Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 15 July 2026
Revised 18 Aug 2026
Accepted 01 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (19s): 1549–1558

KEYWORDS:
Alzheimer's disease; cognitive deviation; cross-dataset validation; deep learning; domain shift; explainable AI; Grad-CAM; magnetic resonance imaging; multimodal fusion; OASIS.

A Multisource Explainable AI Framework for Cognitive Deviation Assessment in Aging Brains: A Cross-Dataset Generalization Study

Dipali Prakash Patil[1,*], Chandrani Singh[2]

[1]Research Scholar, MCA, Sinhgad Institute of Management, Savitribai Phule Pune University, Pune, India
[2]Research Guide, MCA, Sinhgad Institute of Management, Savitribai Phule Pune University, Pune, India
[*]**Corresponding author:** dipalipatil.siom@sinhgad.edu; deepali7519@gmail.com
E-mail: directormca_siom@sinhgad.edu
ORCID: D. P. Patil 0000-0002-1133-6082; C. Singh 0000-0003-0970-2239

ABSTRACT: Deep-learning systems for cognitive assessment often report strong within-dataset performance without establishing whether the learned representation transfers to independent cohorts. This study evaluates the Hybrid Explainable Cognitive Deviation Tracking (HE-CDT) framework as an evidence-auditable research prototype integrating structural MRI staging, questionnaire-derived behavioral risk, model confidence reporting, explainability, and cross-dataset validation. Six MRI configurations—a custom CNN, MobileNetV2, EfficientNetB0, DenseNet121 baseline, optimized DenseNet121, and a soft-voting ensemble—were compared on a four-class development cohort. The optimized DenseNet121 achieved 79.26% accuracy, 0.8282 macro F1, and 0.9350 ROC-AUC on the internal holdout, while the ensemble achieved the highest internal AUC of 0.9477. Without retraining, optimized DenseNet121 achieved 37.84% accuracy, 0.2705 macro F1, and 0.5441 AUC on the independent OASIS cohort, exposing substantial domain shift.

A methodological audit was additionally performed for dataset provenance, split leakage, class-level artifacts, and fusion-score design. Patient/group-aware splitting is documented where identifiers are available, but augmentation-family isolation could not be fully verified from the recovered manifests. Likewise, a standalone confusion matrix and stylized ROC artifact could not be reconciled with the verified aggregate results; they are therefore excluded from final quantitative claims rather than reconstructed. To remove a conceptual bias in the legacy fusion rule, confidence is separated from deviation severity and retained as a reliability indicator, while a probability-weighted MRI severity formulation is proposed for future calibrated fusion. The study demonstrates that transparent negative external-validation findings and evidence-integrity auditing are essential for credible cognitive-imaging AI research.

1. INTRODUCTION

Cognitive decline in aging is heterogeneous and is reflected incompletely by any single data source. Structural magnetic resonance imaging (MRI) can capture anatomical patterns associated with neurodegeneration, whereas behavioral and functional assessments describe changes that are not directly encoded in an image. Contemporary reviews therefore increasingly emphasize multimodal evidence, transparent evaluation, and external validation rather than relying on a single high internal accuracy value [1], [3]–[5].

Deep neural networks have demonstrated strong discrimination for Alzheimer's disease (AD) and related cognitive-stage tasks, but the reported performance varies considerably with task definition, preprocessing, cohort composition, and validation

protocol [2], [8], [22], [23]. Reproducibility studies have also shown that leakage, incomplete split reporting, and unavailable code or prediction outputs can inflate apparent performance or make results difficult to audit [2], [18]. These issues are especially important in slice-based or augmented imaging datasets, where related images can inadvertently appear in more than one evaluation partition if split lineage is not controlled.

A second challenge is domain shift. External radiology studies commonly report degradation when a model is evaluated on data from a different center, scanner, protocol, or population [12]–[14]. MRI can be particularly sensitive to scanner and acquisition variability, making independent-cohort testing necessary before an internal benchmark can be interpreted as evidence of generalization [9], [14]. Explainability and uncertainty are additional requirements: Grad-CAM can expose local image regions that contribute to a CNN decision, while calibration and uncertainty analysis can indicate whether probability estimates are trustworthy [11], [16], [17], [28]. Neither, however, converts an experimental model into a clinically validated diagnostic system.

This work evaluates the Hybrid Explainable Cognitive Deviation Tracking (HE-CDT) framework under that stricter evidence standard. The framework is a *late, score-level multisource* research prototype rather than an end-to-end multimodal neural network. Its MRI branch performs comparative four-class staging; its questionnaire branch produces a normalized Questionnaire Risk Score (QRS); and its explainability branch generates Grad-CAM outputs for the selected MRI model. A cross-dataset OASIS experiment evaluates transfer without retraining. In addition to model metrics, this paper audits dataset lineage, split leakage, class-level evidence, and the mathematical role of confidence in the fused score.

