± 2.6	80 ± 0.05	81 ± 0.07	81 ± 0.03	80 ± 0.04	78 ± 0.04
	WM	78 ± 2.4	76 ± 2.5	76 ± 2.5	77 ± 2.3	73 ± 2.5	**82 ± 0.02**	**81 ± 0.02**	**81 ± 0.03**	**80 ± 0.02**	**78 ± 0.03**
	CSF	75 ± 2.6	74 ± 2.6	74 ± 2.7	75 ± 2.4	70 ± 2.7	82 ± 0.04	80 ± 0.06	81 ± 0.04	81 ± 0.05	79 ± 0.03
ViT	GM	74 ± 2.6	72 ± 2.7	73 ± 2.7	74 ± 2.5	70 ± 2.8	81 ± 0.03	79 ± 0.03	80 ± 0.03	81 ± 0.03	78 ± 0.03
	WM	76 ± 2.7	74 ± 2.8	75 ± 2.8	75 ± 2.6	71 ± 2.8	**82 ± 0.03**	**80 ± 0.03**	**81 ± 0.03**	**81 ± 0.03**	**78 ± 0.03**
	CSF	74 ± 2.5	73 ± 2.5	73 ± 2.6	74 ± 2.6	70 ± 2.9	81 ± 0.03	80 ± 0.02	80 ± 0.02	80 ± 0.03	78 ± 0.03
Proposed base	GM	80 ± 2.4	78 ± 2.4	79 ± 2.5	80 ± 2.4	75 ± 2.5	82 ± 0.04	80 ± 0.03	81 ± 0.05	82 ± 0.03	82 ± 0.04
	WM	79 ± 2.3	77 ± 2.3	78 ± 2.3	79 ± 2.2	74 ± 2.4	**84 ± 0.05**	**81 ± 0.04**	**82 ± 0.05**	**83 ± 0.04**	**79 ± 0.03**
	CSF	78 ± 2.4	76 ± 2.5	77 ± 2.4	78 ± 2.3	73 ± 2.4	**83 ± 0.03**	**83 ± 0.02**	**82 ± 0.03**	**83 ± 0.05**	**80 ± 0.04**
Proposed EL	–	81 ± 2.1	79 ± 2.2	80 ± 2.1	81 ± 2.0	75 ± 2.2	**92 ± 0.02**	**89 ± 0.02**	**90 ± 0.02**	**89 ± 0.03**	**88 ± 0.03**

BR: Brain region corresponding to the segmented tissue (GM, WM, or CSF), Pre: Precision, Rec: Recall, F1: F1-score, AUC: Area Under ROC Curve, Acc: Accuracy.

The proposed framework maintains strong discriminative performance on NACC, with only a modest reduction in mAUC relative to ADNI. This decline is consistent with expected domain shift effects and does not indicate overfitting to ADNI specific characteristics. Notably, performance remains competitive across both convolutional and transformer based baselines, and no catastrophic degradation is observed. These findings demonstrate that MRI-DeepTrust preserves stable and reliable behavior across independent clinical cohorts, reinforcing its translational potential for deployment in heterogeneous real-world medical environments.

The proposed MRI-DeepTrust framework demonstrates consistent and meaningful improvements over both individual CNN backbones and conventional ensemble configurations. Compared to the strongest single tissue-specific model, the proposed ensemble achieves higher mAUC and accuracy while simultaneously maintaining improved stability under PGD-based adversarial perturbations. In contrast to baseline models that exhibit larger performance fluctuations across tissues and demographic subgroups, the integrated framework preserves robustness and gender fairness within narrow margins. Furthermore, external validation on the NACC cohort confirms that these gains are not restricted to the training distribution, but extend across independent clinical settings. These results indicate that the proposed framework offers not only improved predictive accuracy, but also more stable, equitable, and clinically reliable diagnostic behavior for distinguishing sMCI from pMCI. Unlike prior approaches that primarily optimize a single objective, the proposed framework achieves balanced gains across multiple trustworthiness dimensions within a unified, tissue-aware diagnostic pipeline, thereby strengthening its suitability for real-world clinical deployment.

