Journal of Medical and Health Studies (JMHS)
ISSN: 2710-1452
DOI: 10.32996/jmhs
Journal Homepage: www.al-kindipublisher.com/index.php/jmhs

JMHS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Explainable and Trustworthy Deep Learning for MRI-Based Auxiliary Diagnosis of Alzheimer's Disease

Mingxuan Zhang
Sino-British Joint College, China Medical University, Shenyang 110000, China

| ABSTRACT

Alzheimer's disease (AD) and its prodromal stage, mild cognitive impairment (MCI), are characterized by a long preclinical phase. Early identification is crucial for timely intervention and prognostic assessment. Structural MRI has become an important imaging modality for auxiliary diagnosis of AD due to its non-invasive nature, repeatability, and strong clinical accessibility. Deep learning has made significant progress in AD/MCI/CN classification, staging, differential diagnosis, and clinical risk prediction. However, the "black-box" nature of these models, insufficient cross-center generalization, and lack of trustworthy explanations severely limit their clinical translation. This review focuses on four main threads—"model pipeline—explanation methods—explanation evaluation—generalization and deployment"—to systematically summarize representative directions of MRI-based deep learning in AD-related tasks. It also outlines the primary pathways of explainable artificial intelligence (XAI) (saliency attribution, counterfactual explanation, and inherently interpretable architectures) and the key dimensions of trustworthiness assessment (fidelity, stability, anatomical/pathological plausibility, and human-factor usability). Methodological risks such as real-world data challenges, domain generalization, cross-scanner harmonization, and data leakage are discussed in depth. Key contributions include: (1) emphasizing that "explainability ≠ heatmap visualization" and proposing an explanation evaluation framework centered on fidelity and stability; (2) integrating risks such as shortcut learning, data leakage, and cross-scanner differences into a unified discussion of explainability and generalization; and (3) summarizing clinically translatable pathways from the perspectives of real-world application and privacy compliance. MRI-based deep learning is shifting from "pursuing accuracy" toward "trustworthy explanation + cross-center generalization + clinical usability." Promising future directions include standardized explanation evaluation, domain generalization and harmonization techniques tailored to real-world distribution shifts, and human-interpretable explanations validated through multi-center real-world studies.

| KEYWORDS

Alzheimer's disease; structural MRI; deep learning; explainable artificial intelligence; trustworthiness assessment; generalization and real-world application

| ARTICLE INFORMATION

ACCEPTED: 01 March 2026 **PUBLISHED:** 31 March 2026 **DOI:** 10.32996/jmhs.2026.7.5.13

I. Introduction

The clinical value of early screening for Alzheimer's disease is well established. However, real-world clinical implementation requires a robust evidence chain that is "explainable, verifiable, and transferable." On one hand, existing studies frequently report high performance on research cohorts such as ADNI. On the other hand, when models are transferred to routine clinical data, performance often declines substantially. Moreover, models are easily influenced by non-pathological confounding factors such as image quality and contrast agents, leading to typical Clever Hans or shortcut learning phenomena. These issues can mislead both performance evaluation and interpretation results [1]. Furthermore, even within the same public dataset, preprocessing steps can inadvertently induce shortcut learning. For example, skull-stripping may introduce contour cues that serve as unintended "hints" for the model, allowing it to maintain high metrics without relying on gray-white matter texture features. Consequently, the resulting explanations may appear plausible yet deviate significantly from the underlying pathological

mechanisms [2]. In addition, inappropriate data partitioning strategies applied to longitudinal MRI or repeated scan data can cause data leakage or identity confounding, resulting in artificially inflated performance and compromised explanation trustworthiness [3]. Therefore, a comprehensive review is needed that adopts an integrated perspective across "model—explanation—evaluation—deployment" to summarize key advances and existing deficiencies in the field.

