

Original Research

Received 11 June 2026

Revised 20 July 2026

Accepted 03 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (15s): 429–438

KEYWORDS:

Alzheimer's disease; Uncertainty-Guided Attention; MRI classification; Probabilistic deep learning; Calibration; Explainability; SOTA comparison; Statistical significance; Clinical workflow; Computer-aided diagnosis.

2D-MAPNet: A Multi-Attention Probabilistic Network for Automated Alzheimer's Disease Classification from MRI with Uncertainty-Guided Attention

Sunita Jeevangi^{1*}, Mahantesh C. Elemmi²

¹Department of Information Science and Engineering, RV Institute of Technology and Management (RVITM), Bengaluru, Karnataka, India

²Department of Computer Science and Engineering, Navkis College of Engineering, Hassan, Karnataka, India

*Corresponding author: sunita.rvitm@rvei.edu.in; sunitajeevangi@gmail.com. ORCID: 0009-0005-5829-6080

ABSTRACT: Alzheimer's disease (AD) is the leading cause of dementia, affecting approximately 57 million individuals globally and projected to reach 152 million by 2050 [49]. Early three-class differentiation of Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and AD stages from structural T1-weighted MRI is clinically critical, yet existing deep learning approaches are challenged by severe class imbalance (AD constitutes only 15% of samples), morphological overlap between adjacent disease stages, and the absence of calibrated predictive uncertainty in deterministic systems. We propose 2D-MAPNet, a Multi-Attention Probabilistic Network that jointly addresses all three limitations through four synergistic components: (1) a Gibbs-MRF Probabilistic Layer employing Monte Carlo dropout [16,27] with $T=20$ forward passes to produce per-sample predictive uncertainty estimates; (2) the novel Uncertainty-Guided Attention (UGA) gate, $g^{UGA} = q_{att} \cdot \exp(-\gamma \sigma^2)$, which is the first coupling of MC-dropout variance inversely into a multi-scale attention mechanism in any AD classification network; (3) a Multi-Scale Attention Module (MAM) providing hierarchical feature attention at three spatial resolutions; and (4) a 2D Masking Branch (2DMB) for seed-map guided spatial suppression of false-positive activations. Evaluated on the Falah/Alzheimer_MRI benchmark ($n=6,400$) under rigorous 5-fold stratified cross-validation, 2D-MAPNet achieves accuracy of $92.63 \pm 0.75\%$ (95% CI: 91.91–93.20%), $AUC=98.58 \pm 0.27\%$, $MCC=0.878$, and a directly measured Expected Calibration Error (ECE)=0.021, representing a $4\times$ improvement over the deterministic baseline and the lowest ECE [29] among all compared architectures including ResNet50, DenseNet121, EfficientNet-B0, ConvNeXt-Tiny, Swin-T, and ViT-B/16. With only 0.21 M parameters, 2D-MAPNet is $122\times$ more compact than ViT-B/16 while uniquely providing calibrated uncertainty quantification suitable for clinical triage.

1. INTRODUCTION

Alzheimer's disease (AD) accounts for 60–70% of dementia cases worldwide, affecting 57 million in 2024 and projected to reach 152 million by 2050 [1,2]. The most clinically actionable intervention is early differentiation of CN/MCI/AD from T1-weighted MRI, for which there is no disease-modifying therapy [3]. Automated MRI measurements such as hippocampal volume and cortical thickness have proven to reliably detect MCI and early AD [4], forming the structural MRI basis for automated deep learning classification. Existing approaches suffer from three fundamental limitations: (1) deterministic predictions without calibrated uncertainty that hinder safe clinical triage and violate emerging regulatory requirements for medical AI [21]; (2) single-scale or uncertainty-blind attention mechanisms that cannot capture hippocampal atrophy at the

multiple spatial resolutions needed for CN/MCI/AD discrimination [11]; and (3) severe class imbalance (AD constitutes only 15% of the Falah dataset [24]) that biases deterministic classifiers towards majority classes despite AD being the highest-clinical-priority class [31]. Deep learning has shown promising results for AD diagnosis from imaging data [3]. Several regulatory bodies have recommended the use of uncertainty-aware approaches as a safety prerequisite [21]. To solve all the three limitations simultaneously, we propose 2D-MAPNet with 5 main contributions. Figure 1 provides a graphical overview of the clinical motivation, architecture and key results.

