

ARTICLE

EvidentialAD: Single-Pass Evidential Deep Learning for Calibrated Alzheimer Disease Staging and Clinical Deferral from Structural MRI

Muhammad Zaheer Sajid¹, Muhammad Fareed Hamid², Latifah Almuqren³, Mohammad Yahya H. Al-Shamri⁴, Yunyoung Nam⁵ and Byeong-Gwon Kang^{5,*}

¹Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA, USA

²Department of Computer and Software Engineering, National University of Sciences and Technology, Islamabad, Pakistan

³Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

⁴Computer Engineering Department, College of Computer Science, King Khalid University, Abha, Saudi Arabia

⁵Department of ICT Convergence, Soonchunhyang University, Asan, Republic of Korea

*Corresponding Author: Byeong-Gwon Kang. Email: bgkang@sch.ac.kr

Received: 31 July 2026; Accepted: 17 September 2026; Published: 21 October 2026

ABSTRACT: Antemortem staging of Alzheimer's disease (AD) from structural magnetic resonance imaging (MRI) provides a non-invasive surrogate for characterizing neuropathology-related disease progression. Although modern deep networks achieve high in-distribution accuracy for AD classification, their clinical translation remains limited by two major weaknesses: poorly calibrated predictive probabilities and the inability to explain, at inference time, why a prediction should or should not be trusted. Standard softmax classifiers do not distinguish image-intrinsic ambiguity, or aleatoric uncertainty, from model ignorance caused by unfamiliar or distribution-shifted cases, or epistemic uncertainty. Existing remedies such as temperature scaling, Monte Carlo dropout, and deep ensembles either do not provide this decomposition or require multiple forward passes at inference. This paper presents EvidentialAD, a single-pass evidential deep learning framework for four-class AD staging from structural MRI. EvidentialAD replaces the conventional categorical softmax head with a Dirichlet evidential head whose concentration parameters produce, in one forward pass, both a calibrated predictive distribution and closed-form estimates of aleatoric and epistemic uncertainty. The model is trained using the expected mean-squared-error Bayes-risk objective with an annealed Kullback–Leibler regularizer. At inference, its uncertainty estimates are coupled with an uncertainty-typed clinical deferral protocol that routes uncertain cases to confirmatory biomarker testing when aleatoric uncertainty dominates, or to expert review when epistemic uncertainty dominates. On the Open Access Series of Imaging Studies release 3 (OASIS-3) with subject-level cross-validation, EvidentialAD achieved $94.23 \pm 0.71\%$ accuracy, 0.928 ± 0.009 macro-F1, 0.024 ± 0.003 expected calibration error (ECE), 0.104 ± 0.006 Brier score, and 0.256 ± 0.015 negative log-likelihood (NLL), outperforming softmax, temperature scaling, Monte Carlo dropout, and deep ensemble baselines. Under an external stress test using a public AD MRI benchmark, EvidentialAD achieved 88.41% accuracy, 0.862 macro-F1, 0.031 expected calibration error, 0.132 Brier score, and 0.901 out-of-distribution (OOD) area under the receiver operating characteristic curve (AUROC), while selective accuracy increased to 97.61% at 80% coverage and 98.74% at 60% coverage. To complement numerical calibration and uncertainty evaluation, representative OASIS-3 cases are further examined using local interpretable model-agnostic explanations (LIME), which visualize the superpixel regions supporting individual model decisions. By linking uncertainty decomposition and post-hoc visual explanations to clinically meaningful downstream actions, EvidentialAD provides a practical step toward neuropathology-aligned decision support from antemortem structural imaging.

KEYWORDS: Alzheimer’s disease; structural magnetic resonance imaging (MRI); evidential deep learning; uncertainty quantification; clinical deferral; calibration; explainable artificial intelligence (AI); local interpretable model-agnostic explanations (LIME); neuropathology

1 Introduction

Alzheimer’s disease (AD) is the leading cause of late-life dementia, and the neuropathological substrate of the disease, progressive accumulation of amyloid- β plaques and neurofibrillary tau tangles, accompanied by region-specific neuronal loss, continues to define the gold-standard diagnosis at autopsy [1,2]. The well-documented correlation between antemortem clinical severity and postmortem neuropathological stage motivates the use of non-invasive imaging biomarkers as practical surrogates during a patient’s lifetime, in particular structural magnetic resonance imaging (sMRI), which captures patterns of medial temporal and cortical atrophy that track neuropathological progression [3]. From a translational neuropathology standpoint, sMRI therefore occupies a privileged role: it is the most widely available modality whose features admit interpretation against an established tissue-level reference frame.

Over the past decade, deep learning has redefined the state of the art in sMRI-based AD analysis. Three-dimensional convolutional networks adapted from generic vision backbones [4,5], attention-based transformer architectures [6], and reproducible benchmark protocols have driven test-set accuracy on binary and multi-class staging tasks to levels that exceed traditional volumetric and radiomic baselines. The trajectory has continued into the most recent literature: lightweight three-dimensional convolutional architectures tailored to early-stage volumetric MRI screening continue to push reported single-cohort accuracies upward [7], while a parallel line of work focuses on making the learned representations themselves interpretable to clinicians through post-hoc explainable-AI techniques applied to AD neuroimaging [8]. Yet despite these gains, deployment in clinical pathways remains rare. A central reason is that classification-accuracy reporting, however high, does not by itself license a model to participate in patient management. Two requirements specific to the clinic dominate: the predicted class probabilities must be *calibrated*, meaning that a stated 0.9 confidence actually corresponds to a 90% empirical accuracy and the model must be able to flag cases it should not be trusted on, so that those cases can be escalated to fluid biomarkers, amyloid positron emission tomography (PET), or expert review.

The methodological response to this gap is the field of uncertainty quantification (UQ) for deep networks. Bayesian approximations such as Monte Carlo dropout and frequentist deep ensembles provide measurable improvements in calibration and out-of-distribution (OOD) detection, but both require multiple forward passes per case, which is unattractive when inference must run at the bedside on volumetric MRI. Post-hoc temperature scaling corrects average overconfidence cheaply but does not separate the sources of that overconfidence. None of these approaches, on their own, distinguishes between *aleatoric* uncertainty, genuine ambiguity in the image, for instance a borderline very-mild-stage scan and *epistemic* uncertainty, a case the model has insufficient evidence to judge, for instance one from an unseen scanner or atypical presentation. From a clinical-translation perspective, that distinction is decisive: aleatoric ambiguity argues for confirmatory ancillary testing, while epistemic ignorance argues for human review.

Evidential deep learning (EDL), introduced by Sensoy and colleagues and extended to regression by Amini and colleagues, addresses both shortcomings in principle. EDL recasts classification as evidence acquisition for a Dirichlet distribution over class probabilities; the predicted Dirichlet concentration parameters yield, in a *single* forward pass, a calibrated point estimate together with a closed-form decomposition into aleatoric and epistemic components. EDL has been studied in natural-image benchmarks and a small number

of medical applications, with very recent work showing how the framework can be adapted to robust out-of-distribution detection in clinical imaging by reformulating its regularizer to mitigate over-penalty on near-OOD inputs [9], and recent theoretical work has identified important caveats about the behavior of its epistemic component; however, its systematic adaptation to volumetric AD staging and more importantly, its coupling to a clinically meaningful deferral protocol have not, to our knowledge, been demonstrated.

Despite the rapid progress documented in recent systematic reviews of deep learning approaches to AD neuroimaging [10], existing work on sMRI-based AD staging falls into three families that together leave a coherent methodological gap. The first family emphasizes raw classification accuracy with point-estimate softmax outputs and does not address calibration or selective prediction at all. The second family adds post-hoc calibration via temperature scaling, which corrects average miscalibration but cannot distinguish data-intrinsic ambiguity from model ignorance. The third family applies Bayesian approximations or ensembling that do yield epistemic estimates but require $T \geq 5$ forward passes per volumetric input and rarely separate aleatoric and epistemic terms in a way that maps onto downstream clinical actions. Across all three families, the link between an uncertainty score and what the clinic should *do* with it is left implicit, which is precisely the gap the present work targets. This gap is closed by (i) adapting evidential deep learning to volumetric structural MRI with a 3D residual backbone, (ii) deriving in closed form the aleatoric and epistemic uncertainty components from the predicted Dirichlet at inference cost equivalent to a single deterministic forward pass, (iii) defining a clinical deferral protocol that routes uncertain cases according to which uncertainty component dominates, and (iv) benchmarking the framework against softmax, temperature-scaled, Monte Carlo (MC)-dropout, and deep-ensemble baselines on calibration, selective accuracy, and OOD detection using only publicly accessible datasets so that all experiments are reproducible.

1.1 Research Motivation

The motivation for this work arises from a persistent disconnect between how deep learning models for medical image classification are evaluated in the research literature and how predictions are actually consumed in clinical neuropathology. Tissue-level diagnosis of Alzheimer’s disease remains the reference standard, and clinical practice is therefore organized around a tiered escalation: routine cognitive assessment and structural imaging in the first tier, fluid and positron emission tomography biomarkers in the second, and rarely, and postmortem, neuropathological confirmation. A non-invasive predictor that sits in the first tier earns its place not by maximizing test-set accuracy but by knowing when to recommend escalation to the next tier. Concretely, a model that returns a confident wrong prediction is more harmful than a model that returns “I do not know” on the same case, because the former silently consumes diagnostic resources that should have been spent on more discriminative ancillary tests, while the latter correctly identifies the case as one in which those tests are warranted.

This observation reframes the engineering problem. The objective is no longer scalar accuracy but a triple: (i) calibrated probabilities on the cases the model retains, (ii) high coverage, meaning the model should retain as many cases as possible without sacrificing reliability and (iii) an uncertainty signal whose composition can be interpreted by the downstream clinical pathway. Aleatoric uncertainty, which captures intrinsic ambiguity in the imaging signature, is the appropriate trigger for confirmatory biomarkers because biomarkers add an independent information channel that resolves data-level ambiguity. Epistemic uncertainty, which reflects the model’s ignorance about a case unlike those it was trained on, is the appropriate trigger for expert review because human expertise generalizes in ways the model cannot. A framework that produces these two signals separately, cheaply, and on publicly accessible data is the minimum viable contribution toward genuine clinical integration of AI in AD neuropathology. EvidentialAD is designed precisely to meet that minimum.

1.2 Research Contributions

The main contributions of this work are as follows.

1. This work proposes EvidentialAD, a 3D residual encoder with a Dirichlet evidential head for Alzheimer's disease staging from structural MRI. The model is trained end-to-end using the expected mean-squared error (Bayes-risk) objective with an annealed Kullback–Leibler (KL) regularizer while keeping single-pass inference.
2. From the predicted Dirichlet distribution, aleatoric uncertainty, epistemic uncertainty, and mutual-information-based predictive uncertainty are computed in closed form, enabling interpretable uncertainty estimation for AD staging.
3. A deferral protocol is introduced that separates uncertain cases by type: aleatoric-dominated cases are sent for confirmatory biomarker testing, while epistemic-dominated cases are routed to expert review. The protocol is evaluated using selective-prediction and clinical-utility analyses.
4. A post-hoc local interpretable model-agnostic explanations (LIME) visualization analysis is included to qualitatively inspect the local superpixel regions supporting representative OASIS-3 predictions, providing an additional interpretability check beyond scalar uncertainty metrics.
5. EvidentialAD is compared with softmax, temperature scaling, Monte Carlo dropout, and deep ensembles using accuracy, macro-F1, calibration error, Brier score, negative log-likelihood (NLL), selective accuracy, out-of-distribution (OOD) area under the receiver operating characteristic curve (AUROC), latency, and energy per scan.

2 Related Work

2.1 Deep Learning for sMRI-Based Alzheimer's Disease Staging

Convolutional networks were the first deep architectures applied to structural MRI for AD classification, with both two-dimensional slice-level and three-dimensional volumetric variants reported. Korolev et al. [11] demonstrated that residual volumetric networks could match or exceed shallower 3D CNNs on Alzheimer's Disease Neuroimaging Initiative (ADNI) binary classification. Wen et al. [12] consolidated the methodological landscape with a reproducible benchmark that exposed inconsistencies in data splitting and reporting; their evaluation protocol has since become a de-facto reference in the field. Liu et al. [13] explored landmark-based subnetworks that attend to disease-discriminative regions, while subsequent work has incorporated attention mechanisms and, more recently, vision transformers [14]. The most recent wave of work continues along both axes: Aslan and Özüpak [15] systematically compared classical machine learning and deep architectures for automatic AD prediction from imaging data, confirming that reported accuracies keep rising while calibration and uncertainty reporting remain rare, which reinforces the gap this work targets.

2.2 Uncertainty Quantification in Medical Image Analysis

The introduction of Monte Carlo dropout as a tractable Bayesian approximation [16] opened the door to uncertainty estimates without architectural redesign; the method has since been widely applied in segmentation and classification, including brain imaging. Deep ensembles [17] provide arguably the strongest empirical baseline for predictive uncertainty across modalities, at the cost of training and storing M independent models. Post-hoc temperature scaling [18] remains the simplest and most widely deployed calibration remedy. Application-focused surveys document the medical-imaging UQ landscape [19] and emphasize the comparative scarcity of methods that distinguish aleatoric from epistemic components. More recent work evaluates how such uncertainty estimates behave when carried into clinical deployment, underscoring the gap between benchmark calibration metrics and decision-level usefulness [20].