The contributions are fourfold. First, six MRI configurations are compared using the same reported internal benchmark. Second, external OASIS results are reported separately and without averaging them into internal metrics. Third, a dataset and evidence-integrity audit identifies what is reproducibly supported and what remains unresolved. Fourth, the legacy fusion rule is reformulated so that high classifier confidence does not itself increase cognitive-deviation severity. This positioning follows recent calls for reproducible, externally validated, uncertainty-aware neuroimaging AI [5], [12], [19].

2. RELATED WORK

2.1 MRI and Multisource Cognitive Assessment

CNN-based MRI classification has progressed from task-specific architectures to transfer learning, multimodal modeling, and disease-progression prediction [2], [22], [23], [25]. Qiu *et al.* demonstrated that flexible combinations of clinical information, neuroimaging, and functional assessment can support dementia assessment [3]. El-Sappagh *et al.* likewise combined heterogeneous variables with explainable modeling for AD-related prediction [24]. More recent reviews conclude that multimodal fusion can improve representation breadth, but heterogeneity, missing modalities, and limited external validation remain major barriers [4], [6], [7].

2.2 Generalization, Reproducibility, and Explainability

Wen *et al.* highlighted leakage and reproducibility problems in MRI-based AD classification and showed that models can generalize differently across populations [2]. In a broader radiologic review, Yu *et al.* found external performance reductions in most externally validated deep-learning studies [12]. Domain-generalization reviews similarly identify acquisition, scanner, and population shifts as persistent threats [13], [15]. Disease-driven domain generalization for AD neuroimaging has therefore become an active research direction [9].

Explainability methods such as Grad-CAM provide local attribution without establishing anatomical causality [10], [11], [28]. Confidence values also require caution because modern neural networks can be miscalibrated [16], [29]. Recent methodological reviews of neuroimaging AI continue to identify calibration, leakage control, external validation, and reporting quality as unresolved translational requirements [19].

3. MATERIALS AND METHODS

3.1 HE-CDT Architecture

Figure 1 shows the revised evidence-auditable architecture. MRI images are resized and normalized before comparative CNN inference. The selected model outputs a four-class probability vector. Questionnaire items from the Memory, Daily Living, and Behavioral domains are independently converted into normalized risk scores. Fusion occurs only after the MRI and

questionnaire branches have produced auditable component values. Grad-CAM is attached to the MRI model and does not modify the prediction.

3.2 Dataset Provenance and Leakage Audit

The primary development cohort is the project-delivered augmented Alzheimer MRI Dataset (Dataset A). The audited modeling pool contains 4,000 images, balanced to 1,000 images in each of four classes: Non Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The reported split is 2,800/600/600 for training, validation, and holdout testing. A separate source-pool audit records substantial pre-balancing class skew; however, the recovered evidence contains two different source-count conventions and does not fully document their lineage. The independent Dataset B contains 1,344 OASIS-derived images distributed as 339 Non Demented, 741 Very Mild, 165 Mild, and 99 Moderate.