6.9. Comparison and discussion

Table 11 summarizes recent state-of-the-art methods for AD diagnosis using brain structure segmentation, covering diverse modalities,

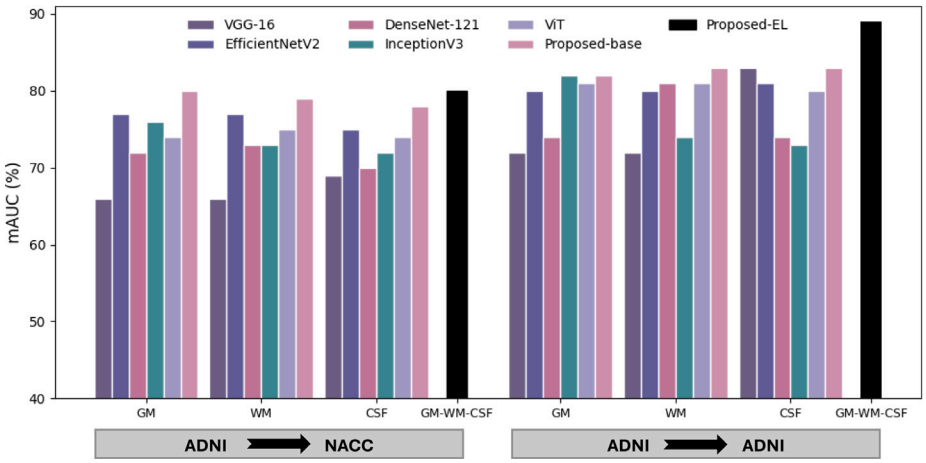

Fig. 13. Comparison of mAUC performance across tissue-specific models and the proposed framework under internal (ADNI→ ADNI) and external (ADNI→ NACC) evaluation settings. Performance is reported for GM, WM, CSF, and their combined representation (GM-WM-CSF). The proposed ensemble and its base variant consistently outperform competing models, highlighting the effectiveness of tissue-aware learning and ensemble fusion for robust cross-dataset generalization.

Table 11
Comparison of the state-of-the-art methods based on brain structure segmentation in the field of AD diagnosis using deep learning.

Study	NTD	Modality	Method	Data(size)	Pre.	Re.	F1	AUC	Acc.	Exp.	Fair.	Rob.
(Ortiz et al., 2016)	GM	MRI, PET	DBN	ADNI(818)	–	–	–	0.95	0.83	–	–	–
(Feng et al., 2019)	GM	MRI, PET	FSBi-LSTM	ADNI(397)	–	0.85	0.86	0.88	0.86	–	–	–
(Cui and Liu, 2018)	HC	MRI	3DCNN	ADNI(811)	–	0.74	–	0.79	0.75	–	–	–
(Ahmed et al., 2019)	HC-patches	MRI	EL-Model	ADNI (352)	0.86	0.86	0.86	–	0.86	–	–	–
(Shi et al., 2019)	CSF	MRI, PET	ML-Model	ADNI(202)	–	0.80	–	0.83	0.80	–	–	–
(Oh et al., 2022)	3D volumes	MRI	LEAR	ADNI(1538)	–	–	–	–	0.77	■	–	–
(Zhang et al., 2023)	3D volumes	MRI	DAUF	ADNI(576)	0.76	0.75	0.76	0.78	0.76	■	–	–
(Leonardsen et al., 2024)	3D volumes	MRI	CNN, LRP	ADNI(1708)	0.92	0.82	–	0.90	0.84	■	–	–
(Zhou et al., 2024)	ROI	MRI, PET	SGCN	ADNI(739)	–	0.63	–	0.73	0.70	■	–	–
(Vlontzou et al., 2025)	ROI	MRI, SNPs	SVM	ADNI(1463)	–	–	0.90	–	0.87	■	–	–
(Tong et al., 2025)	ROI	MRI	ML-Models	ADNI(1592)	–	–	–	–	0.76	■	■	–
(Ding et al., 2025)	ROI	MRI	GCN-KAN	ADNI(134)	–	–	0.60	0.64	0.62	■	–	–
Ours	GM, WM, CSF	MRI	MRI-DeepTrust	ADNI(592)	0.92	0.89	0.90	0.89	0.88	■	■	–
Ours*	**GM, WM, CSF**	**MRI**	**MRI-DeepTrust**	**ADNI(592)**	**0.85**	**0.83**	**0.83**	**0.82**	**0.81**	■	■	■