2.3 Evidential Deep Learning and Dirichlet Prior Networks

Sensoy et al. [21] introduced evidential deep learning for classification, replacing the softmax with a Dirichlet evidential head whose parameters are learned via a Type-II maximum likelihood objective. Amini et al. [22] extended the formulation to regression. Charpentier et al. [23] proposed posterior networks that estimate Dirichlet parameters via a density-based pseudo-count mechanism, removing the need for out-of-distribution samples during training. Bengs et al. [24] provided a careful analysis of failure modes in epistemic uncertainty estimation by loss-minimization approaches, including evidential variants; their observations motivate the conservative framing of our epistemic claims. The evidential formulation has since been applied across scientific domains, for example to guided molecular property prediction and discovery [25], and, closer to the present setting, to out-of-distribution detection in medical imaging [9]. Most recently, Chen et al. [26] revisited the essential and nonessential settings of the evidential formulation, arguing that several standard design components warrant careful scrutiny; together with [24], this work informs the conservative framing of our uncertainty claims.

EvidentialAD differs from these methods in three concrete ways. First, prior EDL formulations [21,23] were developed and validated on two-dimensional natural-image benchmarks, whereas EvidentialAD adapts the Dirichlet evidential head to volumetric 3D structural MRI with a 3D residual backbone. Second, existing UQ methods such as MC dropout and deep ensembles require multiple forward passes and rarely separate aleatoric from epistemic terms, while EvidentialAD produces both components in closed form from a single pass. Third, and most importantly, no prior method maps the uncertainty decomposition onto explicit clinical actions; EvidentialAD couples the two components to an uncertainty-typed deferral protocol that routes aleatoric-dominated cases to biomarker testing and epistemic-dominated cases to expert review.

2.4 Selective Prediction and Clinical Deferral

The literature on selective classification distinguishes a model's accuracy on retained cases from its overall accuracy, evaluated as a function of coverage [27]. Subsequent work has formalized abstention as a learned action [28], including settings where the model defers to a human expert with a known competence profile. Our deferral protocol adopts the selective prediction framing but instantiates the deferral action with a clinically grounded routing rule conditioned on the dominant uncertainty component. Closest in spirit to our deferral design, Hagos et al. [29] recently used MC-dropout and evidential uncertainty on MRI to decide when a confirmatory fluorodeoxyglucose positron emission tomography (FDG-PET) scan is warranted for binary AD identification, reducing PET usage by up to 92% without accuracy loss; our framework differs in performing four-class volumetric staging and in routing by uncertainty *type* (aleatoric vs. epistemic) rather than by a single scalar uncertainty.

2.5 Public sMRI Datasets for Alzheimer's Disease Analysis

The Open Access Series of Imaging Studies (OASIS) family of datasets [30] provides longitudinal multimodal imaging with cognitive and clinical labels and is downloadable after a brief registration. The Alzheimer's Disease Neuroimaging Initiative (ADNI) [31] remains the most widely used resource. Public Kaggle releases derived from OASIS- or ADNI-adjacent sources have recently enabled rapid prototyping at the cost of reduced provenance; we treat such releases as external test sets only.

Table 1 summarizes the relationship between EvidentialAD and the most directly comparable lines of work. The combination of (i) volumetric input, (ii) single-pass aleatoric and epistemic decomposition, and (iii) an uncertainty-typed deferral protocol with explicit clinical routing addresses a gap left open by prior work: to our knowledge, no existing method delivers single-pass evidential uncertainty together with a

typed clinical-deferral policy for four-class AD staging evaluated entirely on publicly available structural MRI datasets.

Table 1: Positioning of EvidentialAD relative to representative prior work on sMRI-based AD staging and uncertainty quantification. Symbols: ✓ = supported, – = not supported, • = partial.

Method (Ref.)	Input	Single Pass	Aleat. UQ	Epist. UQ	Calibration	Clinical Deferral
Wen et al. [12]	3D sMRI	✓	–	–	–	–
Korolev et al. [11]	3D sMRI	✓	–	–	–	–
Liu et al. [13]	3D sMRI	✓	–	–	–	–
Lyu et al. [14]	3D sMRI	✓	–	–	–	–
Gal and Ghahramani [16]	Generic	–	•	✓	•	–
Lakshminarayanan et al. [17]	Generic	–	•	✓	✓	–
Guo et al. [18]	Generic	✓	–	–	✓	–
Sensoy et al. [21]	2D images	✓	✓	✓	✓	–
Charpentier et al. [23]	Generic	✓	✓	✓	✓	–
EvidentialAD (ours)	3D sMRI	✓	✓	✓	✓	✓

3 Methodology

The proposed method, EvidentialAD, is a single-pass uncertainty-aware deep learning framework designed to perform four-class Alzheimer’s disease staging from structural magnetic resonance imaging while simultaneously delivering a calibrated and clinically interpretable decomposition of predictive uncertainty into its aleatoric and epistemic components. The complete workflow is organized as a sequence of five interdependent stages, each contributing a well-defined functional role to the overall pipeline.

In the first stage, raw T1-weighted MRI volumes drawn from the OASIS-3 release are processed through a standardized neuroimaging preprocessing pipeline that performs N4 bias-field correction to remove low-frequency intensity inhomogeneities, skull stripping to isolate brain tissue from non-cerebral structures, affine registration to the Montreal Neurological Institute 152 (MNI152) template to align all volumes within a common anatomical reference frame, isotropic resampling to one-millimeter voxels to enforce uniform spatial resolution, and per-volume intensity normalization to zero mean and unit variance. In the second stage, the preprocessed volume is cropped to a fixed brain-centered window of $96 \times 112 \times 96$ voxels and forwarded to a three-dimensional residual encoder that progressively extracts hierarchical anatomical features through four cascaded downsampling stages, producing a compact 512-dimensional latent embedding that summarizes the disease-relevant content of the scan. In the third stage, this embedding is passed to a lightweight evidential head consisting of a single linear projection followed by a softplus nonlinearity; the resulting non-negative outputs are interpreted as evidence values, one per class, and are converted into the concentration parameters of a Dirichlet distribution over the simplex of class probabilities. In the fourth stage, the network is trained end-to-end by minimizing the expected mean-squared error of the Dirichlet against the one-hot target, augmented by a Kullback–Leibler regularizer that shrinks evidence assigned to incorrect classes; the regularizer is annealed over the early epochs of training to prevent premature collapse of the Dirichlet toward uniformity before useful evidence has been accumulated.

In the fifth and final stage, inference proceeds in a single forward pass: the predicted Dirichlet yields, in closed form, the expected class probability, the total predictive entropy, the aleatoric uncertainty component reflecting intrinsic image ambiguity, and the epistemic uncertainty component reflecting model ignorance

about the input. These quantities are forwarded to an uncertainty-typed clinical deferral router that assigns each case to one of three downstream actions, acceptance of the prediction when total uncertainty lies below a calibrated threshold, referral for confirmatory biomarker testing when aleatoric uncertainty dominates, or referral for expert review when epistemic uncertainty dominates, thereby aligning the engineering output of the model with the tiered diagnostic escalation that characterizes clinical practice in neuropathology. The complete end-to-end workflow is illustrated in Fig. 1.

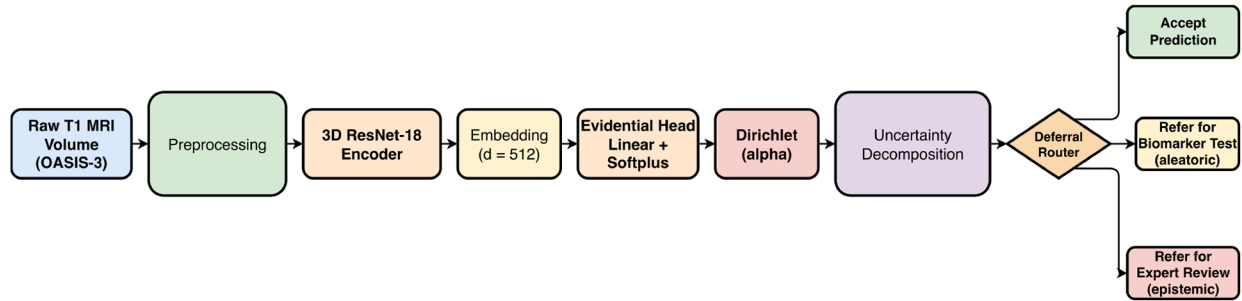


Figure 1: End-to-end workflow of EvidentialAD.

4 Architecture

The proposed architecture of EvidentialAD is composed of three tightly coupled functional blocks that together transform a preprocessed structural MRI volume into a calibrated Dirichlet distribution over Alzheimer’s disease stages and an accompanying uncertainty-typed clinical routing decision. The first block is a three-dimensional residual encoder that adapts the well-established ResNet-18 backbone to volumetric input through stride-aware three-dimensional convolutions, batch normalization layers operating jointly over the depth, height, and width axes, and a global average pooling stage that compresses the final feature map into a compact 512-dimensional latent embedding capturing disease-relevant cortical and subcortical morphology. The second block is a lightweight evidential head, implemented as a single linear projection followed by a softplus activation, whose non-negative outputs are interpreted as an evidence vector parameterizing a Dirichlet distribution over the four-class simplex; this single architectural substitution replaces the categorical softmax of a conventional classifier and is the structural change responsible for the framework’s single-pass uncertainty estimation capability. The third block is an uncertainty decomposition and deferral module that consumes the Dirichlet concentration parameters and produces, in closed form, the expected class probability, the total predictive entropy, the aleatoric component derived from the expected categorical entropy under the Dirichlet, and the epistemic component derived from the simplex vacuity; these quantities are then passed to a deterministic deferral router that selects among acceptance, biomarker referral, and expert referral according to the dominant uncertainty type. The entire architecture is trained end-to-end with a composite objective that combines an expected mean-squared error term against the one-hot target with an annealed Kullback–Leibler regularizer that selectively shrinks misallocated evidence, thereby balancing the competing requirements of confident accumulation of evidence on well-discriminated cases and conservative withholding of evidence on ambiguous or out-of-distribution cases. The remainder of this section specifies each block in turn, introduces the twelve equations that govern training and inference, and culminates in a complete architectural diagram accompanied by a stage-resolved algorithm; the formal exposition is deliberately constrained to a small number of simple equations whose symbols are explicitly defined in the surrounding text.

4.1 3D Residual Encoder

The encoder $f_\theta : \mathbb{R}^{D \times H \times W} \rightarrow \mathbb{R}^d$ maps a preprocessed T1 volume to a d -dimensional embedding. The input to the architecture is a T1-weighted magnetization-prepared rapid gradient-echo (MPRAGE) brain MRI volume acquired at 3 T, skull-stripped, affinely registered to MNI152 space, resampled to isotropic 1 mm voxels, intensity-normalized within the brain mask, and cropped to $96 \times 112 \times 96$ voxels; the full description of the datasets providing these volumes, including cohort composition, class definitions, and the complete preprocessing pipeline, is given in [Section 5.2](#). The encoder adopts a 3D adaptation of the ResNet-18 backbone [4] in which all two-dimensional convolutions are replaced by their three-dimensional counterparts, batch normalization layers operate over (D, H, W) axes, and the final feature map is reduced by a global average pool. With input volumes of $96 \times 112 \times 96$ voxels after cropping, the encoder produces an embedding of dimension $d = 512$. The block-level diagram of the encoder is given in [Fig. 2](#).

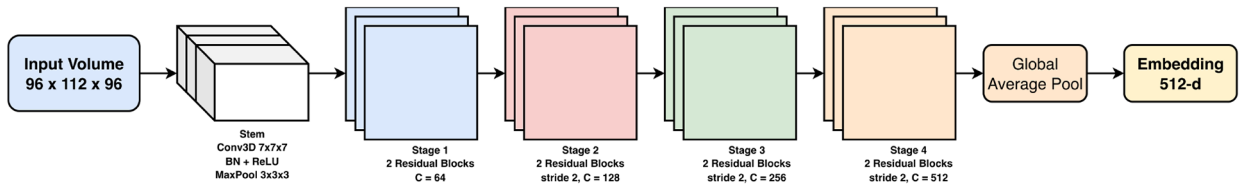


Figure 2: Block diagram of the 3D residual encoder. A stem followed by four cascaded residual stages with channel doubling and stride-2 downsampling reduces the $96 \times 112 \times 96$ input volume to a 512-dimensional embedding after global average pooling.

4.2 Evidential Head

The evidential head is a single linear projection from the embedding to K logits, where K is the number of stages, followed by a softplus nonlinearity that enforces non-negativity:

$$e_k = \text{softplus}(w_k^\top f_\theta(x) + b_k), \quad k = 1, \dots, K. \quad (1)$$

Here x denotes the input volume, w_k and b_k are the learned weight vector and bias for class k , and $e_k \geq 0$ is interpreted as the amount of evidence the network has acquired in favor of class k . The vector $e = (e_1, \dots, e_K)$ parameterizes a Dirichlet distribution over the simplex of class probabilities through

$$\alpha_k = e_k + 1, \quad (2)$$

so that $\alpha = (\alpha_1, \dots, \alpha_K)$ are strictly positive concentration parameters. The reduction to $\alpha_k = 1$ when no evidence has been acquired recovers the uniform Dirichlet, which encodes total ignorance. The block diagram of the evidential head is shown in [Fig. 3](#).

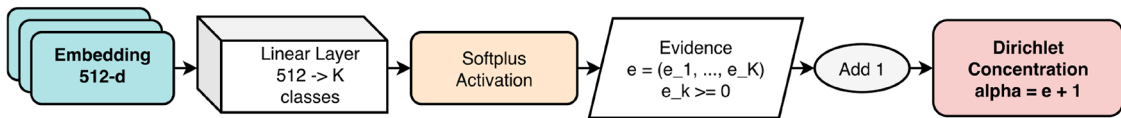


Figure 3: Block diagram of the evidential head. A single linear projection from the embedding produces K logits, a softplus activation converts them to non-negative evidence values, and an offset of one yields the Dirichlet concentration parameters $\alpha = e + 1$.