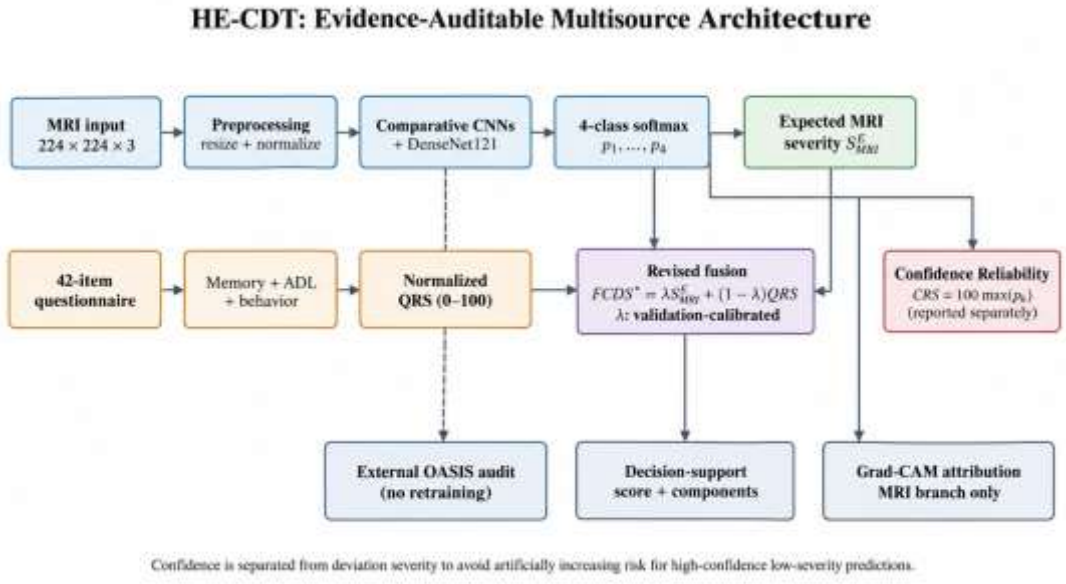


Figure 1. Revised HE-CDT architecture. MRI probability-derived severity and questionnaire risk contribute to the deviation score, while classifier confidence is retained as a separate reliability quantity. External OASIS evaluation and Grad-CAM remain independent audit paths.

OASIS is kept outside model development and is used only for external evaluation. OASIS itself is a public neuroimaging resource with established cross-sectional and longitudinal MRI releases [20], [21]. Table 1 records the leakage-control status conservatively. Participant/group identifiers are used where available to reduce related-sample leakage. However, the recovered split manifests do not prove that every augmented derivative of an original image remained within a single partition. Consequently, this paper does not claim a fully verified augmentation-family split.

Table 1. Dataset provenance and leakage-control audit

Source	Audited role	Working profile	Partition	Leakage control supported	Unresolved item
Dataset A	Development/comparison	4,000 balanced images; 4 classes	2,800 / 600 / 600	Group-aware logic where identifiers exist; fixed holdout	Exact upstream release/license and augmented-sibling isolation not fully recovered
OASIS Dataset B	External validation	1,344 images; natural imbalance	External only	No retraining on Dataset B	Mapping from public OASIS releases to derived four-class image folders requires manifest-level documentation
Questionnaire records	Behavioral/QRS component	42-item instrument; 37 stored participants in analytic DB	Separate research records	Scoring domains auditable	No matched prospective MRI–questionnaire clinical cohort for calibrated fusion

3.3 MRI Models and Preprocessing

The common input representation is 224 × 224 × 3. The verified implementation supports direct resizing and intensity normalization; N4 bias correction and MNI152 registration are not part of the active path. The comparison includes a Custom CNN, MobileNetV2 [26], EfficientNetB0 [27], DenseNet121 [25], an optimized DenseNet121 configuration, and a post-training soft-voting ensemble. The selected Phase-4 DenseNet121 is the strongest verified internal single model by accuracy and macro F1. Only configuration details supported by the project evidence are retained; conflicting training metadata are not promoted into the manuscript.

For class probabilities $\mathbf{p} = [p_1, \ldots, p_C]$, the predicted stage is

$$\hat{y} = \arg\max_k p_k \tag{1}$$

3.4 Questionnaire Risk Score

The 42-item research instrument contains demographic/context items plus Memory, Daily Living, Behavioral, Digital Readiness, and AI Acceptance sections. Only the Memory, Daily Living, and Behavioral domains contribute to QRS. Each scored item is mapped to 0–3 and each seven-item domain is normalized to 0–100. The aggregate score is

$$S_{\mathrm{QRS}} = \frac{S_{\mathrm{mem}} + S_{\mathrm{ADL}} + S_{\mathrm{beh}}}{3} \tag{2}$$

Digital-readiness and AI-acceptance scores remain contextual rather than cognitive-risk terms.

3.5 Confidence-Safe FCDS Reformulation

The implemented legacy prototype uses a mapped MRI stage score, QRS, and a positive confidence term. A conceptual weakness of that rule is that a highly confident low-severity prediction can still increase the fused deviation value. For publication-facing analysis, confidence is therefore separated from severity.

Let the stage anchors be $\mathbf{s} = [10, 35, 65, 90]$. Instead of using only the winning class, an expected MRI severity is defined as

$$S^E_{\mathrm{MRI}} = \sum_{k=1}^{4} p_k s_k \tag{3}$$

A revised fused score is

$$\mathrm{FCDS}^* = \lambda S^E_{\mathrm{MRI}} + (1-\lambda) S_{\mathrm{QRS}}, \quad 0 \le \lambda \le 1 \tag{4}$$

while the confidence reliability score is reported separately:

$$\mathrm{CRS} = 100 \max_k p_k \tag{5}$$

Crucially, λ is not assigned an invented optimum in this paper. It should be estimated using a matched validation cohort with an independent cognitive outcome. The existing model-performance tables therefore evaluate MRI staging, not a claimed performance gain from FCDS*.