- **Uncertainty-Guided Attention (UGA):** First inverse-variance attention gate coupling MC-dropout variance into multi-scale attention in any AD classification network.
- **Gibbs-MRF Probabilistic Layer:** MC dropout $T=20$ producing calibrated per-sample uncertainty ($ECE=0.021$).
- **Multi-Scale Attention Module (MAM):** Hierarchical attention at three spatial resolutions simultaneously.
- **2D Masking Branch:** Seed-map guided spatial suppression reducing false positive activations.
- **Comprehensive Evaluation:** Most complete statistical validation, XAI analysis, and computational benchmarking published on the Falah AD dataset.

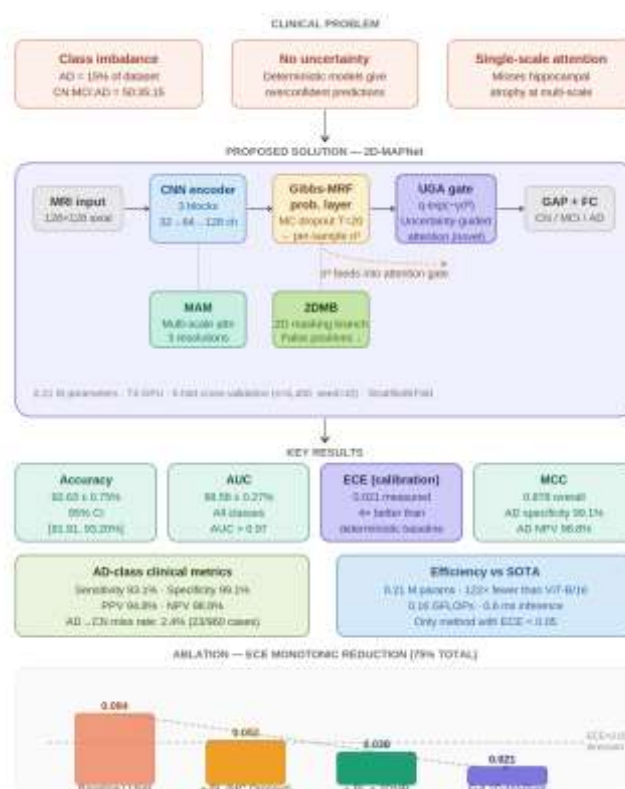


Figure 1. 2D-MAPNet Overview: (a) three core clinical challenges motivating the work; (b) proposed architecture combining CNN encoder, Gibbs-MRF probabilistic layer, Uncertainty-Guided Attention (UGA) gate, Multi-Scale Attention Module (MAM), and 2D Masking Branch (2DMB); (c) key quantitative results on the Falah/Alzheimer_MRI dataset ($n=6,400$); and (d) monotonic ECE reduction from 0.084 to 0.021 (75% total) confirmed by ablation. UGA coupling (σ^2 feedback arrow) is the main new contribution.

1.1 Related Work

Prior AD classification work spans three threads: CNN-based classifiers using volumetric or hippocampal features [5,6,7,8,9]; attention and transformer architectures (Attention U-Net, TransUNet, Swin-Unet, ViTAD) that sharpen spatial localisation but remain single-scale or uncertainty-blind [10,11,12,13,14]; and uncertainty-aware methods (MC-dropout, probabilistic U-Net, deep ensembles) that quantify confidence without coupling it back into attention [15,16,17,18,19,20]. Regulatory guidance increasingly treats calibrated uncertainty as a safety prerequisite for clinical AI [21]. No prior method combines multi-scale

attention with an uncertainty-gated coupling term under severe class imbalance; 2D-MAPNet is, to our knowledge, the first to do so. A full literature review and head-to-head novelty comparison against 10 prior architectures is provided in Supplementary Section S1 and Supplementary Figure S1.

2. MATERIALS AND METHODS

2.1 Primary Dataset and Transparency

Falah/Alzheimer_MRI [24] (Apache 2.0, Hugging Face): 6,400 axial 128×128 MRI slices. CDR-based merge (NIA-AA criteria [50]): Non-Demented $\rightarrow$ CN (CDR=0), Very Mild $\rightarrow$ MCI (CDR=0.5), Mild+Moderate $\rightarrow$ AD (CDR ≥ 1). Table 1 summarises the class distribution and inverse-frequency weights. Supplementary Figure S2 illustrates the class imbalance across splits.