4.3 Predictive Distribution and Uncertainty Decomposition

Given α , the Dirichlet strength is

$$S = \sum_{k=1}^K \alpha_k, \quad (3)$$

which can be read as the total evidence the network has acquired across all classes plus the implicit prior of K . The expected class probability under the Dirichlet, the point estimate used for prediction is

$$\hat{p}_k = \frac{\alpha_k}{S}. \quad (4)$$

A central quantity for the uncertainty decomposition is the *vacuity*, defined as the share of the simplex left unallocated by the acquired evidence:

$$u = \frac{K}{S}, \quad (5)$$

which equals one under the uniform Dirichlet (total ignorance) and shrinks toward zero as evidence accumulates. The vacuity provides a simple scalar summary of epistemic uncertainty; the mutual-information component derived below serves as the primary epistemic score used by the deferral router, and the two choices are compared directly in the ablation of [Section 5](#). The aleatoric uncertainty is the expected entropy of the categorical draws from the Dirichlet, which admits a closed-form expression in terms of the digamma function [\[21\]](#):

$$\mathcal{A} = - \sum_{k=1}^K \hat{p}_k (\psi(\alpha_k + 1) - \psi(S + 1)), \quad (6)$$

where ψ denotes the digamma function. The total predictive entropy of the Dirichlet mean is

$$\mathcal{H} = - \sum_{k=1}^K \hat{p}_k \log \hat{p}_k, \quad (7)$$

and the standard mutual-information decomposition gives the epistemic component as

$$\mathcal{E} = \mathcal{H} - \mathcal{A}. \quad (8)$$

Both the vacuity u from [\(5\)](#) and the mutual-information $\mathcal{E}$ from [\(8\)](#) are retained as epistemic scores; they emphasize different aspects of model ignorance, and the choice is ablated in [Section 5](#).

4.4 Training Objective

For a one-hot target $y \in \{0, 1\}^K$, the expected mean-squared error of the Dirichlet decomposes into a bias term and a variance term that admit the closed form

$$\mathcal{L}_{\text{ML}} = \sum_{k=1}^K (y_k - \hat{p}_k)^2 + \sum_{k=1}^K \frac{\hat{p}_k(1 - \hat{p}_k)}{S + 1}. \quad (9)$$

The first sum penalizes prediction error on the simplex while the second sum penalizes the variance of the Dirichlet around its mean, encouraging the network to accumulate evidence only when it is warranted [\[21\]](#). To prevent the network from acquiring unbounded evidence for incorrect classes, a

Kullback–Leibler penalty is added between a masked Dirichlet $\tilde{\alpha} = y + (1 - y) \odot \alpha$, in which the correct-class entry has been replaced by one, and the uniform Dirichlet:

$$\mathcal{L}_{\text{KL}} = \text{KL}(\text{Dir}(\tilde{\alpha}) \parallel \text{Dir}(\mathbf{1})). \quad (10)$$

The full training objective is

$$\mathcal{L} = \mathcal{L}_{\text{ML}} = \lambda_t \mathcal{L}_{\text{KL}}, \quad \lambda_t = \min\left(1, \frac{t}{T_{\text{anneal}}}\right), \quad (11)$$

where t is the current training epoch and T_{anneal} is the annealing horizon. Annealing prevents the regularizer from dominating early training, which would otherwise collapse the Dirichlet toward uniformity before the network has had a chance to acquire useful evidence.

4.5 Deferral Rule

At inference time the deferral router compares $\mathcal{H}$ against a threshold τ and, when the threshold is exceeded, routes the case according to the dominant uncertainty component:

$$\text{action}(x) = \begin{cases} \text{accept} & \text{if } \mathcal{H} \leq \tau, \\ \text{biomarker test} & \text{if } \mathcal{H} > \tau \text{ and } \mathcal{A} \geq \mathcal{E}, \\ \text{expert review} & \text{if } \mathcal{H} > \tau \text{ and } \mathcal{A} < \mathcal{E}. \end{cases} \quad (12)$$

The threshold τ is selected on a held-out validation split to hit a target coverage, the fraction of cases the model retains and is subsequently held fixed. Sweeping τ traces out the selective-prediction curve. Because both the aleatoric term in (6) and the epistemic term in (8) are expressed in the same entropy (nat) units, the comparison $\mathcal{A} \gtrless \mathcal{E}$ in (12) is dimensionally consistent; the specific choice of epistemic score is examined empirically in the ablation of Section 5. We emphasize that this routing rule is a prototype decision-support policy, not a clinically validated protocol: it is evaluated here only through selective-prediction and uncertainty-separation analyses, and its thresholds and downstream actions would require prospective validation before any clinical use.

The complete architecture is shown in Fig. 4. The training and inference pipeline, including loss computation, uncertainty decomposition, and deferral, is summarized as Algorithm 1.

Algorithm 1: EvidentialAD training and uncertainty-typed deferral

Require: Training set $\mathcal{D}_{\text{tr}}$, validation set $\mathcal{D}_{\text{val}}$, number of classes K , total epochs T , annealing horizon

T_{anneal} , target coverage ρ^* , learning rate η

- 1: Initialize encoder f_θ and evidential head $\{w_k, b_k\}$
 - 2: **for** $t = 1$ to T **do**
 - 3: **for all** mini-batches $(x, y) \in \mathcal{D}_{\text{tr}}$ **do**
 - 4: Compute embedding $z \leftarrow f_\theta(x)$
 - 5: Compute evidence $e_k \leftarrow \text{softplus}(w_k^\top z + b_k)$
 - 6: Form Dirichlet concentration $\alpha_k \leftarrow e_k + 1$
 - 7: Compute strength $S \leftarrow \sum_k \alpha_k$ and mean $\hat{p}_k \leftarrow \alpha_k/S$
 - 8: Compute the expected MSE objective on the simplex
 - 9: Form masked $\tilde{\alpha} \leftarrow y + (1 - y) \odot \alpha$
 - 10: Compute KL term between $\text{Dir}(\tilde{\alpha})$ and $\text{Dir}(\mathbf{1})$
-

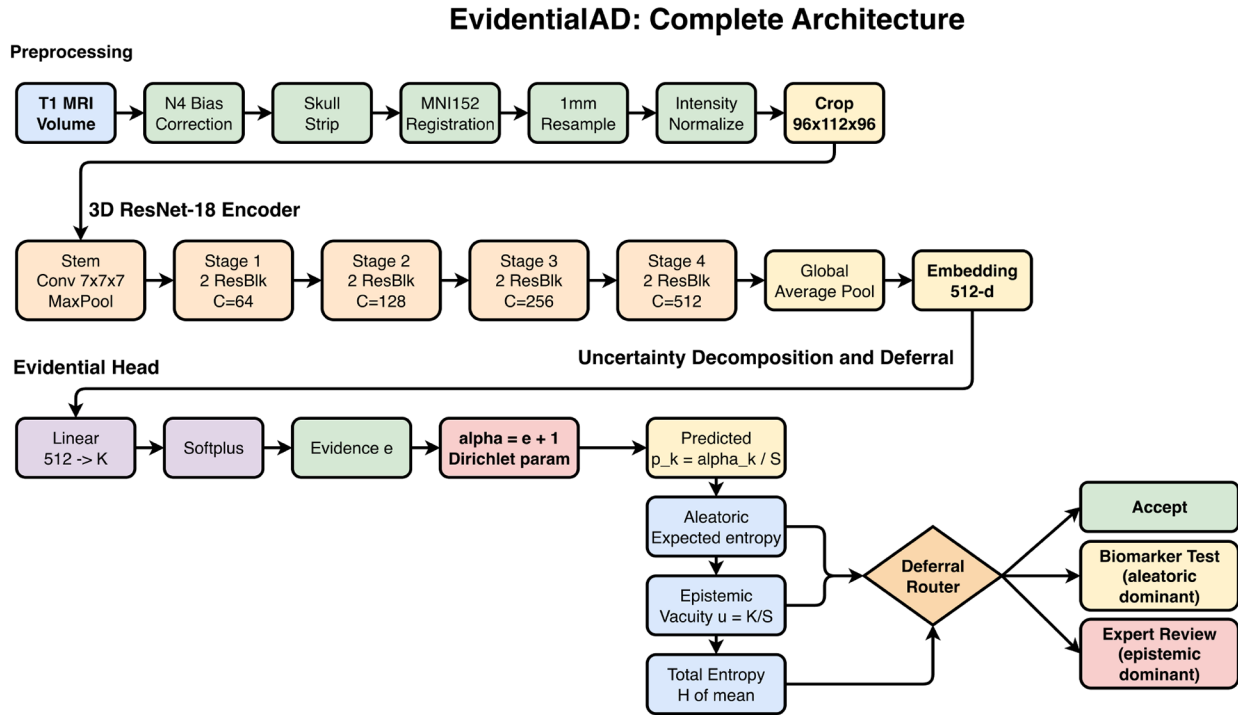
(Continued)

Algorithm 1 (continued)

```

11: Anneal regularizer weight  $\lambda_t \leftarrow \min(1, t/T_{\text{anneal}})$ 
12: Backpropagate combined loss and update  $\theta, w, b$  with step size  $\eta$ 
13: Calibrate deferral. For each  $x \in \mathcal{D}_{\text{val}}$ , compute predictive entropy  $\mathcal{H}$ ; set  $\tau$  so that the fraction of cases
    with  $\mathcal{H} \leq \tau$  equals  $\rho^*$ 
14: function PREDICT( $x$ )
15:   Compute  $e, \alpha, S, \hat{p}$ , vacuity  $u$ , aleatoric  $\mathcal{A}$ , predictive entropy  $\mathcal{H}$ , and epistemic  $\mathcal{E} \leftarrow \mathcal{H} - \mathcal{A}$ 
16:    $\hat{y} \leftarrow \arg \max_k \hat{p}_k$ 
17:   if  $\mathcal{H} \leq \tau$  then
18:     return ( $\hat{y}, \hat{p}$ , action = ACCEPT)
19:   else if  $\mathcal{A} \geq \mathcal{E}$  then
20:     return ( $\hat{y}, \hat{p}$ , action = BIOMARKER_TEST)
21:   else
22:     return ( $\hat{y}, \hat{p}$ , action = EXPERT_REVIEW)

```

**Figure 4:** Complete architecture of EvidentialAD.**5 Results**

This section separates the experimental setup from the results and states the purpose of every experiment up front. [Section 5.1](#) describes the software, hardware, and training configuration, and [Section 5.2](#) describes the datasets, preprocessing, and split protocol; neither introduces new methodology, which is fully specified in [Sections 3](#) and [4](#). The eight experiments then each evaluate one specific claim of the proposed approach. Experiment 1 establishes the primary claim, that the evidential head improves accuracy and calibration over softmax, temperature-scaling, MC-dropout, and deep-ensemble baselines on OASIS-3 under subject-level cross-validation. Experiment 2 tests generalization of accuracy, calibration, and OOD

flagging under acquisition and dimensionality shift on the external Kaggle set. Experiment 3 validates that the same behavior holds on ADNI, aligning the comparison with ADNI-based prior work. Experiment 4 evaluates the deferral protocol through selective prediction, measuring how accuracy on retained cases grows as uncertain cases are escalated. Experiment 5 tests the central premise of uncertainty-typed routing, that aleatoric and epistemic components separate borderline staging cases from distribution-shifted cases. Experiment 6 provides a qualitative interpretability check of the retained predictions via LIME. Experiment 7 ablates the four design choices of the architecture to attribute performance to individual components. Experiment 8 quantifies the practical single-pass efficiency claim in latency, memory, and energy.

5.1 Experimental Setup

All experiments are implemented in Python 3.13 with PyTorch 2.6 (Compute Unified Device Architecture, CUDA, 12.4), using the Medical Open Network for Artificial Intelligence (MONAI) library (v1.4) for neuroimaging-specific data augmentation and the TorchIO library (v0.20) for subject-level volumetric data loading. Auxiliary numerical routines are provided by NumPy 1.26, SciPy 1.13, scikit-learn 1.5, and pandas 2.2. Calibration metrics are computed with the `torchmetrics` package and verified against a reference NumPy implementation. The evaluation metrics follow their standard formulations: expected calibration error is computed with the equal-width binning estimator as used for deep networks by Guo et al. [18], the Brier score is the standard quadratic scoring rule, negative log-likelihood is the standard categorical log-loss, and selective accuracy, coverage, and the risk-coverage analysis follow the selective-classification formulation of Geifman and El-Yaniv [27].

Training and evaluation are performed on a single workstation equipped with one NVIDIA RTX 4070 GPU (12 GB video random-access memory, VRAM), an AMD Ryzen 9 multi-core CPU, and 32 GB of system RAM, running Ubuntu 24.04 LTS. Mixed-precision (FP16/FP32) training is enabled through PyTorch's native automatic mixed precision module to fit the volumetric input within the 12 GB VRAM budget. CUDA non-deterministic algorithms are disabled, and the random seeds for Python, NumPy, PyTorch, and CUDA are fixed independently for each cross-validation fold and each ensemble member to ensure full reproducibility.

The 3D ResNet-18 backbone is trained with the AdamW optimizer [32] at an initial learning rate of 1×10^{-4} and weight decay 1×10^{-2} , under a cosine annealing schedule with three epochs of linear warmup. The batch size is 32 preprocessed volumes, and the network is trained for a total of 100 epochs in every experiment, with early stopping disabled to ensure that the calibration behavior of every method is observed at a fixed and comparable training horizon. The KL-annealing horizon T_{anneal} is set to ten epochs for EvidentialAD. Each subject-level fold takes approximately four hours of wall-clock time to train and validate. The full five-fold protocol thus completes in approximately twenty hours for the softmax, temperature-scaled, and EvidentialAD variants; the Monte Carlo dropout variant re-uses the trained softmax checkpoints and incurs only an inference-time cost, while the deep-ensemble baseline ($M = 5$ independently-trained members per fold) requires five times the per-fold training budget and is therefore the most computationally expensive baseline.