II. MRI Deep Learning Model Pipelines and Task Expansion

Existing MRI-based deep learning approaches for AD diagnosis have generally evolved from conventional CNNs toward architectures that place greater emphasis on structural constraints and cross-domain robustness. Clinical tasks have also expanded beyond the simple "AD vs CN" binary classification to scenarios that more closely reflect real-world clinical practice:

(1) Differential diagnosis (e.g., AD vs FTD) that better aligns with outpatient differential needs. Studies on structural MRI-based differential diagnosis have proposed interpretable intermediate representations such as "grading maps" or coordinate maps, which enhance clinical communicability and have undergone external validation [4].

(2) Preclinical or surrogate-label prediction: Using MRI combined with demographic and scale-based proxy information to predict amyloid-related status. This serves as one promising non-invasive early screening pathway, with specific attention to handling missing labels, class imbalance, and real-world transferability [5].

(3) Real-world differential diagnosis (multimodal, hospital-derived data): Performing multi-disease differentiation on hospital EHR imaging data. These studies emphasize the need for models to avoid treating confounding factors as shortcuts and employ attention mechanisms or explanation techniques to analyze the regions the model focuses on [6].

(4) Cross-center generalization: Domain generalization frameworks attempt to steer models toward focusing on "disease-relevant regions" and validate their extrapolation ability across multiple cohorts [7]. Cross-scanner harmonization studies reveal that current methods still exhibit limitations in longitudinal and multi-device scenarios, highlighting the need for more systematic evaluation and standardization [8].

III. Classification of Explainable Methods

Explainable methods can be broadly categorized into three types:

(1) Post-hoc attribution: Methods such as Grad-CAM and SHAP, which generate voxel-wise heatmaps or feature importance scores. Their advantages include simplicity and model-agnostic compatibility; however, the explanations can be sensitive to perturbations, and different methods may localize saliency to extra-brain regions or produce results inconsistent with known anatomical patterns, necessitating systematic comparative evaluation [9].

(2) Counterfactual explanation: By generating samples or features that involve "minimal changes sufficient to flip the prediction," these methods answer the question "what changes would affect the classification?" Diffusion-model-driven counterfactual approaches have improved upon traditional generative priors in terms of specificity, plausibility, and fidelity, leading to higher-quality counterfactual explanation frameworks [10].

(3) Intrinsic interpretability: Interpretable mechanisms are embedded directly into the model architecture (e.g., attention or constraint modules, interpretable intermediate representations, or visualization-prior alignment). For instance, in domain generalization, aligning class-specific attention with a unified saliency prior forces the model to "look in the right places" during training, thereby improving cross-cohort robustness and consistency with pathological evidence [7]. In differential diagnosis tasks, interpretable intermediate atlases have been used to enhance readability [4].

IV. Explanation and Trustworthiness Evaluation

Merely displaying heatmaps is insufficient to support clinical trustworthiness. A more rigorous evaluation should at minimum include the following dimensions: (1) Fidelity: Whether the explanation faithfully reflects the model's actual decision basis, which can be tested through perturbation, occlusion, or minimal counterfactual changes [10]. (2) Stability/Robustness: Whether the explanation remains consistent under small perturbations or randomness. This can be examined using "explanation robustness assessment" or "unified explanation" approaches, evaluating stability from the perspectives of necessity and sufficiency [11]. (3) Anatomical plausibility: Whether the explanation localizes to credible brain regions. Systematic comparisons of multiple saliency methods have shown significant differences among attribution techniques in anatomical localization and robustness; some methods may assign saliency to extra-cerebral areas and should be used with caution [9]. (4) Pathological plausibility: Mapping large-scale histological tau information onto a common MRI space to validate imaging biomarkers, providing a critical pathway for moving "explanations toward biological truth" [12].

In addition, trustworthiness evaluation must be tightly coupled with data pipeline auditing. Assessments of clinical data warehouses have shown that confounding factors such as image quality and contrast agents can substantially inflate apparent performance; after debiasing, model performance may drop dramatically [1]. Skull-stripping can induce shortcut learning [2], and inappropriate partitioning of longitudinal data can introduce leakage and identity confounding [3]. Systematic comparative studies have further conducted quantitative comparisons between deep learning models and traditional models in terms of explanations, combining internal and external cohort validation with VBM controls to discuss differences in interpretations, thereby providing a paradigm for constructing comprehensive evaluation frameworks [13].