NTD: Neuroimaging Target Data; Pre:Precision; Re:Recall; F1:F1-score; Acc:Accuracy; Fair: Fairness; Exp: Explainability; Rob: Robustness; Ours: results before adversarial attacks; Ours*: results after adversarial attacks.

neuroimaging targets, and evaluation strategies. Earlier works primarily focused on predictive accuracy, with only a few more recent studies extending into interpretability or fairness. However, none explicitly evaluated adversarial robustness, leaving a significant gap in trustworthiness analysis. Studies such as (Ortiz et al., 2016; Cui and Liu, 2018) demonstrated the utility of ROI-based strategies, achieving accuracies in the range of 75 to 83%. Multimodal approaches, including (Feng et al., 2019; Shi et al., 2019), integrated MRI with PET and CSF biomarkers, reporting accuracy between 81 to 86% and AUC values up to 88%. Other contributions, such as (Leonardsen et al., 2024) with CNN-LRP explanations, further extended modeling approaches while beginning to incorporate explainability. Frameworks like (Ding et al., 2025; Zhang et al., 2023) addressed pMCI vs sMCI directly, with reported accuracies of 76 to 77%. More recent work, including (Vlontzou et al., 2025; Tong et al., 2025), highlighted interpretability and fairness, though robustness remained unexplored. Against this backdrop, our proposed MRI-DeepTrust framework distinguishes itself in two ways. First, for fair comparison with prior works presented in Table 11, we report results before adversarial attacks ("Ours"), which already achieve superior predictive performance, with 92% precision, 89% recall, 90% F1-score, 89% AUC, and 88% accuracy. These values surpass or match the best performing baselines that consider only one or two pillars of trustworthiness, such as explainability or fairness. Second, we additionally evaluate performance after adversarial attacks ("Ours*"), where performance remains consistently strong with 85% precision, 83% recall, 83% F1-score, 82% AUC, and 81% accuracy. This robustness evaluation, absent in all prior works, demonstrates that our model not only leads in accuracy under standard conditions but also preserves stability under adversarial perturbations. Taken together, these results highlight that while earlier studies incrementally advanced predictive accuracy, explainability, or fairness, none comprehensively addressed all three pillars of trustworthiness. The proposed framework uniquely achieves high predictive performance, offers model explainability, reduces fairness gaps, and demonstrates robustness through resilience to adversarial attacks, thereby establishing a new benchmark for trustworthy AI in AD diagnosis.

6.10. Practical impact of the proposed framework

The proposed MRI-DeepTrust framework offers several practical advantages that support its potential adoption in real clinical settings. First, the use of a two-dimensional tissue-based modeling strategy enables efficient training and inference while avoiding the high computational and memory demands associated with three-dimensional architectures, making the framework suitable for deployment in resource-constrained clinical environments. Second, the tissue-aware design, which independently models GM, WM, and CSF, aligns naturally with established neuroimaging workflows and enhances interpretability by allowing clinicians to reason about disease progression at the tissue level. Third, the integration of adversarial robustness analysis, fairness evaluation and visual explainability within a single unified pipeline reduces the need for post hoc validation steps and supports transparent

and reliable decision making. Collectively, these characteristics position the proposed framework as a clinically meaningful and practically deployable diagnostic support system rather than a purely experimental model.

7. Limitations and future research directions

This study contributes to the prediction of MCI using segmented brain MRI regions while embedding core principles of trustworthy AI. Nevertheless, several limitations should be acknowledged. First, fairness analysis in the present work was limited to gender as a protected attribute, as demographic information such as ethnicity and socioeconomic factors was not consistently available across datasets. Second, although robustness was evaluated under adversarial attack scenarios, further strengthening of resilience against a wider range of real-world perturbations and the integration of privacy-preserving mechanisms remain important for secure clinical deployment. In addition, while explainability techniques were incorporated to support interpretability, the uncertainty quantification was not explicitly modeled, which may limit confidence in high-stakes clinical decision support scenarios (Nimmy et al., 2025). Finally, both ADNI and NACC cohorts predominantly represent North American populations, and further validation across more geographically and ethnically diverse datasets would be essential to ensure broader demographic generalizability and equitable clinical deployment.