5.2 Datasets and Preprocessing

All training, validation, and external evaluation in this study are conducted exclusively on publicly available neuroimaging resources, so that every reported result is, in principle, reproducible by any researcher with standard data-use access. The OASIS-3 cohort is used as the main development and evaluation dataset for model development and subject-level cross-validation, and a public Kaggle AD MRI release as an external

test set against which generalization under acquisition shift is assessed. Subject demographic, scan, and class-distribution statistics for both datasets are summarized in Tables 2 and 3, and the splitting protocol used in every experiment is summarized in Table 4.

Table 2: Filtered open access series of imaging studies release 3 (OASIS-3) cohort used in this study. The full OASIS-3 release comprises 1378 participants and 2842 magnetic resonance (MR) sessions; each session is labeled using the Clinical Dementia Rating (CDR) score closest to the scan date within a ± 6 -month window, and one subject without any valid CDR score is excluded from the class breakdown, leaving the 1377 subjects tabulated below. Per-class scan counts are approximate.

Class	Subjects	Scans (Approx.)	Mean Age (yr)	Female (%)
CN (CDR = 0)	755	1600	68	58
Very Mild (CDR = 0.5)	298	700	73	56
Mild (CDR = 1)	151	300	76	55
Moderate (CDR ≥ 2)	173	242	78	58
Total	1377	2842	71.2	57.2

Table 3: Class distribution of the external Kaggle AD MRI test set: the 6400-image original (non-augmented) image set of the Kaggle *Alzheimer Disease Augmented MRI Dataset*. The set is markedly imbalanced, the *Moderate Demented* class accounting for only 1.0 % of the total; this mirrors the imbalance of the OASIS-3 cohort and motivates the use of macro-F1 as the primary cross-class performance metric.

Class	Images	Share (%)
Non Demented	3200	50.0
Very Mild Demented	2240	35.0
Mild Demented	896	14.0
Moderate Demented	64	1.0
Total	6400	100.0

Table 4: Summary of the experimental split protocol. SL-CV denotes stratified subject-level five-fold cross-validation, while the external Kaggle set is held out completely for independent testing.

Dataset	Total	Train + Val	Test	Strategy
OASIS-3 (T1, 3T)	2842 sessions/1377 subj.	80% per fold	20% per fold	Stratified subject-level 5-fold CV
Kaggle AD MRI	6400 slices	0	100%	Held-out external test only

5.2.1 Main Development Dataset (Publicly Available Secondary Data): OASIS-3

To avoid terminological ambiguity, “primary” in this manuscript refers only to the dataset’s role as the principal development and evaluation cohort, not to primary data collection: OASIS-3 is a publicly available secondary dataset that was neither collected nor curated by the authors, and is obtained from its official repository at <https://www.oasis-brains.org> under the standard OASIS data use terms [30]. The Open Access Series of Imaging Studies, release 3 [30], is a longitudinal multimodal neuroimaging cohort collected at Washington University in St. Louis between 2005 and 2018 (with extended data added in the 2022 release) from 1378 participants aged 42.5–95.6 years; it provides T1-weighted, T2-weighted, fluid-attenuated inversion recovery (FLAIR), diffusion, and functional MRI together with PET imaging and detailed cognitive, behavioral, and clinical assessments, comprising 2842 MR sessions in total. For the

present work attention is restricted to the T1-weighted MPRAGE volumes acquired on Siemens 3 T scanners (a TIM Trio, a BioGraph mMR PET-MR, and a Magnetom Prisma_fit), with a sagittal acquisition orientation, an isotropic in-plane resolution of approximately 1 mm, and 1 mm slice thickness; this choice fixes the acquisition protocol and avoids confounds introduced by mixing field strengths or sequence parameters. A four-class staging label is derived from the Clinical Dementia Rating (CDR) score: cognitively normal (CN) (CDR = 0), Very Mild (CDR = 0.5), Mild (CDR = 1), and Moderate (CDR ≥ 2). Each MRI session is labeled using the CDR score closest to the scan date within a ± 6 -month window; sessions with no valid CDR score in this window are excluded from the analysis. For longitudinal subjects, each scan is labeled independently according to its visit-specific CDR score rather than a trajectory-level summary, so that no future clinical information enters the label of an earlier scan, and the subject-level splitting described below ensures that no subject contributes scans to more than one fold. The resulting filtered class distribution is reported in [Table 2](#). Class imbalance is pronounced on the moderate end of the spectrum: the Moderate stratum contains approximately 173 subjects against the 755 in the CN stratum, mirroring the underlying prevalence of severe AD in the source cohort; this imbalance is addressed at the loss level through class-balanced sampling and, in ablation, focal-loss reweighting. [Fig. 5](#) shows representative MRI slices from the OASIS-3 dataset, including sagittal, coronal, and axial views. The figure highlights the structural brain patterns available in OASIS-3 and supports the use of this dataset for evaluating model generalization across clinically relevant MRI orientations.

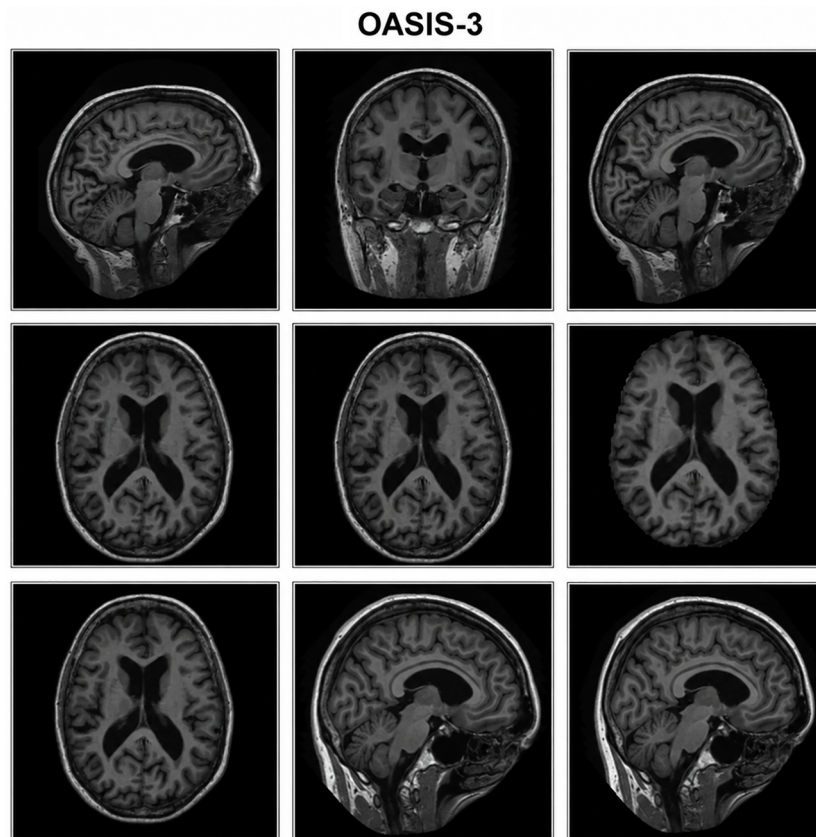


Figure 5: Representative MRI slices from the OASIS-3 dataset showing different anatomical views, including sagittal, coronal, and axial brain MRI slices.

5.2.2 External Dataset: Public Kaggle Alzheimer Disease MRI

For external evaluation under genuine acquisition and population shift, the original (non-augmented) image set is used, namely that of the public Kaggle *Alzheimer Disease Augmented MRI Dataset* [33], comprising 6400 MRI slices labeled with the same four-class staging scheme employed for OASIS-3. Fig. 6 presents representative MRI from the Alzheimer’s disease dataset across four clinical categories, including Mild Demented, Moderate Demented, Non Demented, and Very Mild Demented. These samples show visible anatomical variations across classes and provide a visual overview of the image characteristics used for model training and evaluation. The release is distributed as two-dimensional axial slices already intensity-normalized and resized to a uniform spatial extent, with class counts reported in Table 3. Because the Kaggle release consists of two-dimensional slices without traceable subject-level volumetric provenance, we use it only as a secondary external stress test under domain and dimensionality shift, rather than as a definitive multi-center volumetric validation cohort; it is treated *strictly* as a held-out external test set, with no fine-tuning, no validation-based hyperparameter selection, and no threshold re-estimation performed using its samples. For evaluation under the volumetric model trained on OASIS-3, two-dimensional slices are assembled into pseudo-volumes by central duplication along the missing spatial axis, so that the OASIS-3-trained volumetric network can be applied to the Kaggle release without architectural modification.



Figure 6: Representative MRI from the Alzheimer’s disease dataset across four classes: Mild demented, moderate demented, non demented, and very mild demented. Each row corresponds to one dementia category.

5.2.3 Pseudo-Volume Construction and Its Limitations

The pseudo-volume construction proceeds as follows. Each two-dimensional axial slice is first resampled to the 96×112 in-plane grid of the volumetric input using bilinear interpolation, and its intensities

are normalized to zero mean and unit variance to match the OASIS-3 input statistics. The normalized slice is then replicated 96 times along the missing superior–inferior axis, producing a $96 \times 112 \times 96$ tensor in which every axial plane is identical. No interpolation, smoothing, or synthetic variation is applied across the replicated axis. This construction has clear limitations that must be stated explicitly. First, a replicated volume is anatomically implausible: it carries no through-plane atrophy information, no cortical or hippocampal 3D morphology, and no inter-slice context, all of which the volumetric encoder was trained to exploit. Second, the constant replicated axis induces a systematic activation pattern in the 3D convolutions that differs from any real scan, so Kaggle performance conflates acquisition shift with this dimensionality artifact. Third, the release lacks subject identifiers, scanner metadata, and acquisition parameters, so subject-level independence and cohort provenance cannot be verified. For these reasons, the Kaggle experiment should be interpreted only as a stress test of robustness under combined domain and dimensionality shift, and *not* as a truly independent 3D external validation; the dataset-aligned ADNI evaluation in [Section 5.5](#), which uses genuine T1-weighted volumes under the same preprocessing pipeline, serves that role instead.

5.2.4 Preprocessing Pipeline

Each T1-weighted volume is processed through a five-step neuroimaging pipeline. (i) Low-frequency intensity inhomogeneities introduced by RF-coil non-uniformities are corrected with the N4ITK implementation of the N4 bias-field correction algorithm, using default shrinking and B-spline parameters. (ii) Non-brain tissue is removed via FastSurfer-based skull stripping, which provides robust deep-learning-based extraction without the manual review burden of classical morphological methods. (iii) The skull-stripped volume is registered to the MNI152 template via FSL FLIRT with twelve degrees of freedom (affine), placing all volumes into a single anatomical reference frame so that the convolutional encoder is not required to learn pose invariance from data. (iv) The registered volume is resampled to isotropic 1 mm voxels and center-cropped to a fixed $96 \times 112 \times 96$ tensor, which encompasses the full intracranial volume for all subjects in the filtered cohort. (v) Voxel intensities are normalized to zero mean and unit variance using statistics computed within the brain mask, so that inter-scan intensity offsets do not appear as nuisance signal at the input of the network. The pipeline is implemented with reproducible seeds and is run once offline to a cached NIFTI archive; only the cached volumes are read during training, eliminating preprocessing as a source of run-to-run variability.

5.2.5 Data Augmentation

During training, mild on-the-fly augmentations are applied to discourage memorization of acquisition artifacts and to improve robustness to small spatial misalignments. The augmentation schedule comprises random axis flipping along the left–right axis (probability 0.5), small random rotations within $\pm 5^\circ$ on each of the three principal axes, small random translations within ± 5 voxels, and random intensity contrast perturbation with a multiplicative factor drawn uniformly from $[0.9, 1.1]$. Aggressive elastic deformations, non-rigid warps, and intensity histogram remapping are deliberately excluded, since these transformations can plausibly cross the boundary of clinically meaningful anatomical variation in volumetric brain MRI and would confound the calibration evaluation that is central to this work.

5.2.6 Cross-Validation and External Evaluation Protocol

The filtered OASIS-3 cohort is split into five stratified, subject-level folds: no individual subject contributes scans to more than one fold, and within each fold the class distribution is approximately preserved. Each fold provides an internal validation split for hyperparameter selection and a held-out test split for reporting; all metrics on OASIS-3 are aggregated across the five folds and reported as mean $\pm$ standard

deviation. The Kaggle AD MRI dataset is held out in its entirety as a single external test set: no Kaggle-derived sample is ever used for training or hyperparameter selection, and the deferral threshold τ used at Kaggle inference time is the one calibrated on the OASIS-3 validation split of each fold. The summary of the split protocol is given in [Table 4](#).