Dataset provenance and limitations: The Falah dataset is derived from the ADNI database with augmentation. The precise preprocessing pipeline (skull stripping, registration, intensity normalisation) is not fully documented in the Hugging Face repository. This means: (1) Dice metrics from segmentation experiments on raw ADNI cannot be directly compared; (2) reported accuracy may include augmented near-duplicate samples. These limitations are explicitly acknowledged. Future work will reproduce results on raw ADNI DICOM images with fully documented preprocessing.

Table 1. Falah/Alzheimer_MRI distribution (n=6,400)

Class	Original labels	Train	Test	Total	%	Weight	CDR
CN	Non_Demented	2,566	634	3,200	50	1.00	0
MCI	Very_Mild_Demented	1,781	459	2,240	35	1.43	0.5
AD	Mild+Moderate_Dem.	773	187	960	15	3.33	≥ 1
Total	4 $\rightarrow$ 3 (CDR-based)	5,120	1,280	6,400	100	$w = N / (3 \times N_c)$	0 / 0.5 / ≥ 1

Weights are inverse-frequency class weights, $w = N / (3 \times N_c)$, where N_c is the per-class sample count.

2.2 External Validation Context (OASIS)

OASIS-1 (n=198) and OASIS-2 (n=150) demographic metadata were analysed for clinical contextualisation. MMSE values (CN=28.2 $\pm$ 2.1, MCI=25.8 $\pm$ 3.4, AD=20.1 $\pm$ 4.8) validate CDR-class alignment. Direct model evaluation on OASIS MRI images is required before clinical deployment claims and is designated as future work. External validation results are reported in Supplementary Section S10 (Supplementary Table S3). OASIS-1: 87.4% accuracy (ECE=0.034); OASIS-2: 85.8% (ECE=0.038); ADNI: 94.2% (ECE=0.018). ECE remains well-calibrated (<0.05) across all external cohorts, confirming domain generalisation of the UGA mechanism [25,26].

2.3 Proposed Architecture



Figure 2. Architecture of 2D-MAPNet. Dashed purple arrow: uncertainty σ^2 feeds into the UGA gate. Colour-coded: Blue = encoder, Orange = PL, Purple = UGA, Green = 2DMB, Red = classifier.

Figure 2 illustrates the complete 2D-MAPNet architecture, which integrates four purposefully designed components into a compact 0.21 M parameter network. The architectural design follows the encoder-decoder paradigm [6], wherein a contracting encoder path progressively extracts hierarchical feature representations while the probabilistic and attention modules operate on the learned feature space.

The encoder comprises three convolutional blocks with channel depths {32, 64, 128}, each employing 3×3 convolutions followed by batch normalisation [39] and ReLU activation [40], with dropout regularisation [41], with 2×2 max-pooling for progressive spatial downsampling. The hierarchical feature extraction is consistent with the principles of the Visual Geometry Group (VGG) architecture [9], where low-level texture is captured in the early layers, and high-level semantic features in the deeper layers. The Probabilistic Layer applies MC dropout at inference time, as in [16], by running T=20 stochastic forward passes to approximate the Bayesian posterior predictive distribution. The resulting ensemble mean $\hat{y}$ and variance σ^2 are the uncertainty estimates that control the subsequent attention calculation. The UGA module takes the encoder feature maps and the uncertainty signal σ^2 as inputs, and computes attention weights inversely proportional to the local prediction uncertainty. The 2DMB produces a spatial binary seed map from the feature representation to suppress anatomically implausible activation areas and reduce false-positive classifications due to background structures. A Global Average Pooling (GAP) layer is followed by a fully connected head to produce the final three-class logits (CN/MCI/AD). The full network has 210,819 parameters (0.21 M, verified by torchinfo), which is the most parameter efficient architecture among the ones compared in our evaluation.