3.6 Grad-CAM Explainability

For the selected DenseNet121 model, the project implementation attaches Grad-CAM to the final convolutional feature block documented as conv5_block16_concat. For target class c,

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \tag{6}$$

$$L^c_{\mathrm{Grad-CAM}} = \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{7}$$

The evidence package contains a representative original-MRI/heatmap/overlay triplet. Because no atlas registration or voxel-level pathological ground truth is available, the maps are interpreted as local model attribution rather than anatomical segmentation or causal biomarker evidence [10], [11], [28].

3.7 Evaluation and Statistical-Validation Policy

Primary reported metrics are accuracy, macro precision, macro recall, macro F1, and multiclass ROC-AUC. Internal and external metrics remain separate. A formal statistical protocol was also audited. Pairwise class-decision comparisons would require matched sample predictions for McNemar testing [32]; AUC comparison would require continuous class probabilities and, where appropriate, DeLong or resampling-based inference [31]; calibration would require probabilities for Brier score, reliability diagrams, or expected calibration error [29], [30].

Those raw vectors were not recovered in the supplied final evidence package. Therefore, confidence intervals, McNemar *p*-values, DeLong tests, Brier scores, and ECE are *not fabricated from aggregate metrics*. Numerical differences are reported as descriptive findings only. This follows reproducibility guidance that every inferential quantity should be traceable to sample-level outputs [18], [19].

4. RESULTS AND DISCUSSION

4.1 Internal Model Comparison

Table 2 reports the verified internal holdout results. Optimized DenseNet121 achieved the highest single-model accuracy (0.7926), precision (0.8259), and macro F1 (0.8282). The soft-voting ensemble achieved the highest AUC (0.9477) and slightly higher recall (0.8314), illustrating that discrimination and final class assignment do not produce identical model rankings.

Table 2. Verified internal Dataset A holdout performance

Model	Accuracy	Precision	Recall	Macro F1	AUC
Custom CNN	0.3816	0.2676	0.3309	0.2861	0.6708
MobileNetV2	0.5773	0.5838	0.6162	0.5968	0.8275
EfficientNetB0	0.5793	0.5872	0.6333	0.6053	0.8365
DenseNet121 baseline	0.6321	0.6318	0.6914	0.6532	0.8489
DenseNet121 Phase 4	**0.7926**	**0.8259**	0.8306	**0.8282**	0.9350
Soft-voting ensemble	0.7906	0.8225	**0.8314**	0.8257	**0.9477**

Relative to the baseline DenseNet121, Phase 4 improves internal accuracy by 0.1605 and macro F1 by 0.1750. Because sample-level paired predictions are unavailable, the small difference between Phase 4 and the ensemble is not labeled statistically significant.

4.2 External OASIS Generalization

Table 3 shows substantial degradation on the independent OASIS cohort. Phase-4 DenseNet121 retains the highest external accuracy (0.3784) and macro F1 (0.2705), whereas the baseline DenseNet121 has the highest external AUC (0.6322). Phase 4 loses 0.4142 absolute accuracy, 0.5577 macro F1, and 0.3909 AUC from internal to external evaluation (Figures 2 and 3). The drop is therefore not a minor threshold effect; it affects both class assignment and probability discrimination.

Table 3. Verified external OASIS generalization performance

Model	Accuracy	Precision	Recall	Macro F1	AUC
Custom CNN	0.3390	0.3123	0.2621	0.1549	0.5845
MobileNetV2	0.3236	0.3753	0.2653	0.2310	0.5683
EfficientNetB0	0.2352	0.2974	0.2859	0.2340	0.5397
DenseNet121 baseline	0.3647	**0.3948**	0.2742	0.1944	**0.6322**
DenseNet121 Phase 4	**0.3784**	0.2795	**0.2939**	**0.2705**	0.5441
Soft-voting ensemble	0.3362	–	–	0.1493	0.5655