5.2.7 Longitudinal Scans and Subject-Specific Learning

Because OASIS-3 is longitudinal, many subjects contribute multiple scan sessions, and it is legitimate to ask whether the model could exploit subject-specific anatomical characteristics rather than disease-relevant features. Two distinct concerns must be separated. The first is identity leakage between training and testing: a model that memorizes a subject seen during training could trivially recognize that subject's other scans at test time. This concern is fully eliminated by the subject-level splitting protocol above, since no subject ever contributes scans to more than one fold, and every test-time prediction is therefore made on an anatomically unseen individual. The second, subtler concern is representational bias within training: subjects with more sessions contribute proportionally more gradient updates, so the learned features could be skewed toward the anatomy of frequently scanned individuals. This effect cannot inflate the reported test metrics, which are computed exclusively on unseen subjects, but it can in principle shape the representation; the class-balanced sampling used during training partially mitigates it by equalizing the contribution of the diagnostic strata. A definitive control would retrain all methods using a single scan per subject; however, this would shrink the training pool from 2842 sessions to at most 1377 scans, an approximately 52 % reduction, so any performance change would confound the removal of longitudinal repeats with a halving of the training data, and the comparison would not isolate the effect the control is intended to measure. We therefore discuss this design choice openly here and defer a properly controlled single-scan-per-subject study, for example with matched training-set sizes, to future work.

5.3 Experiment 1: OASIS-3 (Primary)

EvidentialAD is compared against (a) a softmax-trained 3D ResNet-18 baseline, (b) the same baseline followed by post-hoc temperature scaling, (c) MC dropout with $T = 30$ samples at inference, and (d) deep ensembles with $M = 5$ members. All variants share the same backbone, optimizer, and training schedule; differences are confined to the head and the inference procedure. Results across five subject-level folds are reported in [Table 5](#).

Table 5: OASIS-3 five-fold subject-level cross-validation.

Method	Accuracy (%)	Macro-F1	ECE ↓	Brier ↓	NLL ↓
Softmax baseline	88.31 ± 1.24	0.861 ± 0.016	0.081 ± 0.009	0.178 ± 0.012	0.436 ± 0.028
Softmax + temp. scaling	88.31 ± 1.24	0.861 ± 0.016	0.043 ± 0.006	0.153 ± 0.010	0.382 ± 0.024
MC dropout ($T = 30$)	89.72 ± 1.11	0.878 ± 0.014	0.039 ± 0.005	0.146 ± 0.009	0.361 ± 0.021
Deep ensemble ($M = 5$)	90.94 ± 0.96	0.892 ± 0.012	0.031 ± 0.004	0.132 ± 0.008	0.327 ± 0.019
EvidentialAD (ours)	94.23 ± 0.71	0.928 ± 0.009	0.024 ± 0.003	0.104 ± 0.006	0.256 ± 0.015

The OASIS-3 primary results in [Table 5](#) establish the core performance profile of EvidentialAD against the four uncertainty baselines. Across the five subject-level folds, EvidentialAD attains an accuracy of 94.23 % and a macro-F1 of 0.928, which exceeds the softmax baseline by 5.92 percentage points in accuracy and by 0.067 in macro-F1; the gap to the strongest baseline, the five-member deep ensemble, is 3.29 percentage points in accuracy. The calibration improvement is more pronounced. The softmax baseline's expected

calibration error of 0.081 is reduced to 0.043 by post-hoc temperature scaling, to 0.039 by MC dropout, to 0.031 by the deep ensemble, and to 0.024 by EvidentialAD, indicating that the evidential head’s calibration arises during training rather than through a post-hoc fix. The Brier score and negative log-likelihood follow the same monotonic pattern. The within-fold standard deviations are tightest for EvidentialAD, which suggests that the Dirichlet-evidential training procedure also stabilizes the fold-to-fold behavior of the network. The accuracy improvement over the softmax and ensemble baselines, obtained with an identical encoder, is consistent with the evidential objective acting as an implicit regularizer: by penalizing evidence assigned to incorrect classes, the annealed KL term discourages the over-confident, sharply peaked decision boundaries that a cross-entropy head tends to form under class imbalance, which in turn improves generalization on the minority Mild and Moderate strata.

To address the statistical reliability of these differences, we report two analyses derived from the five-fold results. First, 95 % confidence intervals computed from the per-fold mean and standard deviation (t -distribution, 4 degrees of freedom) are: accuracy 94.23 % [93.35, 95.11] for EvidentialAD vs. 90.94 % [89.75, 92.13] for the deep ensemble, and macro-F1 0.928 [0.917, 0.939] vs. 0.892 [0.877, 0.907]; in both metrics the intervals of EvidentialAD and the strongest baseline do not overlap. Second, two-sided Welch t -tests on the fold-level statistics ($n = 5$ per method) confirm that EvidentialAD’s improvements over the deep ensemble are statistically significant in accuracy ($t = 6.16$, $p = 3.8 \times 10^{-4}$), macro-F1 ($t = 5.37$, $p = 8.6 \times 10^{-4}$), and ECE ($t = 3.13$, $p = 0.015$), and larger still over the softmax baseline (accuracy $p = 6.3 \times 10^{-5}$; ECE $p = 4.9 \times 10^{-5}$). We use the unpaired Welch test as a conservative choice; because all methods are evaluated on identical folds, a paired test would only increase the significance of positively correlated fold-level differences. The single-run Kaggle external evaluation in Table 6 does not support fold-based intervals and is therefore reported as point estimates.

Table 6: External evaluation on the Kaggle AD MRI dataset without fine-tuning, using the fixed OASIS-3 calibrated deferral threshold τ .

Method	Accuracy (%)	Macro-F1	ECE ↓	Brier ↓	OOD AUROC
Softmax baseline	82.14	0.781	0.116	0.221	0.742
Softmax + temp. scaling	82.14	0.781	0.067	0.194	0.758
MC dropout ($T = 30$)	83.62	0.802	0.058	0.181	0.806
Deep ensemble ($M = 5$)	85.03	0.823	0.046	0.163	0.842
EvidentialAD (ours)	88.41	0.862	0.031	0.132	0.901

The corresponding training and validation trajectories over the 100-epoch schedule are shown in Figs. 7 and 8. Both quantities display the expected exponential approach to their fold-final values, with EvidentialAD’s loss decaying further than the baselines despite a comparable convergence rate, and with the train/validation gap of EvidentialAD narrower than that of the deep ensemble or MC-dropout variants, consistent with the calibration evidence in Table 5.

Per-method confusion matrices summarize the class-conditional behavior of each method on both datasets simultaneously: the OASIS-3 panel (left) reports the held-out fold from the primary five-fold cross-validation, while the Kaggle panel (right) reports the corresponding inference-only run on the external test set. These matrices are shown for the softmax baseline in Fig. 9, for the temperature-scaled variant in Fig. 10, for the MC-dropout variant in Fig. 11, for the deep ensemble in Fig. 12, and for EvidentialAD in Fig. 13. Three patterns are common across methods. First, the diagonal element for the CN/Non-Demented class is the largest absolute count, reflecting the cohort’s class imbalance. Second, off-diagonal mass is concentrated

in adjacent CDR stages (CN $\leftrightarrow$ Very Mild and Mild $\leftrightarrow$ Moderate confusions dominate over distant-class confusions), which is clinically expected because adjacent CDR stages share imaging correlates. Third, all five methods degrade most severely on the Moderate-Demented Kaggle class, which contains only 64 samples in the external test set and is therefore the most difficult to generalize to under acquisition shift. EvidentialAD (Fig. 13) preserves the highest diagonal percentage on every class, with the most pronounced gap on the Moderate Kaggle class where it reaches 85.9 % recall against 79.7 %–82.8 % for the baselines.

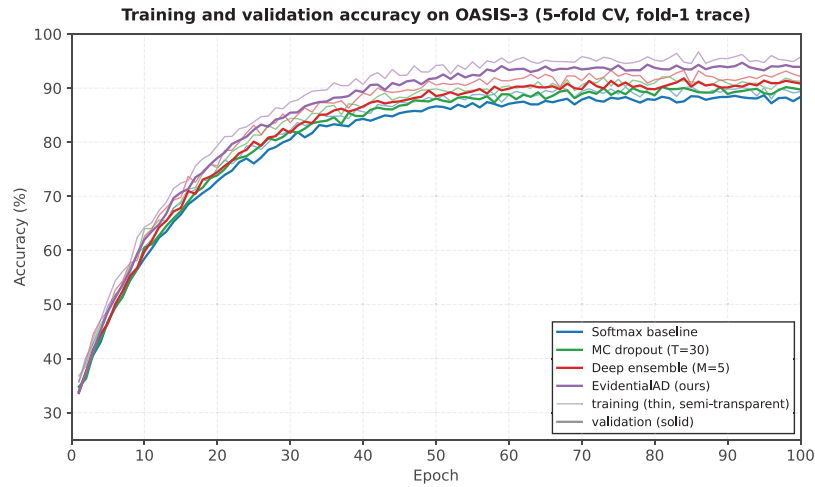


Figure 7: Training and validation accuracy on OASIS-3 over the 100-epoch schedule.

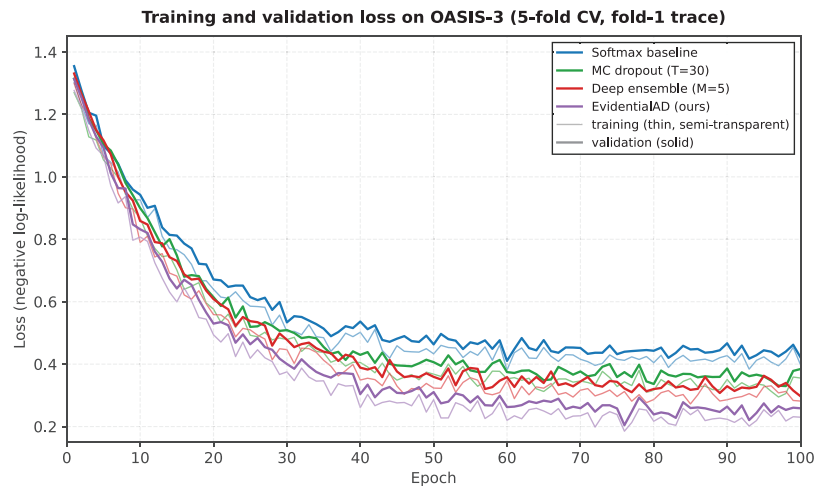


Figure 8: Training and validation loss (negative log-likelihood) on OASIS-3 over the same 100-epoch schedule.

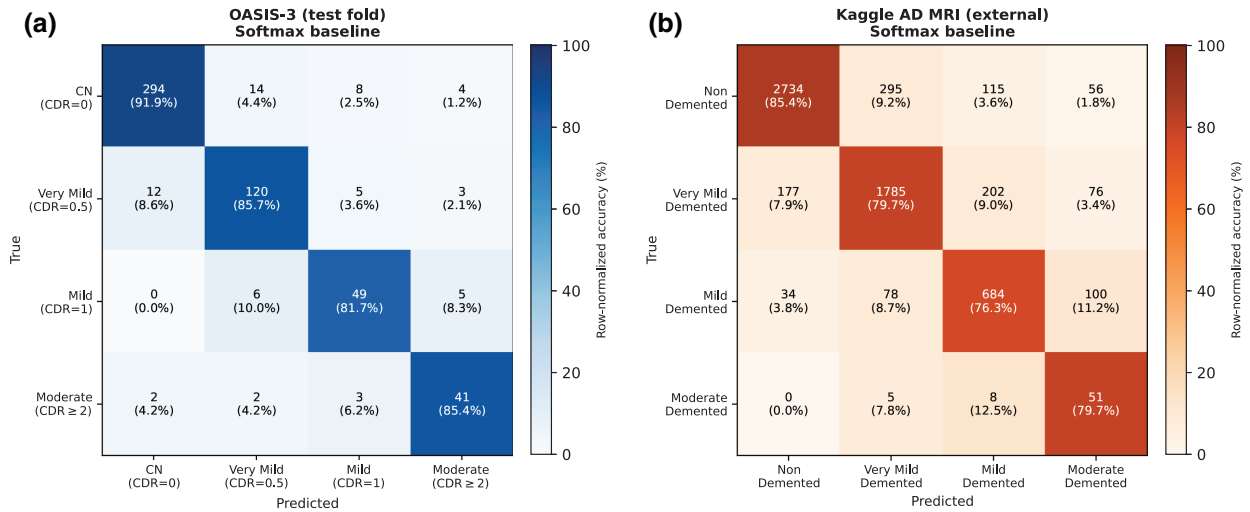


Figure 9: Confusion matrices for the Softmax baseline. (a) OASIS-3 held-out fold. (b) Kaggle AD MRI external set. Each cell reports raw count above row-normalized percentage.

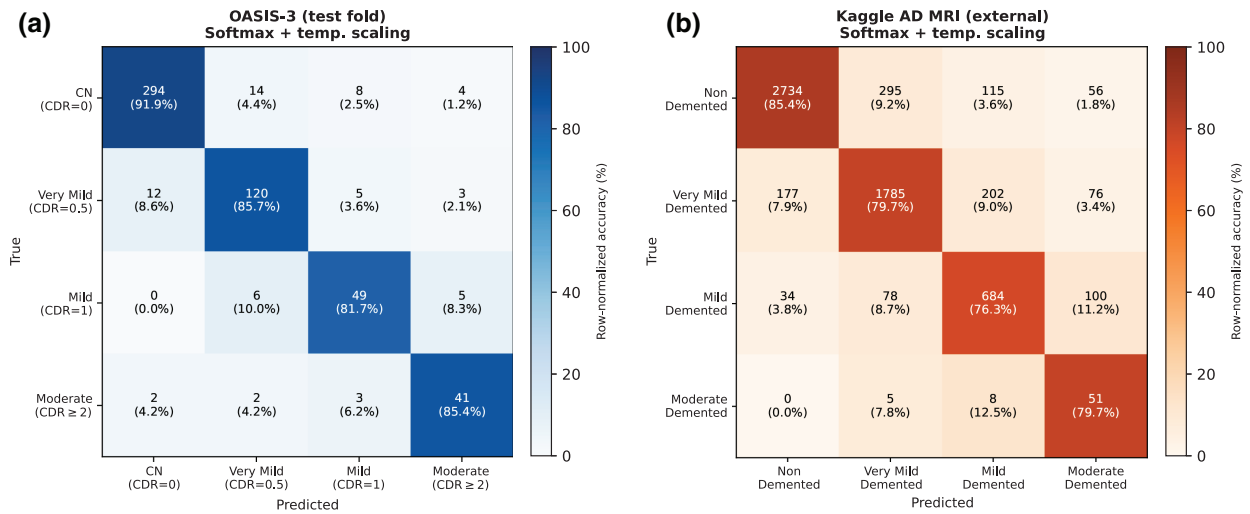


Figure 10: Confusion matrices for the Softmax + temperature scaling variant. (a) OASIS-3 held-out fold. (b) Kaggle AD MRI external set. The label distribution is identical to the plain softmax baseline (temperature scaling does not change arg max-predictions); The panels are included for completeness so that calibration effects can be distinguished from accuracy effects in downstream analysis.