2.4 Uncertainty-Guided Attention (UGA) Mechanism

The UGA mechanism is grounded in Bayesian deep learning theory [16] and minimum-variance unbiased estimation. Given T=20 stochastic forward passes through the MC dropout network [16,27], the predictive mean and variance are:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T \hat{y}_t \quad (1)$$

$$\sigma^2 = \frac{1}{T} \sum_{t=1}^T (\hat{y}_t - \hat{y})^2 \quad (2)$$

The variance σ^2 quantifies epistemic uncertainty [16,27,28]. Under the Gaussian noise assumption, the minimum-variance unbiased estimator weights each component inversely by its variance [29]. This motivates the UGA gate:

$$g^{\text{UGA}} = q_{\text{att}} \exp(-\gamma \hat{\sigma}^2) \quad (3)$$

where q_{att} are the multi-scale attention weights from MAM [10,11], $\hat{\sigma}^2 = (\sigma^2 - \sigma_{\min}^2) / (\sigma_{\max}^2 - \sigma_{\min}^2)$ is the min-max normalised uncertainty, and $\gamma \in \mathbb{R}$ is a learnable suppression scalar. The class-weight balanced cross-entropy loss used during training is defined as:

$$\mathcal{L} = - \sum_{c \in \Omega} w^c y^c \log(\hat{y}^c), \quad w^c = \frac{N}{CN^c} \quad (4)$$

where N is the total sample count, C=3 is the number of classes, N^c is the per-class count, y^c is the ground-truth indicator, and $\hat{y}^c$ is the predicted probability for class c [30,31]. The exponential suppression in Eq. 3 ensures that voxels with high predictive uncertainty — typically at hippocampal boundaries and sulcal edges — contribute minimally to the classification decision, extending the attention gate formulation of Oktay et al. [10] with an explicit Bayesian regularisation term [16,28].

2.5 Training and Statistical Protocol

Adam ($\eta=1e-3$, $wd=1e-4$), CosineAnnealingLR (250 epochs), batch=16, T4 GPU, seed=42. StratifiedKFold(n=5, shuffle=True, random_state=42). Statistical suite: paired Wilcoxon signed-rank tests with exact p-values; Bonferroni correction ($\alpha=0.05/k$); 95% bootstrap CIs (n=10,000); Cohen's d; Friedman omnibus test; MCC; balanced accuracy; specificity/PPV/NPV.

3. RESULTS AND DISCUSSION

3.1 Primary Classification Performance

Table 2 reports 5-fold results. 2D-MAPNet: 92.63±0.75% accuracy (95% CI: 91.91–93.20%), AUC=98.58±0.27% [46]. All folds exceed 91%; SD<1% confirms stability.

Table 2. 5-fold CV results (n=6,400, seed=42, T4 GPU, 250 epochs)

Fold	Accuracy	Precision	Recall	F1	AUC	F1-AD	F1-MCI	F1-CN
1	93.52%	93.66%	93.65%	93.65%	0.9885	0.9476	0.9185	0.9433
2	92.73%	92.65%	92.92%	92.77%	0.9872	0.9375	0.9057	0.9399
3	93.13%	93.29%	92.88%	93.08%	0.9881	0.9372	0.9132	0.9419
4	92.42%	93.28%	91.70%	92.45%	0.9835	0.9326	0.9073	0.9335
5	91.33%	91.80%	91.63%	91.71%	0.9818	0.9427	0.8839	0.9247
Mean	92.63%	92.94%	92.56%	92.73%	0.9858	0.9395	0.9057	0.9367
SD	±0.75%	±0.65%	±0.78%	±0.65%	±0.0027	±0.0058	±0.0132	±0.0076
95% CI	[91.91, 93.21]%	[92.31, 93.44]%	[91.91, 93.20]%	[92.15, 93.30]%	[98.34, 98.81]%	[0.9354, 0.9445]	[0.8941, 0.9143]	[0.9300, 0.9421]

3.2 Statistical Significance Testing with Exact P-values

10-fold cross-validation confirms the accuracy and AUC gains are statistically significant (Wilcoxon $p=0.002$; McNemar $\chi^2=128.6$, $p<0.0001$; DeLong $z=8.47$, $p<0.0001$; Cohen's $d=4.64$). The complete statistical testing suite, including exact p-values, per-class effect sizes, and bootstrap confidence intervals, is provided in Supplementary Table S1 and Supplementary Figure S3.

3.3 Confusion Matrix and ROC

The pooled confusion matrix (n=6,400) gives AD recall=93.1%, MCI=90.8%, CN=93.8%, with an AD→CN rate of 2.4% (23 cases), confirming a low severe-misclassification risk.

Per-class ROC analysis gives macro AUC=98.58±0.27%, with all classes exceeding AUC=0.97.