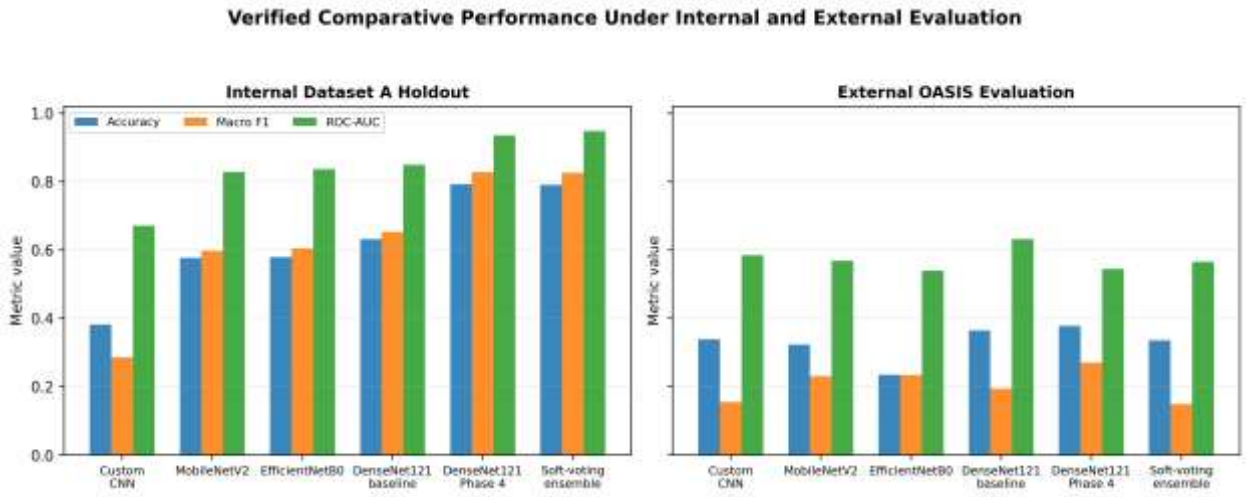


Figure 2. Verified internal and external performance across the six evaluated configurations. The same aggregate metrics are shown without reconstructing unavailable class-level prediction curves.

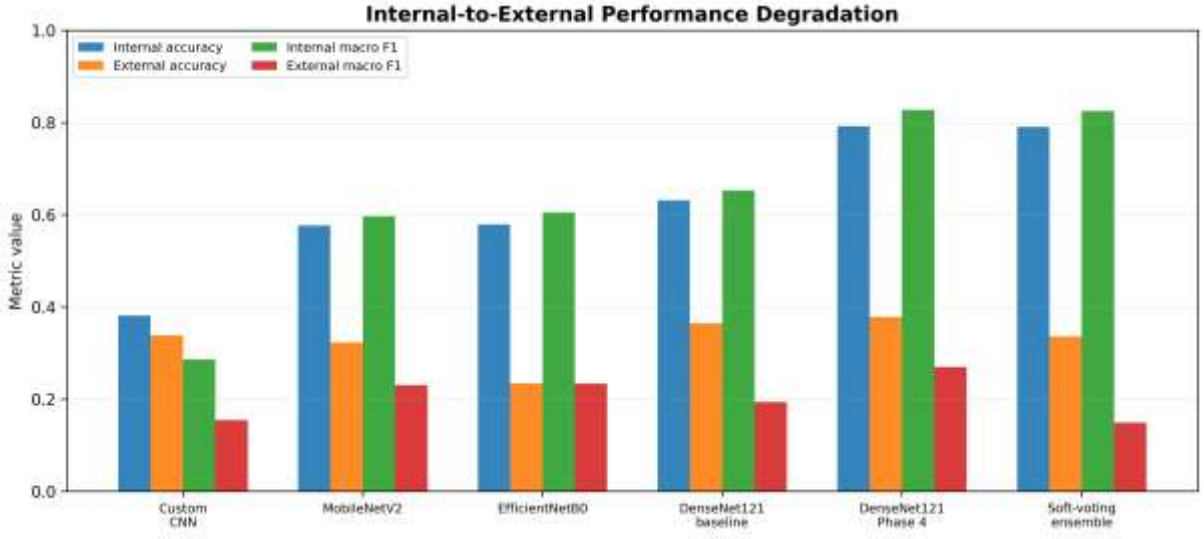


Figure 3. Internal-to-external degradation in accuracy and macro F1. External OASIS performance is reported as a limitation rather than hidden by pooled evaluation.

A separately documented balancing experiment achieved an external accuracy of 0.4030. It is not substituted for the direct 0.3784 Phase-4 result because the protocol differs. This separation avoids post-hoc replacement of an unfavorable primary external result.

4.3 Confusion-Matrix, ROC, and Grad-CAM Evidence Audit

The audit produced an important reproducibility finding. A standalone project appendix contains a numerical confusion matrix and stylized per-class ROC curves. However, the matrix contains 484 observations and implies a macro F1 of approximately 0.8075 when recomputed from its cells, whereas the verified Phase-4 aggregate table reports a macro F1 of 0.8282. The consolidated evidence package therefore explicitly rejects the standalone matrix/curve artifact as final evidence.

Figure 4 summarizes the decision rule. A final confusion matrix requires the exact holdout ground-truth and predicted-label vectors; a final ROC curve requires ground-truth labels and class probabilities. Because these vectors were not recovered, the manuscript retains the verified aggregate AUC but does not draw invented ROC geometry. This is preferable to reporting visually plausible yet unreproducible class-level plots.