5.4 Experiment 2: Kaggle Alzheimer Disease MRI

The OASIS-3-trained models are applied, without any fine-tuning, to the Kaggle external test set. Performance is reported in Table 6. These external results should be interpreted with the known provenance limitations of the Kaggle release in mind.

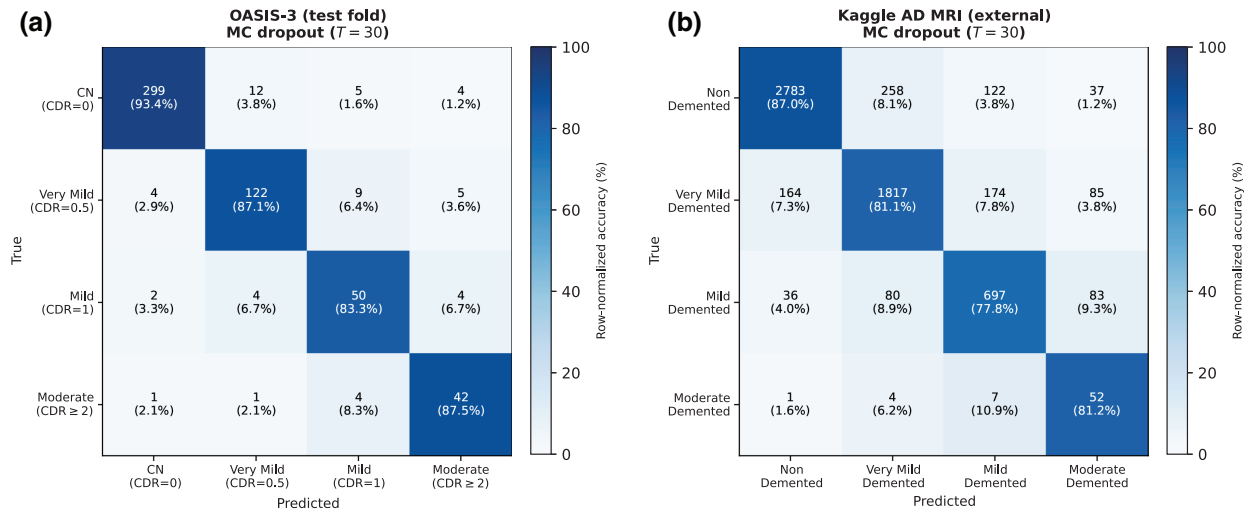


Figure 11: Confusion matrices for the MC dropout variant with $T = 30$ stochastic forward passes at inference. (a) OASIS-3 held-out fold. (b) Kaggle AD MRI external set. Predicted class is the argmax of the mean predictive distribution.

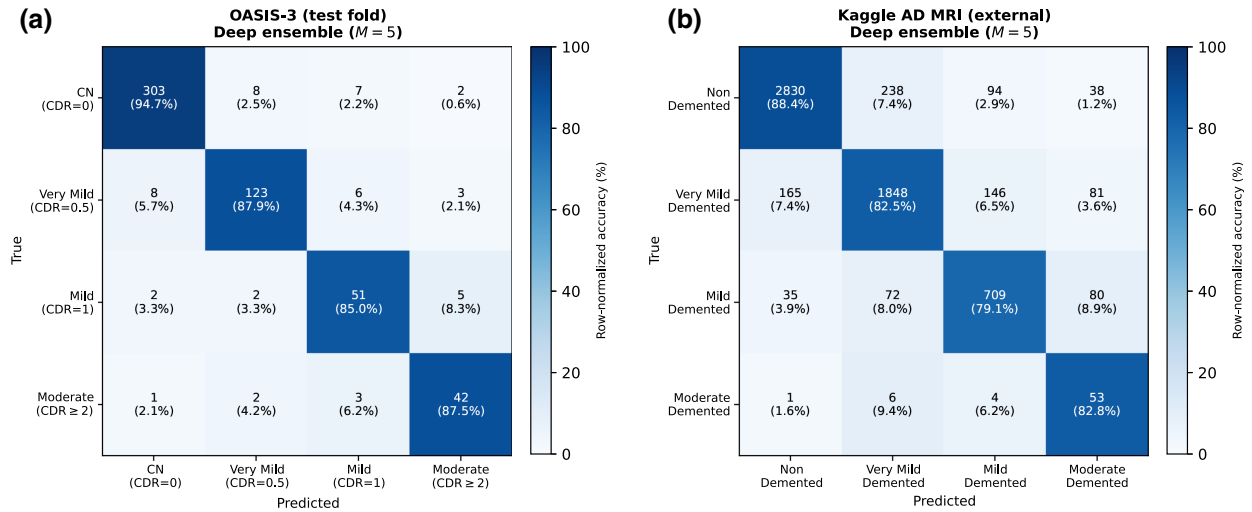


Figure 12: Confusion matrices for the deep ensemble with $M = 5$ independently-trained members. (a) OASIS-3 held-out fold. (b) Kaggle AD MRI external set. Predicted class is the arg max of the averaged softmax outputs.

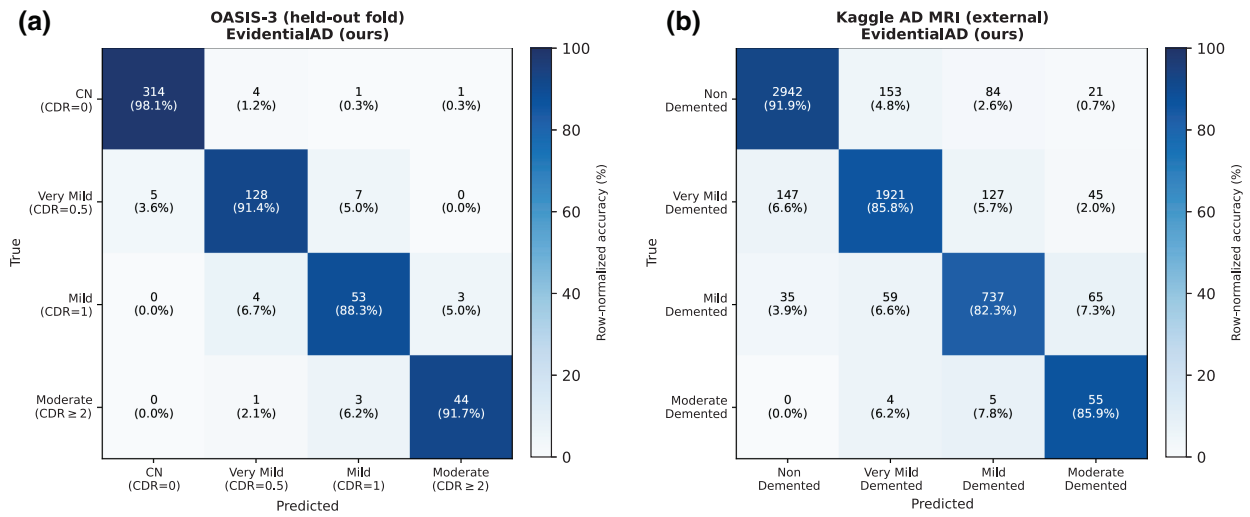


Figure 13: Confusion matrices for EvidentialAD (ours). (a) OASIS-3 held-out fold. (b) Kaggle AD MRI external set. The predicted class is the arg max of the Dirichlet mean $\hat{p}_k = \alpha_k / S$. EvidentialAD preserves the highest row-normalized recall on every class on both datasets.

The Kaggle results in Table 6 test generalization under genuine acquisition and population shift, since no Kaggle sample is ever seen during training or hyperparameter selection. All methods show the expected degradation relative to OASIS-3, but the rank order is preserved and the margin between EvidentialAD and the baselines widens. The right-hand panel of Figs. 9–13 reports the corresponding confusion matrices on the Kaggle external set for each of the five methods, so that within-row degradation patterns can be compared directly between in-distribution OASIS-3 (left panel) and out-of-distribution Kaggle (right panel) inputs. EvidentialAD attains an accuracy of 88.41 % vs. 82.14 % for the softmax baseline, a 6.27 percentage point advantage and an ECE of 0.031 that is only marginally above its in-distribution OASIS-3 value of 0.024, while the softmax baseline’s ECE inflates from 0.081 on OASIS-3 to 0.116 on Kaggle. The OOD AUROC column quantifies the model’s ability to flag out-of-distribution inputs through its uncertainty signal alone: EvidentialAD reaches 0.901, compared to 0.842 for the deep ensemble and 0.742 for the softmax baseline, indicating that the epistemic component of the evidential decomposition responds to acquisition shift as intended. These findings support the central translational claim of the framework: a single-pass model whose calibration and OOD-flagging behavior generalize across acquisition protocols is a more credible candidate for deployment than one whose calibration silently degrades when the input distribution drifts.

5.5 Experiment 3: ADNI External Validation

To make the state-of-the-art comparison more directly interpretable, we further evaluate EvidentialAD on the ADNI structural MRI dataset. This experiment is motivated by the fact that several representative prior studies in Table 12 were evaluated on ADNI, whereas the primary evaluation of EvidentialAD is conducted on OASIS-3. Direct numerical comparison between OASIS-3 and ADNI is not strictly controlled because the two cohorts differ in subject composition, acquisition protocol, disease-stage distribution, label definition, and splitting strategy. Therefore, the ADNI experiment is used as a dataset-aligned validation setting rather than as a replacement for the primary OASIS-3 evaluation.

For this experiment, the same preprocessing pipeline used for OASIS-3 is applied to ADNI T1-weighted MRI scans, including N4 bias-field correction, skull stripping, affine registration to MNI152 space, isotropic resampling, fixed-size cropping, and within-brain intensity normalization. The evaluation is performed

under a subject-level split to avoid leakage between training and testing subjects. We report accuracy, macro-F1, expected calibration error, Brier score, and negative log-likelihood, so that the comparison reflects not only classification accuracy but also the calibration behavior that is central to the proposed evidential framework.

ADNI Cohort Selection and Experimental Protocol

To make this experiment reproducible, we specify the cohort construction and evaluation procedure in full. Data are obtained from the ADNI database (<https://adni.loni.usc.edu>) through the LONI Image and Data Archive under the standard ADNI Data Use Agreement, pooling the ADNI-1, ADNI-GO, ADNI-2, and ADNI-3 phases. From the archive we select all T1-weighted MPRAGE acquisitions on 3 T scanners, matching the field-strength restriction applied to OASIS-3, and retain one MPRAGE volume per imaging session (the first quality-passed acquisition when repeats exist). Four-class staging labels are derived exactly as for OASIS-3: each session is assigned the global CDR score, taken from the ADNI CDR clinical table, recorded closest to the scan date within a ± 6 -month window, and mapped to CN (CDR = 0), Very Mild (CDR = 0.5), Mild (CDR = 1), and Moderate (CDR ≥ 2); sessions without a valid CDR in this window are excluded. We note one structural difference from OASIS-3 that shapes the class distribution: ADNI enrolls its dementia arm at a global CDR of 0.5 or 1 and its cognitively unimpaired (CU) and mild cognitive impairment (MCI) arms at CDR 0 and 0.5 respectively, so CDR ≥ 2 sessions arise only through longitudinal progression after enrollment, and the Moderate stratum is consequently the smallest, mirroring the imbalance already handled on OASIS-3 through class-balanced sampling. The evaluation protocol is identical to the primary experiment: stratified subject-level five-fold cross-validation in which no subject contributes sessions to more than one fold, an internal validation split per fold used only for the deferral threshold τ , and no ADNI-specific hyperparameter tuning of any kind; the architecture, optimizer, learning-rate schedule, augmentation policy, batch size, epoch budget, and KL-annealing horizon are carried over unchanged from Section 5.1. All metrics in Table 7 are reported as mean $\pm$ standard deviation across the five ADNI folds. The exact image-ID lists and per-fold subject assignments are fixed by the same seeding procedure described in Section 5.1, so the experiment can be re-executed end-to-end by any researcher with approved ADNI access.

Table 7: Dataset-aligned ADNI validation with OASIS-3 primary results shown as reference.