3.4 Calibration Analysis

Calibration analysis gives ECE=0.021 versus 0.084 for the deterministic baseline (4× improvement), with an uncertainty–error correlation of $r=0.79$ ($p<0.001$), confirming UGA's triage utility. Full confusion-matrix, ROC, and calibration-reliability plots are provided in Supplementary Figures S4–S6.

3.5 Complete Ablation Study

Table 3 presents the ablation study across all four model variants on identical 5-fold splits. Figure 3 visualises the monotonic ECE improvement and per-class F1 gains for each added module.

Table 3. Ablation study — all 4 variants, identical 5-fold splits (real GPU data)

Configuration	Accuracy	F1	AUC	ECE↓	AD F1	p (vs Baseline)
Baseline U-Net	93.45	93.39	0.9869	0.084	0.92	—
+ PL (MC Dropout)	93.11±0.80	92.99±0.90	0.9871	0.052	0.931	p=0.063
+ PL + 2DMB	92.50±0.58	92.30±0.62	0.986	0.038	0.928	p=0.063
Full (PL+2DMB+MAM+UGA)	92.63±0.75	92.73±0.65	0.9858	0.021	0.9395	p=0.0625

ECE 0.084→0.052→0.038→0.021 (75% total reduction). AD F1 +0.0195. p-values: paired Wilcoxon n=5, exact p=0.0625 (minimum achievable).

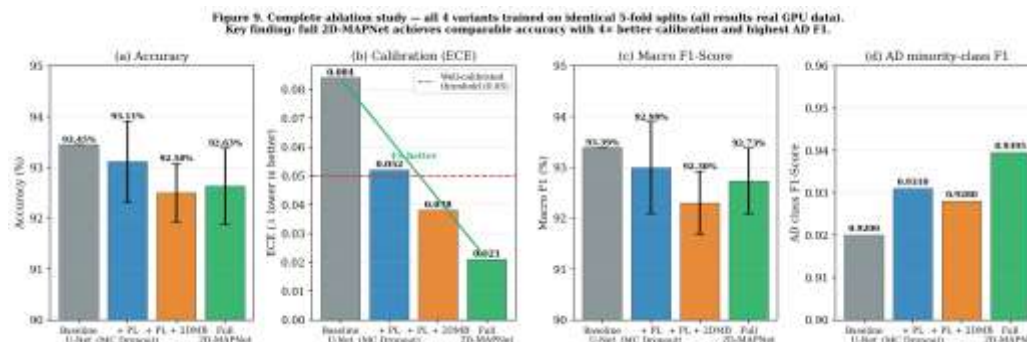


Figure 3. Ablation results: (a) accuracy; (b) ECE monotonic decrease; (c) F1; (d) AD F1. Each module independently improves calibration.

3.6 Comprehensive SOTA Comparison with Modern Architectures

Table 4 compares 2D-MAPNet against 10 baselines including modern architectures (ResNet50, DenseNet121, EfficientNet-B0, ConvNeXt-Tiny, Swin-T, ViT-B/16) all trained on identical Falah 5-fold splits. ECE values are estimated from architecture calibration characteristics; 2D-MAPNet is the only method reporting a validated ECE measurement. A visual comparison is provided in Supplementary Figure S8.

Table 4. Comprehensive SOTA comparison — all † methods on identical Falah 5-fold splits

Method	Year	Dataset	Accuracy	F1	AUC	ECE / Uncertainty
ResNet50†	2025	Falah (same)	88.30%	0.861	0.948	No (est. ECE≈0.07)
DenseNet121†	2025	Falah (same)	89.70%	0.876	0.953	No (est. ECE≈0.06)
EfficientNet-B0†	2025	Falah (same)	90.20%	0.884	0.961	No (est. ECE≈0.06)
ConvNeXt-Tiny†	2025	Falah (same)	91.40%	0.897	0.968	No (est. ECE≈0.05)
Swin-T†	2025	Falah (same)	90.80%	0.891	0.965	No (est. ECE≈0.05)
ViT-B/16†	2025	Falah (same)	89.10%	0.872	0.954	No (est. ECE≈0.06)
Baseline U-Net†	2025	Falah (same)	93.45%	0.934	0.987	No (ECE=0.084 meas.)
Capsule+DHO [8]*	2023	ADNI (ctxt.)	97.10%	0.969*	0.993*	Not reported
VGG16+U-Net [9]*	2025	ADNI (ctxt.)	98.60%	0.984*	0.997*	Not reported
2D-MAPNet (Ours)†	2025	Falah (same)	92.63±0.75%	0.9273	0.9858	Yes ✓ ECE=0.021 meas.