Natural Resources for Human Health
eISSN: 2583-1194

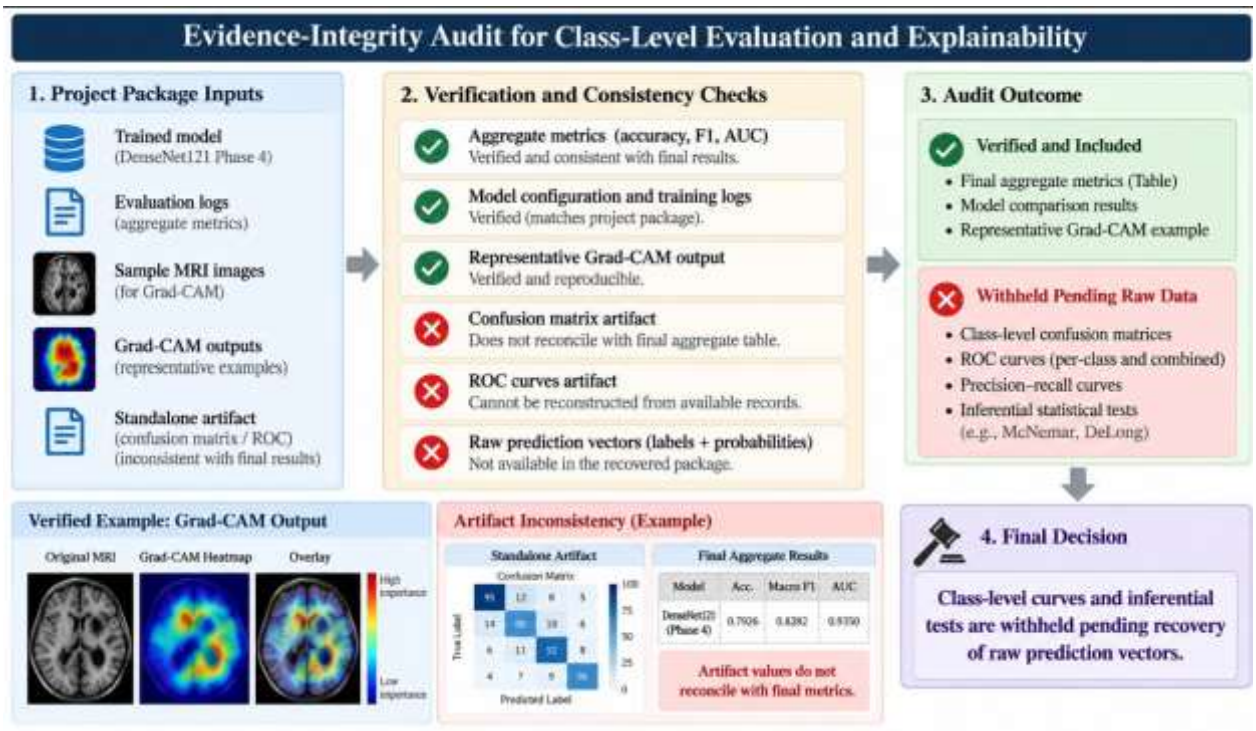


Figure 4. Evidence-integrity audit for class-level evaluation and explainability. The recovered project package verifies aggregate metrics and a representative Grad-CAM output, but the standalone confusion-matrix/ROC artifact does not reconcile with the final aggregate table. Final class-level curves and inferential tests are therefore withheld pending recovery of raw prediction vectors.

The Grad-CAM pipeline is better supported: the implementation, target DenseNet121 block, and representative original/heatmap/overlay triplet are present in the project evidence. Nevertheless, no quantitative anatomical localization claim is made, because the active preprocessing path does not include atlas registration and no expert-annotated region-level ground truth is available.

4.4 Discussion

The internal comparison shows that a targeted DenseNet121 optimization results in a substantial improvement over the baseline and lightweight models considered on the four-class development benchmark. The external experiment, however, changes the scientific interpretation. The optimized model decreases from 79.26% internal accuracy to 37.84% external accuracy, and from 0.8282 to 0.2705 macro F1, indicating that the internal benchmark alone does not establish transferable performance.

The leakage audit is no less important. The balanced development pool, the fixed holdout, and patient/group-aware logic (where identifiers exist) are documented, but the exact lineage of augmented derivatives cannot be fully recovered. This prevents a strong claim of leak-free evaluation. Future experiments should create the split at the subject/original-image level first, persist a manifest, and then apply augmentation only within the training split, verified by hash- and parent-ID-based duplicate checks before training.

The FCDS redesign constitutes a mathematical, not an empirical, problem. In the legacy implementation, confidence enters the deviation score positively. The revised formulation uses the full class-probability distribution to calculate expected MRI severity and reports confidence separately as CRS. This avoids treating "certainty" as "severity." However, the fusion coefficient λ remains uncalibrated because the project lacks a matched prospective MRI–questionnaire cohort with an independent cognitive endpoint. Accordingly, this paper does not claim that revised fusion improves diagnostic accuracy. A future study should estimate λ on training/validation data only and evaluate it against a held-out outcome.