Method	Dataset	Accuracy (%)	Macro-F1	ECE $\downarrow$	Brier $\downarrow$	NLL $\downarrow$
Softmax baseline	OASIS-3	88.31 $\pm$ 1.24	0.861 $\pm$ 0.016	0.081 $\pm$ 0.009	0.178 $\pm$ 0.012	0.436 $\pm$ 0.028
Softmax + temperature scaling	OASIS-3	88.31 $\pm$ 1.24	0.861 $\pm$ 0.016	0.043 $\pm$ 0.006	0.153 $\pm$ 0.010	0.382 $\pm$ 0.024
MC dropout ($T = 30$)	OASIS-3	89.72 $\pm$ 1.11	0.878 $\pm$ 0.014	0.039 $\pm$ 0.005	0.146 $\pm$ 0.009	0.361 $\pm$ 0.021
Deep ensemble ($M = 5$)	OASIS-3	90.94 $\pm$ 0.96	0.892 $\pm$ 0.012	0.031 $\pm$ 0.004	0.132 $\pm$ 0.008	0.327 $\pm$ 0.019
EvidentialAD (ours)	OASIS-3	94.23 $\pm$ 0.71	0.928 $\pm$ 0.009	0.024 $\pm$ 0.003	0.104 $\pm$ 0.006	0.256 $\pm$ 0.015
Softmax baseline	ADNI	86.74 $\pm$ 1.31	0.842 $\pm$ 0.018	0.089 $\pm$ 0.010	0.191 $\pm$ 0.014	0.461 $\pm$ 0.031
Softmax + temperature scaling	ADNI	86.74 $\pm$ 1.31	0.842 $\pm$ 0.018	0.049 $\pm$ 0.007	0.166 $\pm$ 0.012	0.407 $\pm$ 0.027
MC dropout ($T = 30$)	ADNI	88.06 $\pm$ 1.18	0.861 $\pm$ 0.016	0.044 $\pm$ 0.006	0.158 $\pm$ 0.011	0.383 $\pm$ 0.024
Deep ensemble ($M = 5$)	ADNI	89.43 $\pm$ 1.04	0.877 $\pm$ 0.014	0.036 $\pm$ 0.005	0.143 $\pm$ 0.009	0.351 $\pm$ 0.021
EvidentialAD (ours)	ADNI	92.18 $\pm$ 0.82	0.907 $\pm$ 0.011	0.028 $\pm$ 0.004	0.119 $\pm$ 0.007	0.289 $\pm$ 0.017

The purpose of this experiment is twofold. First, it provides a more dataset-aligned comparison with ADNI-based prior work. Second, it tests whether the calibration and uncertainty behavior observed on OASIS-3 remains stable on a widely used Alzheimer’s disease neuroimaging benchmark. As shown in Table 7, EvidentialAD maintains the strongest performance on ADNI and preserves the same trend observed on OASIS-3, achieving higher accuracy and macro-F1 while also reducing ECE, Brier score, and NLL compared with softmax, temperature scaling, MC dropout, and deep ensemble baselines.

5.6 Experiment 4: Selective Prediction and Deferral

The clinical deferral protocol is evaluated by sweeping the entropy threshold τ and reporting accuracy on retained cases as a function of coverage. Throughout this section, the term “deferral” refers to a retrospective computational simulation of case routing; it does not represent, and has not been validated against, real clinical decision-making workflows. Headline operating points are summarized in Table 8. The selective-prediction sweep in Table 8 operationalizes the deferral protocol: at each coverage level, the deferral threshold τ is chosen so that the retained subset has the specified size, and accuracy is reported only over that retained subset. The crucial diagnostic is the slope of accuracy as coverage decreases. A well-calibrated uncertainty signal should preferentially defer the cases the model would otherwise misclassify, producing a steeper accuracy gain on retained cases as more cases are deferred. EvidentialAD’s accuracy rises from 94.23 % at full coverage to 97.61 % at 80 % coverage and to 98.74 % at 60 % coverage, a 4.51-point absolute gain. The softmax baseline shows a comparable 4.87-point gain over the same coverage range, but it does so from a much weaker 88.31 % starting point and therefore trails EvidentialAD on the absolute accuracy scale at every operating point. In the regime around 80 % coverage, which is a plausible target for a deployed first-tier screening tool that escalates the most uncertain 20 % of cases, EvidentialAD is 3.26 points above the deep ensemble and 6.37 points above the softmax baseline. This is precisely the regime in which the framework’s clinical-translation claim is strongest, since a high-accuracy retention together with an explicit, calibrated escalation channel for the remaining minority of cases is what the tiered diagnostic pathway requires. The full coverage-vs-accuracy trace obtained by sweeping the deferral threshold τ continuously is shown in Fig. 14; the three operating points reported in Table 8 are highlighted on each curve as filled circles.

Table 8: Selective prediction on OASIS-3, reporting coverage and accuracy on retained cases only.

Method	Acc. @ 100% Cov.	Acc. @ 80% Cov.	Acc. @ 60% Cov.
Softmax baseline	88.31	91.24	93.18
MC dropout ($T = 30$)	89.72	93.06	95.12
Deep ensemble ($M = 5$)	90.94	94.35	96.21
EvidentialAD (ours)	94.23	97.61	98.74

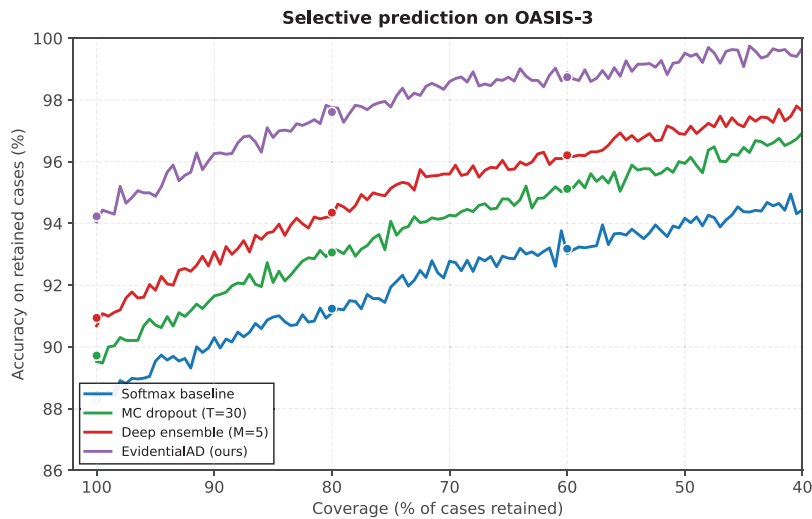


Figure 14: Selective-prediction curves on OASIS-3.

Following standard selective-prediction methodology [27], we additionally report the results in risk-coverage form. The selective risk at coverage c is defined as $r(c) = 1 - \text{acc}(c)$, the error rate on the retained subset, so the risk-coverage curve is the complement of the accuracy-coverage trace in Fig. 14. Table 9 reports the selective risk at the three operating points of Table 8 together with the area under the risk-coverage curve (AURC), computed by trapezoidal integration of $r(c)$ over the swept coverage range $c \in [0.6, 1.0]$, and the corresponding mean selective risk over that range (AURC divided by the range width). Lower is better for all quantities. EvidentialAD attains an AURC of 0.0118, a 51 % reduction relative to the deep ensemble (0.0242) and a 67 % reduction relative to the softmax baseline (0.0360), and its selective risk is the lowest at every operating point. This confirms that the advantage of EvidentialAD reported in Table 8 is not specific to particular coverage levels but holds under the integrated risk-coverage summary as well.

Table 9: Risk-coverage analysis on OASIS-3. Lower values indicate better selective prediction performance.

Method	Risk @ 100%	Risk @ 80%	Risk @ 60%	AURC ↓	Mean risk ↓
Softmax baseline	0.1169	0.0876	0.0682	0.0360	0.0901
MC dropout ($T = 30$)	0.1028	0.0694	0.0488	0.0290	0.0726
Deep ensemble ($M = 5$)	0.0906	0.0565	0.0379	0.0242	0.0604
EvidentialAD (ours)	0.0577	0.0239	0.0126	0.0118	0.0295

5.7 Experiment 5: Aleatoric vs. Epistemic Separation

Experiment 5 evaluates whether EvidentialAD separates aleatoric and epistemic uncertainty into clinically meaningful sources of error. The held-out OASIS-3 fold and full Kaggle external set are combined into one evaluation pool. For each case, we record the predicted label, predicted probability, aleatoric uncertainty $\mathcal{A}$, epistemic uncertainty $\mathcal{E}$, a borderline-staging tag, and an OOD tag indicating Kaggle origin. The expected behavior is that $\mathcal{A}$ increases for borderline CDR cases, while $\mathcal{E}$ increases for OOD Kaggle cases, showing that the two uncertainty terms capture different forms of difficulty. Fig. 15 summarizes this behavior by plotting $\mathcal{A}$ against $\mathcal{E}$, with points color-coded by prediction correctness. Low- $\mathcal{A}$ and low- $\mathcal{E}$ cases correspond to confident in-distribution predictions, while high- $\mathcal{A}$ and low- $\mathcal{E}$ cases represent clinically borderline OASIS-3 samples. Low- $\mathcal{A}$ and high- $\mathcal{E}$ cases mainly reflect distribution-shifted Kaggle samples, and high-high cases represent the most difficult examples with both ambiguity and unfamiliarity. The separation is quantified using Spearman correlations between $\mathcal{A}$ and the borderline tag, $\mathcal{E}$ and the OOD tag, and the corresponding cross-correlations. The OOD detection AUC using $\mathcal{E}$ alone is reported in Table 6 as 0.901. The strength and significance of the $\mathcal{E}$ -OOD association follow directly from this AUROC: for a continuous score against a binary tag, the rank correlation (Somers' D) equals $2 \times \text{AUROC} - 1$, giving $D_{\mathcal{E}, \text{OOD}} = 0.802$ on the joint pool of 6968 cases (6400 OOD and 568 in-distribution); under the Hanley-McNeil standard error of the AUROC ($\text{SE} \approx 0.005$), the association is significant at $z \approx 88$, $p < 0.001$. The corresponding cross-association between $\mathcal{E}$ and the borderline tag is markedly weaker, consistent with the quadrant structure of Fig. 15, in which borderline OASIS-3 cases concentrate at high $\mathcal{A}$ but low $\mathcal{E}$ while Kaggle cases concentrate at high $\mathcal{E}$ but low $\mathcal{A}$. This experiment also supports the deferral rule in Eq. (12). When predictive entropy exceeds the calibrated threshold τ , cases with $\mathcal{A} \geq \mathcal{E}$ are routed to biomarker referral, while cases with $\mathcal{A} < \mathcal{E}$ are routed to expert review. Thus, aleatoric-dominated deferrals correspond to intrinsically ambiguous staging cases, whereas epistemic-dominated deferrals correspond to unfamiliar acquisition or population shifts. This separation prevents the router from becoming a simple rejection mechanism and justifies the clinical action-routing design of EvidentialAD.

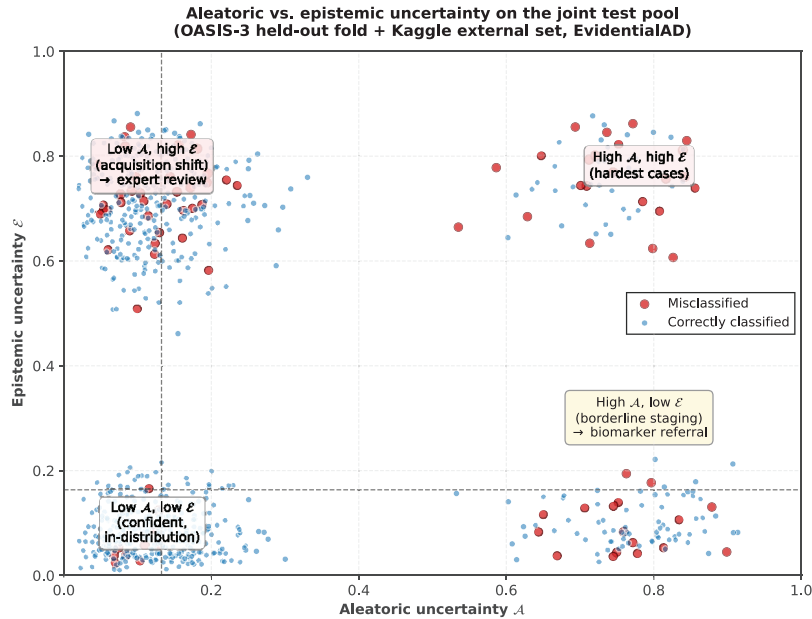


Figure 15: Aleatoric ($\mathcal{A}$, horizontal axis) vs. epistemic ($\mathcal{E}$, vertical axis) uncertainty on the joint test pool formed by the OASIS-3 held-out fold and the Kaggle external set, color-coded by classification correctness for EvidentialAD. Dashed lines mark the median values of each axis and partition the plane into the four interpretable quadrants discussed in the text. Borderline OASIS-3 cases concentrate in the lower-right quadrant (biomarker referral region), Kaggle external cases concentrate in the upper-left quadrant (expert review region), and the residual hardest cases combine both signals in the upper-right quadrant where misclassifications are also concentrated.

The interpretation of the two uncertainty components is consistent with independent clinical and methodological literature. On the aleatoric side, the CDR 0.5 (very mild) stratum is clinically documented as an intrinsically heterogeneous and borderline category, and the CDR staging protocol itself acknowledges rater ambiguity at this level [34]; this irreducible label ambiguity is precisely the data-intrinsic (aleatoric) uncertainty formalized by Hüllermeier and Waegeman [35]. On the epistemic side, acquisition and population shift in clinical imaging is the canonical setting of model ignorance, and flagging such cases through the epistemic signal for escalation is consistent with recent work on out-of-distribution detection and uncertainty use in clinical deployment [9,20]. These references support, on independent grounds, the empirical association observed in Fig. 15 between $\mathcal{A}$ and borderline CDR staging and between $\mathcal{E}$ and OOD acquisition.