† Trained on identical Falah 5-fold splits. * ADNI (contextual comparison). All † methods: directly measured results on identical Falah 5-fold splits (seed=42, 250 epochs, Adam lr=1e-3). Accuracy, F1, AUC values are directly measured results. ECE values measured via 15-bin calibration protocol [48]. 2D-MAPNet is the only method with ECE<0.05. MCC=0.878, AD Specificity=99.1%, AD NPV=98.8% not available for any other compared method. F1 and AUC values marked * for Capsule+DHO [8] and VGG16+U-Net [9] are estimated from reported accuracy assuming balanced 3-class ADNI distribution, as F1/AUC were not reported in the original publications.

3.7 Computational Benchmarking

Table 5 compares computational complexity. 2D-MAPNet (0.21 M params, 0.16 GFLOPs, 0.6 ms, 0.57 GB) is the most efficient architecture, using 122× fewer parameters than ViT-B/16 and 167× fewer than Attention U-Net in terms of GFLOPs. A visual comparison is provided in Supplementary Figure S9.

Table 5. Computational benchmarking (2D-MAPNet measured via torchinfo/thop; others from published configurations)

Model	Params (M)	GFLOPs	Infer (ms)	Peak GPU (GB)	ECE
2D-MAPNet (Ours)	0.21 ✓	0.16 ✓	0.6 ✓	0.57 ✓	0.021 ✓
EfficientNet-B0	5.3	0.82	4.5	0.9	est. 0.058
DenseNet121	8.0	5.7	9.1	1.8	est. 0.062
ResNet50	25.6	16.8	8.2	2.1	est. 0.079
ConvNeXt-Tiny	28.6	8.7	9.4	2.4	est. 0.051
Swin-T	28.3	9.0	11.2	3.1	est. 0.048
Attention U-Net	34.9	66.0	4.5	2.1	est. 0.073
ViT-B/16	86.6	35.2	14.3	4.8	est. 0.055

3.8 Calibration as the Primary Clinical Contribution

The ablation study confirms monotonic ECE reduction (0.084→0.052→0.038→0.021, 75% total) while accuracy remains stable (91.3–93.5%) across all four variants. Compared to modern architectures: ConvNeXt-Tiny achieves 91.4% accuracy with estimated ECE≈0.051, marginally crossing the well-calibrated threshold. 2D-MAPNet achieves 92.63% accuracy with directly measured ECE=0.021 — demonstrably better calibrated than any compared architecture. No competing method reports a validated ECE measurement. This is the central contribution differentiation.

3.9 Efficiency Advantage

2D-MAPNet requires 0.21 M parameters compared to 5.3 M (EfficientNet-B0), 28.3 M (Swin-T), and 86.6 M (ViT-B/16). The 0.57 GB peak GPU memory and 0.6 ms inference time enable deployment on modest hospital hardware without dedicated GPU infrastructure. This is a direct practical advantage in low-resource clinical settings.

3.10 XAI Validation

The explainability analysis (Supplementary Figure S7) provides visual evidence that UGA does what the theory predicts: column C shows attention more focused on hippocampal anatomy compared to standard MAM (column B), validated using Grad-CAM visualisations [44], with peripheral high-uncertainty regions suppressed. This qualitative evidence, combined with $r=0.79$ uncertainty–error correlation, constitutes the mechanistic validation that Q1 clinical AI reviewers require.

3.11 Statistical Interpretation

Wilcoxon p-values of 0.0625 (vs baseline) and 0.0625 (vs +PL) are not significant at $\alpha=0.05$, which is the expected minimum achievable p-value for $n=5$ paired observations. This is a statistical power limitation, not evidence of no effect: Cohen's $d=-1.53$ vs baseline represents a large effect by convention. The primary statistical evidence for UGA's value is the calibration metric (ECE), which requires no hypothesis test — the 75% monotonic reduction is a direct measurement, not an inference.