Building on these limitations, several promising directions for future research can be identified. First, fairness-aware modeling should be extended beyond gender to include additional protected attributes such as ethnicity, age, and socioeconomic status, alongside the development of mitigation strategies to ensure equitable diagnostic performance across diverse populations. Second, future work may explore more advanced robustness and privacy-aware learning strategies, including stronger adversarial defenses and secure model deployment mechanisms suitable for real clinical environments. Third, extending the proposed framework through multimodal integration represents a compelling avenue for further investigation. Combining MRI with complementary data sources such as longitudinal clinical records, genetic information, and demographic factors (Rahim et al., 2023a), as well as leveraging large language model based agents for clinical reasoning and decision support (Li et al., 2025), could enhance predictive accuracy and provide more comprehensive patient-level assessments. Finally, validation on larger and more diverse multi-center datasets, together with closer collaboration with healthcare institutions, will be essential to assess real-world clinical utility and support seamless integration into routine practice (Mahmud et al., 2024).

8. Conclusion

In this study, we presented MRI-DeepTrust, a trustworthy DL framework for the early diagnosis of AD via MCI. By integrating the three pillars of trustworthiness including, adversarial robustness, gender-based fairness and explainability, our approach addresses the long-standing limitations of black-box CNNs in medical imaging. Unlike conventional full-slice or ROI-based techniques, MRI-DeepTrust leverages segmentation brain tissues GM, WM, and CSF, extracting discriminative features through tissue-specific CNNs and fusing them in an ensemble classifier. This tissue-guided strategy enables interpretable region-level visualizations and achieves high predictive performance on the ADNI dataset, with a mean accuracy of 81%, AUC of 82%, F1-score of 83%, recall of 83%, and precision of 85%. Comparative evaluations with other state-of-the-art models further confirm the effectiveness of the proposed method, both in homogeneous and heterogeneous ensemble settings. Beyond performance, MRI-DeepTrust demonstrates that it is possible to build models that are accurate, equitable, and resilient to adversarial perturbations, aligning with the emerging standards of trustworthy AI in healthcare. Our fairness experiments across gender-balanced and imbalanced cohorts highlight the importance of addressing diagnostic disparities, while robustness analyses confirm the stability of predictions under adversarial noise. Importantly, the proposed framework explicitly incorporates a clinician-in-the-loop paradigm, in which medical domain experts assess visual explanations and model behavior, reinforcing clinical relevance and supporting informed decision making. MRI-DeepTrust establishes a foundation for advancing trustworthy DL in neuroimaging, bridging the gap between algorithmic performance and clinical applicability, and paving the way for more reliable, explainable, and fair diagnostic systems in real-world healthcare settings.

CRediT authorship contribution statement

Maria Bashir: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis. **Nasir Rahim:** Writing – review & editing, Methodology, Investigation, Data curation. **Shaker El-Sappagh:** Visualization, Validation, Methodology, Formal analysis. **Omar Amin El-Serafy:** Writing – review & editing, Validation, Methodology, Investigation. **Tamer Abuhmed:** Writing – review & editing, Validation, Supervision, Funding acquisition.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2021-NR058558), (Institute for Information & communications Technology Planning & Evaluation) (IITP) grant funded by the Korea government (MSIT) under the ICT Creative Consilience Program (IITP-2025-RS-2020-II201821) and AI Platform to Fully Adapt and Reflect Privacy-Policy Changes (No. 2022-0-00688).

Data availability

This study is based on the Alzheimer's Disease Neuroimaging Initiative dataset.

References

Abbas, S., Ahmed, F., Khan, W.A., Ahmad, M., Khan, M.A., Ghazal, T.M., 2025. Intelligent skin disease prediction system using transfer learning and explainable artificial intelligence. Sci. Rep. 15 (1), 1746.

Abuhmed, T., El-Sappagh, S., Alonso, J.M., 2021. Robust hybrid deep learning models for alzheimer's progression detection. Knowl.-Based Syst. 213, 106688.

Aderghal, K., Khvostikov, A., Krylov, A., Benois-Pineau, J., Afdel, K., Catheline, G., 2018. Classification of alzheimer disease on imaging modalities with deep CNNs using cross-modal transfer learning. In: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems. CBMS, IEEE, pp. 345–350.