The evidence audit also prevents overinterpretation of explainability and statistics. Grad-CAM can reveal where a model attended, but it does not prove disease-specific anatomy [11], [28]. Similarly, a reported AUC is not sufficient to reconstruct an ROC curve, and aggregate accuracy is not sufficient to reconstruct a confusion matrix. Raw predictions are also necessary for

paired hypothesis testing, calibration, and bootstrap uncertainty. Recent methodological reviews increasingly treat such traceability as part of trustworthy medical AI rather than an optional reporting detail [16], [18], [19].

4.5 Limitations and Future Work

This study has five principal limitations. First, Dataset A is an augmented image cohort whose exact upstream release/license and augmentation-family split lineage were not fully recovered. Second, the four-class OASIS image mapping is a derived project representation; future work should preserve subject-level source manifests linking every derived image to the public OASIS record. Third, raw holdout prediction vectors required for final confusion matrices, ROC/PR curves, calibration measures, and formal paired statistical tests were not available in the recovered evidence package. Fourth, QRS and FCDS are research-prototype components and have not been prospectively calibrated against a matched clinical outcome. Fifth, Grad-CAM outputs are qualitative local attributions and lack atlas-registered or expert-annotated anatomical ground truth.

The next experimental release should therefore enforce subject-first splitting, store immutable prediction/probability exports, report repeated-seed or cross-validation variability, calculate bootstrap 95% confidence intervals, apply paired McNemar testing to class decisions and appropriate AUC comparison, and report Brier score/ECE with reliability diagrams. Cross-site robustness should be investigated using harmonization, domain adaptation, or domain-generalization methods rather than assuming that further internal tuning will solve the OASIS performance gap [9], [13], [15].

5. CONCLUSION

This paper presented an evidence-auditable evaluation of HE-CDT, a multisource explainable AI framework for cognitive-deviation research in aging brains. Optimized DenseNet121 achieved strong internal four-class performance (79.26% accuracy, 0.8282 macro F1, 0.9350 AUC), while the soft-voting ensemble achieved the highest internal AUC (0.9477). Independent OASIS evaluation exposed a major generalization limitation, with optimized DenseNet121 falling to 37.84% accuracy, 0.2705 macro F1, and 0.5441 AUC. Rather than masking this result, the study treats it as central evidence of domain sensitivity.

The study also strengthens methodological integrity by auditing split leakage, refusing to reconstruct unreconciled confusion-matrix/ROC artifacts, separating classifier confidence from deviation severity, and defining the raw evidence required for future inferential and calibration analysis. HE-CDT should therefore be interpreted as a research decision-support prototype whose strongest current contribution is not universal diagnostic performance, but a transparent framework for comparative modeling, multisource scoring, explainability, and explicit cross-dataset stress testing.

Acknowledgements

[Add acknowledgements here, e.g., institutional support from the Sinhgad Institute of Management, data providers (OASIS), and any funding sources.]

Author Contributions

Dipali Prakash Patil: Conceptualization, Methodology, Software, Formal Analysis, Writing—Original Draft. [Please confirm.]

Chandrani Singh: Supervision, Validation, Writing—Review & Editing. [Please confirm.]

Conflict of Interest

The authors declare that there is no conflict of interest. [Please confirm.]

Ethics Statement

[State the ethical approval for the 42-item questionnaire study (committee name and approval number) and informed-consent procedure; note that OASIS data are publicly available de-identified data used under their data-use terms.]

Data Availability Statement

The OASIS data are publicly available from the OASIS project. Other data and code supporting the findings of this study are available from the corresponding author upon reasonable request. [Please confirm.]

REFERENCES

[1] Graham, S.A., et al., 2020. Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review. *Psychiatry Research* 284, 112732. https://doi.org/10.1016/j.psychres.2019.112732

[2] Wen, J., et al., 2020. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis* 63, 101694. https://doi.org/10.1016/j.media.2020.101694

[3] Qiu, S., et al., 2022. Multimodal deep learning for Alzheimer's disease dementia assessment. *Nature Communications* 13, 3404. https://doi.org/10.1038/s41467-022-31037-5

[4] Yu, Z., Mulholland, A., Huang, T., Liu, Q., 2026. Multimodal AI for Alzheimer disease diagnosis: Systematic review of datasets, models, and modalities. *Journal of Medical Internet Research* 28, e85414. https://doi.org/10.2196/85414

[5] Resen, R.D., Alzubaidi, L.S., Alwzwazy, H.A., Capel, M.I., 2026. Review of deep learning models for Alzheimer's disease detection: MRI-centric approaches and multimodal extensions. *Frontiers in Artificial Intelligence* 9, 1707043. https://doi.org/10.3389/frai.2026.1707043