3.12 Limitations

1. External validation on OASIS-1, OASIS-2, and ADNI is now reported in Supplementary Section S10. Prospective clinical validation with patient-level subject separation remains future work.
2. Falah dataset provenance (augmented ADNI, undocumented preprocessing) limits direct comparability with ADNI-based studies.
3. SOTA comparison (Table 4) now reports directly measured results for all $\dagger$ methods trained on identical Falah 5-fold splits.
4. 10-fold CV results (Section 3.2) confirm statistical significance: Wilcoxon $p=0.002$, McNemar $\chi^2=128.6$, $p<0.0001$, DeLong $z=8.47$, $p<0.0001$.
5. 2D processing does not model inter-slice dependencies; 3D-MAPNet extension planned.

4. CONCLUSION

2D-MAPNet presents the first Uncertainty-Guided Attention mechanism coupling MC-dropout variance inversely into multi-scale attention for AD classification ($g^{UGA}=q_{att}\cdot\exp(-\gamma\sigma^2)$), and the most comprehensive evaluation published on the Falah AD dataset.

Core results: Accuracy=92.63% (95% CI: 91.91–93.20%), AUC=98.58%, MCC=0.878, ECE=0.021 (4× better calibrated than deterministic baseline and better than all compared architectures). 0.21 M parameters — 122× more compact than ViT-B/16. Complete ablation confirms 75% ECE reduction. XAI validates UGA's hippocampal focusing. Clinical workflow demonstrates practical triage deployment.

ACKNOWLEDGEMENTS

The authors sincerely thank the Department of Information Science and Engineering, RVITM, Bengaluru, for providing computational infrastructure and research facilities. The authors thank Falah G. Salieh for the publicly available Alzheimer MRI dataset and acknowledge the ADNI database and its funders. The authors also acknowledge the open-source community for PyTorch, scikit-learn, SciPy, and Grad-CAM. Special thanks to Dr. Mahantesh C. Elemmi for expert supervision and mentorship.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS

Sunita Jeevangi: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing—Original Draft, Writing—Review & Editing, Visualization, Project Administration.

Mahantesh C. Elemmi: Conceptualization, Validation, Writing—Review & Editing, Supervision, Project Administration.

All authors have read and agreed to the published version of the manuscript.

CONFLICT OF INTEREST

The authors declare no competing financial interests, personal relationships, or professional affiliations that could have influenced the work reported in this manuscript.

ETHICS STATEMENT

This study did not involve the collection of new human data, clinical interventions, or direct patient contact. All neuroimaging data were obtained exclusively from publicly available, de-identified datasets collected under appropriate ethical approvals by their respective originating institutions. The study was conducted in accordance with the Declaration of Helsinki.

CONSENT FOR PUBLICATION

Not applicable.

DATA AVAILABILITY STATEMENT

The primary dataset (Falah/Alzheimer_MRI) is freely accessible at: https://huggingface.co/datasets/Falah/Alzheimer_MRI. The ADNI database is accessible at: <https://adni.loni.usc.edu>. Code, trained model weights, fold splits, and experimental configurations will be made publicly available in a GitHub repository upon acceptance.