Ahmed, S., Choi, K.Y., Lee, J.J., Kim, B.C., Kwon, G.-R., Lee, K.H., Jung, H.Y., 2019. Ensembles of patch-based classifiers for diagnosis of alzheimer diseases. IEEE Access 7, 73373–73383.

Al-bakri, F.H., Bejuri, W.M.Y.W., Al-Andoli, M.N., Ikram, R.R.R., Khor, H.M., Tahir, Z., Initiative, A.D.N., 2025. A meta-learning-based ensemble model for explainable alzheimer's disease diagnosis. Diagnostics 15 (13), 1642.

Albahri, A., Duhaim, A.M., Fadhel, M.A., Alnoor, A., Baqer, N.S., Alzubaidi, L., Albahri, O., Alamoodi, A., Bai, J., Salhi, A., et al., 2023. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Inf. Fusion.

Beekly, D.L., Ramos, E.M., Lee, W.W., Deitrich, W.D., Jacka, M.E., Wu, J., Hubbard, J.L., Koepsell, T.D., Morris, J.C., Kukull, W.A., et al., 2007. The national alzheimer's coordinating center (NACC) database: the uniform data set. Alzheimer Dis. Assoc. Disord. 21 (3), 249–258.

Castro, E., Ulloa, A., Plis, S.M., Turner, J.A., Calhoun, V.D., 2015. Generation of synthetic structural magnetic resonance images for deep learning pre-training. In: 2015 IEEE 12th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1057–1060.

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 839–847.

Choi, B.K., Madusanka, N., Choi, H.K., So, J.H., Kim, C.H., Park, H.G., Bhattacharjee, S., Prakash, D., 2020. Convolutional neural network-based MR image analysis for alzheimer's disease classification. Curr. Med. Imaging 16 (1), 27–35.

Choung, H., David, P., Ross, A., 2023. Trust in AI and its role in the acceptance of AI technologies. Int. J. Human–Computer Interact. 39 (9), 1727–1739.

Chunxia, Z., Jiangshe, Z., 2011. A survey of selective ensemble learning algorithms. Chinese J. Comput. 34 (8), 1399–1410.

Cui, R., Liu, M., 2018. Hippocampus analysis by combination of 3-D DenseNet and shapes for alzheimer's disease diagnosis. IEEE J. Biomed. Health Informat. 23 (5), 2099–2107.

Ding, T., Xiang, D., Schubert, K.E., Dong, L., 2025. GKAN: Explainable diagnosis of alzheimer's disease using graph neural network with Kolmogorov-arnold networks. arXiv preprint arXiv:2504.00946.

Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q., 2020. A survey on ensemble learning. Front. Comput. Sci. 14 (2), 241–258.

Dosovitskiy, A., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

El-Sappagh, S., Alonso-Moral, J.M., Abuhmed, T., Ali, F., Bugarín-Diz, A., 2023. Trustworthy artificial intelligence in alzheimer's disease: state of the art, opportunities, and challenges. Artif. Intell. Rev. 1–148.

Erdogmus, P., Kabakus, A.T., 2023. The promise of convolutional neural networks for the early diagnosis of the alzheimer's disease. Eng. Appl. Artif. Intell. 123, 106254.

Feng, C., Elazab, A., Yang, P., Wang, T., Zhou, F., Hu, H., Xiao, X., Lei, B., 2019. Deep learning framework for alzheimer's disease diagnosis via 3D-CNN and FSBi-LSTM. IEEE Access 7, 63605–63618.

Gotkowski, K., Gonzalez, C., Bucher, A., Mukhopadhyay, A., 2020. M3d-CAM: A PyTorch library to generate 3D data attention maps for medical deep learning. arXiv:arXiv:2007.00453.

Greene, S.J., Killiany, R.J., Initiative, A.D.N., 2012. Hippocampal subregions are differentially affected in the progression to alzheimer's disease. Anat. Rec.: Adv. Integr. Anat. Evol. Biology 295 (1), 132–140.

Hardt, M., Price, E., Srebro, N., 2016. Equality of opportunity in supervised learning. Adv. Neural Inf. Process. Syst. 29.

van Hoesen, G.W., Augustinack, J.C., Dierking, J., Redman, S.J., Thangavel, R., 2000. The parahippocampal gyrus in alzheimer's disease: clinical and preclinical neuroanatomical correlates. Ann. New York Acad. Sci. 911 (1), 254–274.