As shown in Table 1, this review summarizes the main explainable pathways and trustworthiness assessment dimensions in a corresponding manner, providing a horizontal comparison of relevant methods across four aspects: typical output, primary advantages, potential risks/misuses, and recommended evaluation dimensions.

Table 1 Correspondence between Explainable Methods and Trustworthiness Assessment Dimensions

Method Category	Typical Output	Primary Advantages	Typical Risks/Misuses	Recommended Assessment Dimensions (see Section 4)	Representative References
Post-hoc Attribution (Grad-CAM/SHAP, etc.)	Heatmaps / feature importance	Easy integration, low cost	Unstable explanations; possible explanation bias / shortcut reliance	Fidelity, Stability, Anatomical plausibility	[1][2][9][13]
Counterfactual Explanation	"Minimal change" samples / features	Close to causal questioning; quantifiable	Biologically implausible counterfactuals; unstable generation quality	Fidelity, Feasibility constraints, Stability	[10]
Intrinsic Interpretable Structures (ROI / intermediate representations / constrained attention)	Structured attention regions / intermediate atlases	Easier clinical communication	May appear "superficially explanatory"; requires external validation	Anatomical/Pathological plausibility, External validation	[4][7]
Generalization and Auditing (DG / harmonization / confounding control)	Cross-center performance / bias analysis	Oriented toward deployment; interpretable sources of bias	Experimentally complex; inconsistent reporting	External validation, Confounding factor auditing	[1][7][8]
Pathological/Biological Truth Validation	Imaging–pathology alignment	Enhances medical	Expensive data; alignment	Pathological plausibility, Completeness	[12]

Method Category	Typical Output	Primary Advantages	Typical Risks/Misuses	Recommended Assessment Dimensions (see Section 4)	Representative References
	evidence	credibility	difficulties	of evidence chain	

V. Generalization and Real-World Deployment

The research-to-clinic gap remains one of the most critical barriers to translation. Routine clinical MRI data exhibit substantial variations in population demographics, scanner hardware, imaging protocols, acquisition workflows, and annotation standards, all of which induce distribution shifts. Evaluations on clinical data warehouses have shown that models are prone to "shortcut exploitation" by non-pathological confounding factors such as image quality and contrast agents. Performance reported solely on research cohorts is therefore insufficient to demonstrate clinical utility [1].

On the technical front, several promising pathways have emerged:

(1) Domain generalization aims to learn more robust representations that can adapt to cross-dataset variations. By incorporating explanation or saliency priors for alignment, these methods constrain the model to focus on disease-relevant regions, thereby enhancing extrapolation capability [7].

(2) Cross-scanner harmonization seeks to mitigate differences arising from scanners and protocols; however, systematic comparisons indicate that current approaches still struggle to fully resolve consistency issues in longitudinal and multi-device settings [8].

(3) Real-world differential diagnosis models trained on hospital EHR imaging data and tested at external sites provide validation closer to actual clinical practice. Through attention mechanisms and explanation analyses, these studies have identified potential biomarkers such as subcortical structures that the model attends to [6].

(4) In the context of privacy and regulatory compliance, multi-center collaborative training combined with interpretable and communicable outputs will represent a long-term direction (not elaborated in this article).

In summary, a clinically viable model should simultaneously satisfy the following requirements: rigorous external validation, auditing of confounding factors, explainable sensitivity to preprocessing steps, as well as stability and fidelity of the explanation outputs.