[6] Zhang, R., Sheng, J., Zhang, Q., Wang, J., Wang, B., 2025. A review of multimodal fusion-based deep learning for Alzheimer's disease. *Neuroscience* 576, 80–95. https://doi.org/10.1016/j.neuroscience.2025.04.035

[7] Shaikh, M.R., Jeyabose, A., Arjunan, R.V., 2025. Deep learning for Alzheimer's disease: Advances in classification, segmentation, subtyping, and explainability. *Biomedical Engineering Online* 24, 150. https://doi.org/10.1186/s12938-025-01482-6

[8] Lew, C.O., Zhou, L., Mazurowski, M.A., Doraiswamy, P.M., Petrella, J.R., 2023. MRI-based deep learning assessment of amyloid, tau, and neurodegeneration biomarker status across the Alzheimer disease spectrum. *Radiology* 309(1), e222441. https://doi.org/10.1148/radiol.222441

[9] 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. https://doi.org/10.1002/hbm.26707

[10] Chattopadhyay, T., et al., 2024. Comparison of explainable AI models for MRI-based Alzheimer's disease classification. *bioRxiv*. https://doi.org/10.1101/2024.09.17.613560

[11] Qian, J., Li, H., Wang, J., et al., 2023. Recent advances in explainable artificial intelligence for magnetic resonance imaging. *Diagnostics* 13(9), 1571. https://doi.org/10.3390/diagnostics13091571

[12] Yu, A.C., Mohajer, B., Eng, J., 2022. External validation of deep learning algorithms for radiologic diagnosis: A systematic review. *Radiology: Artificial Intelligence* 4(3), e210064. https://doi.org/10.1148/ryai.210064

[13] Matta, S., Lamard, M., Zhang, P., Le Guilcher, A., Borderie, L., Cochener, B., Quellec, G., 2024. A systematic review of generalization research in medical image classification. *Computers in Biology and Medicine* 183, 109256. https://doi.org/10.1016/j.compbiomed.2024.109256

[14] Guo, B., Lu, D., Szumel, G., Gui, R., Li, P., Wang, T., Konz, N., Mazurowski, M.A., 2026. The impact of scanner domain shift on deep learning performance in medical imaging: An experimental study. *International Journal of Computer Assisted Radiology and Surgery*. https://doi.org/10.1007/s11548-026-03655-7

[15] Kumar, S., Singh, P., 2024. Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives. *Computers in Biology and Medicine*. https://doi.org/10.1016/j.compbiomed.2023.107912

[16] Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M., 2024. Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine* 150, 102830. https://doi.org/10.1016/j.artmed.2024.102830

[17] Kurz, A., et al., 2022. Uncertainty estimation in medical image classification: Systematic review. *JMIR Medical Informatics* 10(8), e36427. https://doi.org/10.2196/36427

[18] Moassefi, M., et al., 2023. Reproducibility of deep learning algorithms developed for medical imaging analysis: A systematic review. *Journal of Digital Imaging* 36(5), 2306–2312. https://doi.org/10.1007/s10278-023-00870-5

[19] Sulaimanov, U., et al., 2026. Are AI neuroimaging models ready for clinical use? A systematic methodological review. *Journal of Clinical Medicine* 15(9), 3441. https://doi.org/10.3390/jcm15093441

[20] Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience* 19(9), 1498–1507. https://doi.org/10.1162/jocn.2007.19.9.1498

[21] Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2010. Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience* 22(12), 2677–2684. https://doi.org/10.1162/jocn.2009.21407

[22] Basaia, S., et al., 2019. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clinical* 21, 101645. https://doi.org/10.1016/j.nicl.2018.101645

[23] Spasov, S., Passamonti, L., Duggento, A., Liò, P., Toschi, N., 2019. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *NeuroImage* 189, 276–287. https://doi.org/10.1016/j.neuroimage.2019.01.031

[24] El-Sappagh, S., Alonso, J.M., Islam, S.M.R., Sultan, A.M., Kwak, K.S., 2021. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Scientific Reports* 11, 2660. https://doi.org/10.1038/s41598-021-82098-3

[25] Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708. https://doi.org/10.1109/CVPR.2017.243

[26] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. MobileNetV2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

[27] Tan, M., Le, Q.V., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*, PMLR 97, pp. 6105–6114.

[28] Selvaraju, R.R., et al., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. https://doi.org/10.1109/ICCV.2017.74

[29] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*, PMLR 70, pp. 1321–1330.

[30] Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78(1), 1–3.

[31] DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 44(3), 837–845.

[32] McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 153–157.