Jahn, A., 2020. FreeSurfer tutorial #6: Freeview & mdash; Andy's Brain Book 1.0 documentation.

Javed, H., El-Sappagh, S., Abuhmed, T., 2024. Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. Artif. Intell. Rev. 58 (1), 12.

Jenkinson, M., Pechaud, M., Smith, S., et al., 2005. BET2: MR-based estimation of brain, skull and scalp surfaces. In: Eleventh Annual Meeting of the Organization for Human Brain Mapping, vol. 17, (3), Toronto., p. 167.

Khan, Z.A., Waqar, M., Chaudhary, N.I., Raja, M.J.A.A., Khan, S., Khan, F.A., Chaudhary, I.I., Raja, M.A.Z., 2024. Fractional gradient optimized explainable convolutional neural network for alzheimer's disease diagnosis. Heliyon 10 (20).

Lee, S.B., 2025. Gradual poisoning of a chest x-ray convolutional neural network with an adversarial attack and AI explainability methods. Sci. Rep. 15 (1), 21779.

Leonardsen, E.H., Persson, K., Grødem, E., Dinsdale, N., Schellhorn, T., Roe, J.M., Vidal-Piñeiro, D., Sørensen, Ø., Kaufmann, T., Westman, E., et al., 2024. Constructing personalized characterizations of structural brain aberrations in patients with dementia using explainable artificial intelligence. NPJ Digit. Med. 7 (1), 110.

Li, R., Wang, X., Berlowitz, D., Mez, J., Lin, H., Yu, H., 2025. CARE-AD: a multi-agent large language model framework for alzheimer's disease prediction using longitudinal clinical notes. Npj Digit. Med. 8 (1), 541.

Mahmud, T., Barua, K., Habiba, S.U., Sharmen, N., Hossain, M.S., Andersson, K., 2024. An explainable ai paradigm for alzheimer's diagnosis using deep transfer learning. Diagnostics 14 (3), 345.

McCarthy, P., 2020. Fsleyes (version 0.34. 0). Zenodo.

Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N., 2022. Explainable artificial intelligence: a comprehensive review. Artif. Intell. Rev. 55 (5), 3503–3568.

Nelson, P.T., Abner, E.L., Patel, E., Anderson, S., Wilcock, D.M., Kryscio, R.J., Van Eldik, L.J., Jicha, G.A., Gal, Z., Nelson, R.S., et al., 2018. The amygdala as a locus of pathologic misfolding in neurodegenerative diseases. J. Neuropathol. Exp. Neurol. 77 (1), 2–20.

Nimmy, S.F., Hussain, O.K., Chakrabortty, R.K., Leshob, A., 2025. Quantifying the trustworthiness of explainable artificial intelligence outputs in uncertain decision-making scenarios. Eng. Appl. Artif. Intell. 141, 109678.

Oh, K., Yoon, J.S., Suk, H.-I., 2022. Learn-explain-reinforce: Counterfactual reasoning and its guidance to reinforce an alzheimer's disease diagnosis model. IEEE Trans. Pattern Anal. Mach. Intell..

Orouskhani, M., Zhu, C., Rostamian, S., Zadeh, F.S., Shafiei, M., Orouskhani, Y., 2022. Alzheimer's disease detection from structural MRI using conditional deep triplet network. Neurosci. Informat. 2 (4), 100066.

Ortiz, A., Munilla, J., Gorriz, J.M., Ramirez, J., 2016. Ensembles of deep learning architectures for the early diagnosis of the alzheimer's disease. Int. J. Neural Syst. 26 (07), 1650025.

Poloni, K.M., Ferrari, R.J., Initiative, A.D.N., et al., 2022. A deep ensemble hippocampal CNN model for brain age estimation applied to alzheimer's diagnosis. Expert Syst. Appl. 195, 116622.

Porsteinsson, A., Isaacson, R., Knox, S., Sabbagh, M., Rubino, I., 2021. Diagnosis of early alzheimer's disease: Clinical practice in 2021. J. Prev. Alzheimer's Dis. 8 (3), 371–386.

Rahim, N., Abuhmed, T., Mirjalili, S., El-Sappagh, S., Muhammad, K., 2023a. Time-series visual explainability for alzheimer's disease progression detection for smart healthcare. Alex. Eng. J. 82, 484–502.