VI. Conclusion

MRI-based deep learning has established a relatively mature spectrum of models for AD-related diagnosis and prediction. However, future breakthroughs will not lie in simply stacking increasingly complex networks, but rather in constructing a solid evidence chain of "trustworthy explanation—cross-center generalization—clinical usability." Future research is recommended to: Establish unified explanation evaluation protocols and systematically report fidelity and stability metrics; Develop domain generalization and harmonization strategies specifically tailored to real-world distribution shifts; Conduct rigorous external validation on multi-center real-world datasets and produce structured, communicable, and verifiable explanations.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Bottani, S., Burgos, N., Maire, A., Saracino, D., Ströer, S., Dormont, D., Colliot, O., & Alzheimer's Disease Neuroimaging Initiative. (2023). Evaluation of MRI-based machine learning approaches for computer-aided diagnosis of dementia in a clinical data warehouse. *Medical Image Analysis*, 89, 102903. DOI: 10.1016/j.media.2023.102903.
[2] Tinauer, C., Sackl, M., Stollberger, R., Schmidt, R., Ropele, S., & Langkammer, C. (2025). Skull-stripping induces shortcut learning in MRI-based Alzheimer's disease classification. *Insights into Imaging*, 16(1), 283. DOI: 10.1186/s13244-025-02158-4.

[3] Rumala, D. J. (2023). How you split matters: Data leakage and subject characteristics studies in longitudinal brain MRI analysis. In *Lecture Notes in Computer Science: Vol. 14242* (pp. 235–245). Springer. DOI: 10.1007/978-3-031-45249-9_23.
[4] Nguyen, H.-D., Clément, M., Planche, V., Mansencal, B., & Coupé, P. (2023). *Deep Grading for MRI-based Differential Diagnosis of Alzheimer's Disease and Frontotemporal Dementia*.
[5] Hwang, U., Kim, S.-W., Jung, D., Kim, S., Lee, H., Seo, S. W., Seong, J.-K., Yoon, S., & Alzheimer's Disease Neuroimaging Initiative. (2023). Real-world prediction of preclinical Alzheimer's disease with a deep generative model. *Artificial Intelligence in Medicine*, 144, 102654. DOI: 10.1016/j.artmed.2023.102654.
[6] Leming, M., & Im, H. (2025). Differential dementia detection from multimodal brain images in a real-world dataset. *Alzheimer's & Dementia*, 21(7), e70362. DOI: 10.1002/alz.70362.
[7] Lteif, D., Sreerama, S., Bargal, S. A., Plummer, B. A., Au, R., & Kolachalama, V. B. (2024). Disease-driven domain generalization for neuroimaging-based assessment of Alzheimer's disease. *Human Brain Mapping*, 45(8), e26707. DOI: 10.1002/hbm.26707.
[8] Gebre, R. K., Senjem, M. L., Raghavan, S., Schwarz, C. G., Gunter, J. L., Hofrenning, E. I., ... Jack, C. R., Jr. (2023). Cross-scanner harmonization methods for structural MRI may need further work: A comparison study. *NeuroImage*, 269, 119912. DOI: 10.1016/j.neuroimage.2023.119912.
[9] Guo, K. H., Chaudhari, N. N., Jafar, T., Chowdhury, N. F., Bogdan, P., Irimia, A., ... & the Alzheimer's Disease Neuroimaging Initiative. (2024). Anatomic interpretability in neuroimage deep learning: Saliency approaches for typical aging and traumatic brain injury. *Neuroinformatics*, 22(4), 591–606. DOI: 10.1007/s12021-024-09694-2.
[10] Bedel, H. A., & Çukur, T. (2023). *DreaMR: Diffusion-driven Counterfactual Explanation for Functional MRI* [Preprint] (arXiv:2307.09547). DOI: 10.48550/arXiv.2307.09547.
[11] 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. *Scientific Reports*, 15(1), 8410. DOI: 10.1038/s41598-025-92577-6.
[12] Ushizima, D., Chen, Y., Alegro, M., Ovando, D., Eser, R., Lee, W. H., ... & the Alzheimer's Disease Neuroimaging Initiative. (2022). Deep learning for Alzheimer's disease: Mapping large-scale histological tau protein for neuroimaging biomarker validation. *NeuroImage*, 248, 118790. DOI: 10.1016/j.neuroimage.2021.118790.
[13] Bloch, L., & Friedrich, C. M. (2024). Systematic comparison of 3D deep learning and classical machine learning explanations for Alzheimer's disease detection. *Computers in Biology and Medicine*, 170, 108029. DOI: 10.1016/j.compbiomed.2024.108029.