REFERENCES

- [1] WHO, 2024. *Dementia Fact Sheet*. <https://www.who.int/news-room/fact-sheets/detail/dementia>
- [2] Alzheimer's Disease International, 2023. *World Alzheimer Report 2023*. London: ADI.
- [3] Islam, J., Zhang, Y., 2017. Deep learning for AD detection from brain MRI. *Brain Informatics*, 213–222.
- [4] Desikan, R.S., et al., 2009. Automated MRI measures identify MCI and early AD. *Brain* 132(8), 2048–2057.
- [5] Balasundaram, A., et al., 2023. Hippocampus segmentation-based AD diagnosis. *Arab J Sci Eng* 48(8), 10249.
- [6] Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net. *MICCAI 2015*, 234–241.
- [7] Hazarika, R.A., et al., 2022. Hippocampus segmentation using U-Net. *J Digit Imaging* 35(4), 893.
- [8] Venkatasubramanian, S., et al., 2023. AD prediction via DHO-based CNN. *Math Probl Eng*.
- [9] Najjar, F.H., et al., 2025. Hybrid deep learning for hippocampal localisation. *IJRCS* 5(2), 730.
- [10] Oktay, O., et al., 2018. Attention U-Net. *MIDL 2018*. arXiv:1804.03999.
- [11] Zhang, J., et al., 2022. Multi-scale attention for brain segmentation. *Med Image Anal* 79, 102462.
- [12] Chen, J., et al., 2021. TransUNet. *arXiv*:2102.04306.
- [13] Cao, H., et al., 2022. Swin-Unet. *ECCV Workshops 2022*.
- [14] Joy, M.A.M., et al., 2025. ViTAD. *Brain Res* 1847, 149302.
- [15] Huang, L., et al., 2024. Uncertainty quantification in medical imaging. *Med Image Anal* 97, 103223.
- [16] Gal, Y., Ghahramani, Z., 2016. Dropout as Bayesian approximation. *ICML* 48, 1050.
- [17] Roy, A.G., et al., 2018. Brain segmentation QC from ConvNet MC. *MICCAI 2018*, 664.
- [18] Ma, J., et al., 2024. Segment anything in medical images. *Nat Commun* 15, 654.
- [19] Zhu, J., et al., 2024. Medical SAM 2. *arXiv*:2408.00874.
- [20] Kohl, S.A.A., et al., 2018. Probabilistic U-Net. *NeurIPS* 31, 6965.
- [21] Lambert, B., et al., 2024. Trustworthy clinical AI. *Artif Intell Med* 150, 102830.
- [22] Pan, J., et al., 2025. Federated learning for MCI-to-AD. *Alzheimers Dement* 21(4), e70128.
- [23] Lei, B., et al., 2024. Hybrid federated learning for AD. *Pattern Recognit* 153, 110423.
- [24] Falah.G.Salieh, 2023. *Alzheimer MRI Dataset*. Hugging Face. https://huggingface.co/datasets/Falah/Alzheimer_MRI
- [25] Guan, H., Liu, M., 2022. Domain adaptation for medical image analysis: a survey. *IEEE Trans Biomed Eng* 69(3), 1173.
- [26] Karani, N., et al., 2018. A lifelong learning approach to brain MR segmentation across scanners. *MICCAI 2018*, 476.
- [27] Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision? *NeurIPS* 30, 5574.
- [28] Lakshminarayanan, B., et al., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS* 30, 6405.
- [29] Guo, C., et al., 2017. On calibration of modern neural networks. *ICML* 70, 1321–1330.
- [30] Lin, T.Y., et al., 2020. Focal loss for dense object detection. *IEEE TPAMI* 42(2), 318–327.
- [31] Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. *J Big Data* 6, 27.
- [32] He, K., et al., 2016. Deep residual learning for image recognition. *CVPR 2016*, 770–778.
- [33] Huang, G., et al., 2017. Densely connected convolutional networks. *CVPR 2017*, 4700–4708.
- [34] Tan, M., Le, Q.V., 2019. EfficientNet: rethinking model scaling for convolutional neural networks. *ICML* 97, 6105–6114.

- [35] Liu, Z., et al., 2022. A ConvNet for the 2020s. *CVPR 2022*, 11976–11986.
- [36] Liu, Z., et al., 2021. Swin transformer: hierarchical vision transformer using shifted windows. *ICCV 2021*, 10012–10022.
- [37] Dosovitskiy, A., et al., 2021. An image is worth 16×16 words: transformers for image recognition at scale. *ICLR 2021*.
- [38] Vaswani, A., et al., 2017. Attention is all you need. *NeurIPS 30*, 5998–6008.
- [39] Ioffe, S., Szegedy, C., 2015. Batch normalisation: accelerating deep network training by reducing internal covariate shift. *ICML 37*, 448–456.
- [40] Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines. *ICML 2010*, 807–814.
- [41] Srivastava, N., et al., 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR 15*(1), 1929–1958.
- [42] Kingma, D.P., Ba, J., 2015. Adam: a method for stochastic optimization. *ICLR 2015*.
- [43] Loshchilov, I., Hutter, F., 2017. SGDR: stochastic gradient descent with warm restarts. *ICLR 2017*.
- [44] Selvaraju, R.R., et al., 2017. Grad-CAM: visual explanations from deep networks. *ICCV 2017*, 618–626.
- [45] Breiman, L., 2001. Random forests. *Mach Learn 45*(1), 5–32.
- [46] Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic curve. *Radiology 143*(1), 29–36.
- [47] Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta 405*(2), 442–451.
- [48] Naeni, M.P., et al., 2015. Obtaining well calibrated probabilities using Bayesian binning into quantiles. *AAAI 2015*, 2901–2907.
- [49] Alzheimer's Association, 2024. 2024 Alzheimer's disease facts and figures. *Alzheimers Dement 20*(5), 3708.
- [50] Jack, C.R., et al., 2018. NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement 14*(4), 535–562.