Rahim, N., Ahmad, N., Ullah, W., Bedi, J., Jung, Y., 2025. Early progression detection from MCI to AD using multi-view MRI for enhanced assisted living. Image Vis. Comput. 157, 105491.

Rahim, N., El-Sappagh, S., Ali, S., Muhammad, K., Del Ser, J., Abuhmed, T., 2023b. Prediction of alzheimer's progression based on multimodal deep-learning-based fusion and visual explainability of time-series data. Inf. Fusion 92, 363–388.

Rahim, N., El-Sappagh, S., Rizk, H., El-serafy, O.A., Abuhmed, T., 2024. Information fusion-based Bayesian optimized heterogeneous deep ensemble model based on longitudinal neuroimaging data. Appl. Soft Comput. 162, 111749.

Reeves, S., Mehta, M., Howard, R., Grasby, P., Brown, R., 2010. The dopaminergic basis of cognitive and motor performance in alzheimer's disease. Neurobiol. Dis. 37 (2), 477–482.

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

Shi, Y., Suk, H.I., Gao, Y., Lee, S.W., Shen, D., 2019. Leveraging coupled interaction for multimodal alzheimer's disease diagnosis. IEEE Trans. Neural Networks Learn. Syst. 31 (1), 186–200.

Shui, C., Xu, G., Chen, Q., Li, J., Ling, C.X., Arbel, T., Wang, B., Gagné, C., 2022. On learning fairness and accuracy on multiple subgroups. Adv. Neural Inf. Process. Syst. 35, 34121–34135.

Tarzanagh, D.A., Hou, B., Tong, B., Long, Q., Shen, L., 2023. Fairness-aware class imbalanced learning on multiple subgroups. In: Uncertainty in Artificial Intelligence. PMLR, pp. 2123–2133.

Tayebi Arasteh, S., Ziller, A., Kuhl, C., Makowski, M., Nebelung, S., Braren, R., Rueckert, D., Truhn, D., Kaissis, G., 2024. Preserving fairness and diagnostic accuracy in private large-scale ai models for medical imaging. Commun. Med. 4 (1), 46.

Tong, B., Edwards, T., Yang, S., Hou, B., Tarzanagh, D.A., Urbanowicz, R.J., Moore, J.H., Ritchie, M.D., Davatzikos, C., Shen, L., 2025. Ensuring fairness in detecting mild cognitive impairment with MRI. In: AMIA Annual Symposium Proceedings, vol. 2024, p. 1119.

Ulloa, A., Plis, S., Erhardt, E., Calhoun, V., 2015. Synthetic structural magnetic resonance image generator improves deep learning prediction of schizophrenia. In: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing. MLSP, IEEE, pp. 1–6.

Valizadeh, G., Elahi, R., Hasankhani, Z., Rad, H.S., Shalbaf, A., 2025. Deep learning approaches for early prediction of conversion from MCI to AD using MRI and clinical data: A systematic review. Arch. Comput. Methods Eng. 32 (2), 1229–1298.

Vlontzou, M.E., Athanasiou, M., Dalakleidi, K.V., Skampardoni, I., Davatzikos, C., Nikita, K., 2025. A comprehensive interpretable machine learning framework for mild cognitive impairment and alzheimer's disease diagnosis. Sci. Rep. 15 (1), 8410.

Wang, D., Wang, L., Zhang, Z., Wang, D., Zhu, H., Gao, Y., Fan, X., Tian, F., 2021. "Brilliant AI doctor" in rural clinics: Challenges in AI-powered clinical decision support system deployment. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–18.

Zhang, Z., Gao, L., Jin, G., Li, P., Wang, J., et al., 2023. DAUF: A disease-related attentional unet framework for progressive and stable mild cognitive impairment identification. Comput. Biol. Med. 107401.

Zhang, H., Ni, M., Yang, Y., Xie, F., Wang, W., He, Y., Chen, W., Chen, Z., 2025. Patch-based interpretable deep learning framework for alzheimer's disease diagnosis using multimodal data. Biomed. Signal Process. Control. 100, 107085.

Zhou, H., He, L., Chen, B.Y., Shen, L., Zhang, Y., 2024. Multi-modal diagnosis of alzheimer's disease using interpretable graph convolutional networks. IEEE Trans. Med. Imaging.