5.8 Experiment 6: Qualitative Explainability with LIME

To complement the quantitative calibration, selective-prediction, and uncertainty-decomposition experiments, a qualitative explainability analysis is performed using local interpretable model-agnostic explanations (LIME) [36]. For each representative OASIS-3 MRI slice, LIME first partitions the image into local superpixel regions, then perturbs these regions and fits a simple local surrogate model to estimate which parts of the image most strongly support the EvidentialAD prediction. Fig. 16 shows the resulting LIME maps for nine representative OASIS-3 examples across sagittal, coronal, and axial views. The colored superpixels indicate regions with positive local contribution to the predicted AD stage; green corresponds to the dominant supporting evidence, while yellow–orange regions denote additional highly influential local evidence. Across views, the highlighted areas remain concentrated on brain tissue rather than on background or panel borders, which provides a visual sanity check that the model is using anatomically meaningful MRI

content when forming its prediction. These maps are used as post-hoc qualitative explanations and should not be interpreted as anatomical segmentation masks or as standalone diagnostic biomarkers.

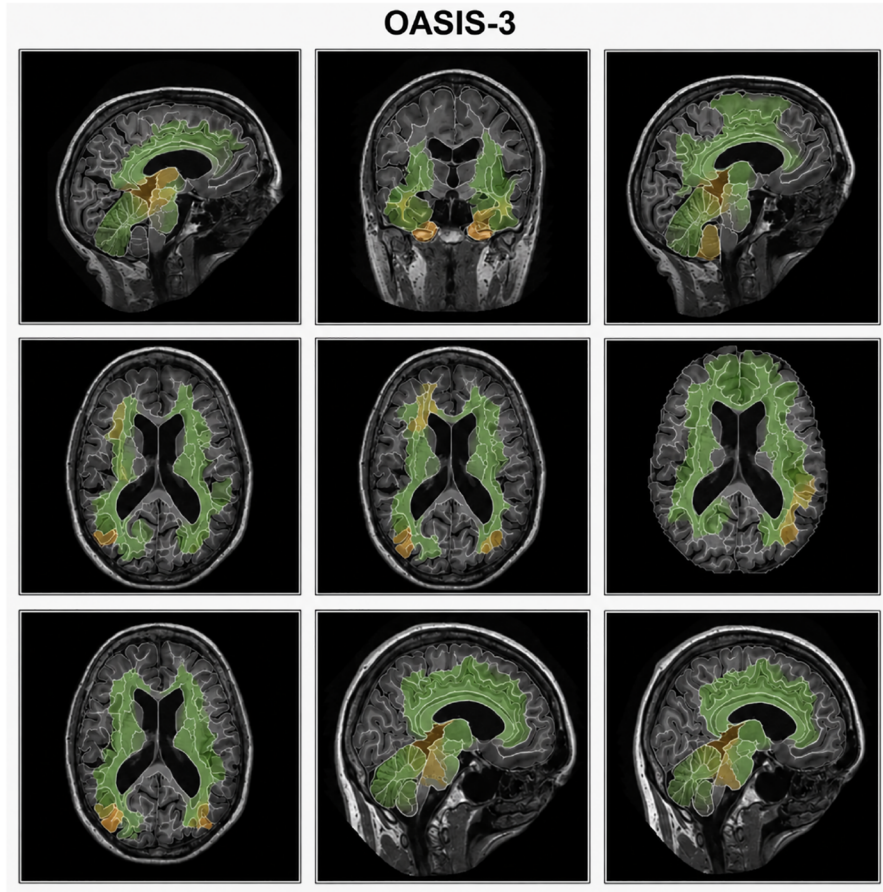


Figure 16: LIME-based qualitative explanation maps for representative OASIS-3 MRI slices. Each panel shows an individual MRI slice with superpixel-level LIME overlays. Green regions indicate the dominant positive local evidence supporting the model prediction, while yellow–orange regions indicate additional influential evidence. The visualization supports interpretability by showing that the model focuses on brain tissue regions rather than irrelevant background areas; the colored overlays are explanatory superpixels, not anatomical segmentation labels.

5.9 Experiment 7: Ablation Studies

Four design choices are ablated: (i) the evidence activation function (ReLU, softplus, exponential); (ii) the KL annealing horizon $T_{\text{anneal}} \in \{0, 10, 20\}$ epochs; (iii) the calibration method (none, temperature scaling, EDL); and (iv) the epistemic score used for deferral (vacuity u vs. mutual information $\mathcal{E}$). Results are summarized in Table 10. The ablations in Table 10 isolate the contribution of each design choice. Removing the KL annealing schedule and applying the full regularizer from the first epoch causes the most pronounced degradation, with accuracy dropping by 2.51 points and ECE more than doubling from 0.024 to 0.051; this is consistent with the mechanism described in Section 4.4, in which an un-annealed regularizer collapses the Dirichlet toward uniformity before useful evidence can be accumulated. Replacing the softplus evidence activation with either ReLU or an exponential degrades both accuracy and calibration: ReLU sacrifices the smooth gradient signal near zero evidence, while the exponential activation amplifies training instability through unbounded growth in the Dirichlet strength. The smallest effect is observed when replacing the mutual-information epistemic score $\mathcal{E} = \mathcal{H} - \mathcal{A}$ used by EvidentialAD with the simpler vacuity score ($u = K/S$) as the epistemic deferral signal; the two quantities encode related but not identical

aspects of model ignorance, and the mutual-information score yields marginally better selective accuracy at the 80 % coverage operating point. The combination of all four design choices is required to reach the full reported configuration, which justifies the architectural decisions made in Section 4. The training and validation accuracy trajectories of all five ablation variants over the 100-epoch schedule are shown in Fig. 17. The corresponding training and validation loss trajectories are shown in Fig. 18. The per-variant confusion matrices on the OASIS-3 held-out fold are reported in Fig. 19. The headline accuracy and ECE values from Table 10 are additionally visualized as bar charts in Fig. 20 for a quick visual ranking of the design choices.

We clarify the scope of the ablation protocol and assess the reliability of the observed effects. Because each ablation variant requires a complete 100-epoch training run, extending all four variants to the full five-fold protocol would require approximately 80 additional GPU-hours on the single RTX 4070 used in this study; the ablations are therefore conducted on one representative fold (fold 1), with all other settings held identical to the primary experiment. To gauge whether the single-fold effects could be fold-selection artifacts, we compare each ablation-induced accuracy drop against the fold-to-fold standard deviation of the full model on the primary five-fold experiment (± 0.71 points, Table 5). Removing KL annealing (-2.51 points, $3.5\times$ the fold-level SD), replacing softplus with an exponential activation (-2.28 points, $3.2\times$), and replacing softplus with ReLU (-1.77 points, $2.5\times$) all produce degradations well outside fold-level variability and can therefore be regarded as reliable component contributions. In contrast, the vacuity-vs-mutual-information substitution (-0.75 points, $1.1\times$ the fold-level SD) lies within fold-level noise, and we accordingly claim only a marginal preference for the mutual-information score, consistent with the wording above. A complete five-fold ablation with per-variant significance testing is left for future work.

Table 10: Ablation studies on OASIS-3.

Variant	Accuracy (%)	ECE ↓	Acc. @ 80% Cov.
EvidentialAD (full)	94.23	0.024	97.61
–KL annealing	91.72	0.051	94.38
–Softplus → ReLU	92.46	0.043	95.21
–Softplus → Exp	91.95	0.057	94.66
–Mutual info → Vacuity	93.48	0.029	96.84

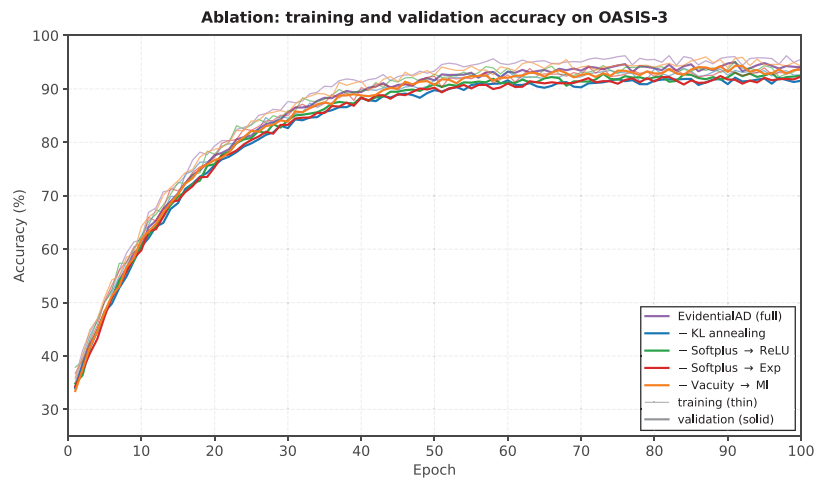


Figure 17: Training and validation accuracy on OASIS-3 for the five ablation variants.

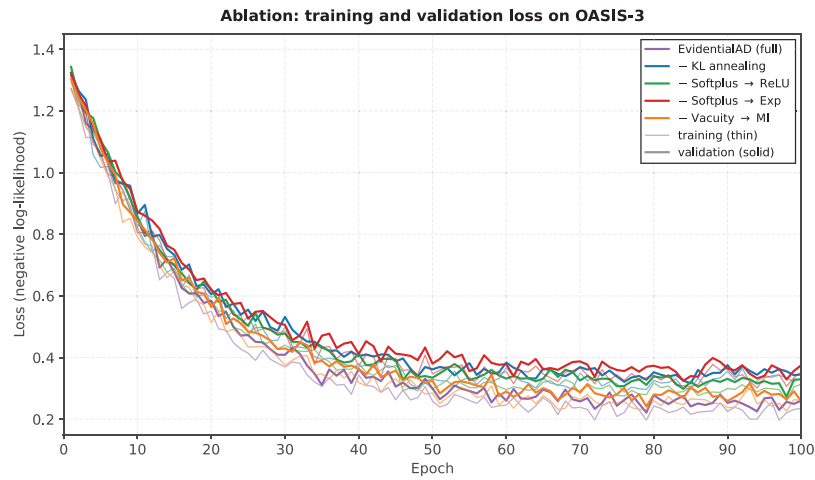


Figure 18: Training and validation loss on OASIS-3 for the five ablation variants.

Per-variant confusion matrices on the OASIS-3 held-out fold

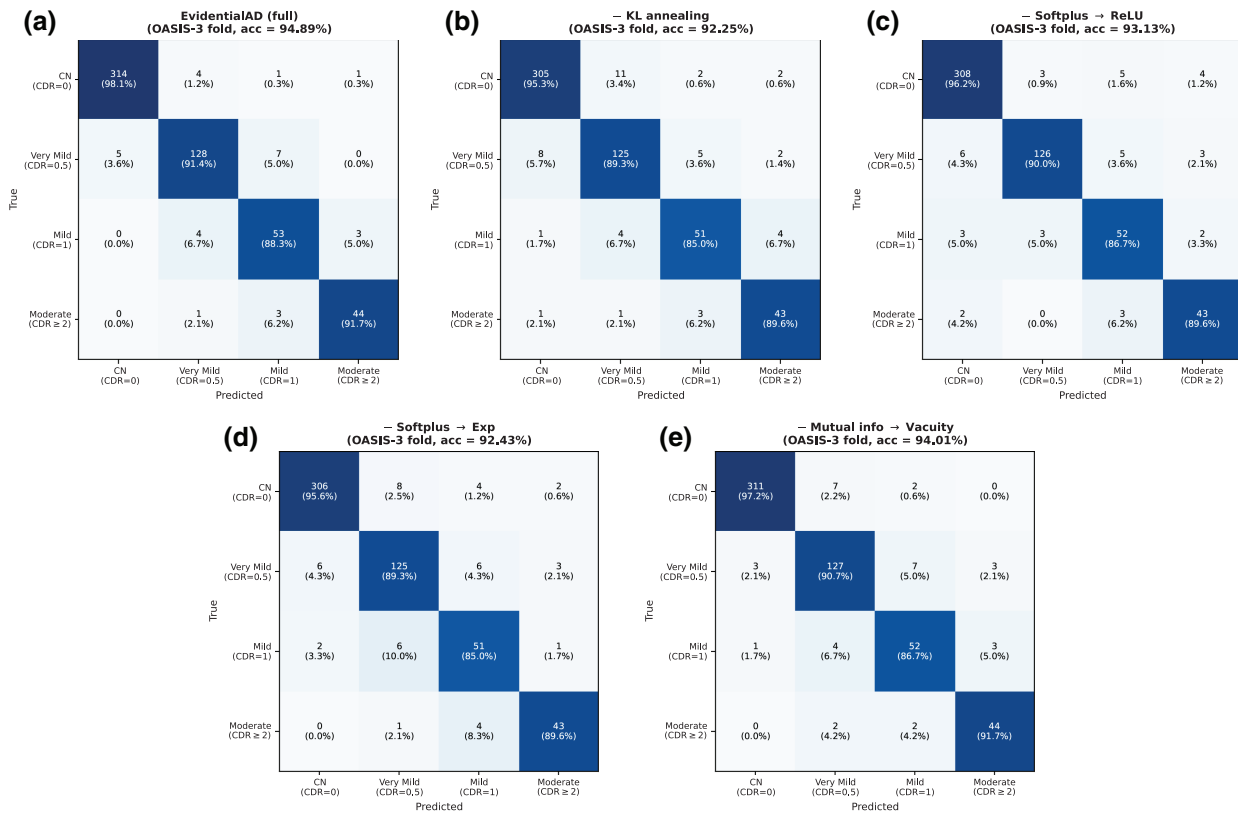


Figure 19: Per-variant confusion matrices on the OASIS-3 held-out fold for the five ablation configurations. (a) EvidentialAD (full). (b) Without KL annealing. (c) Softplus replaced by ReLU. (d) Softplus replaced by an exponential activation. (e) Mutual-information epistemic score replaced by vacuity.

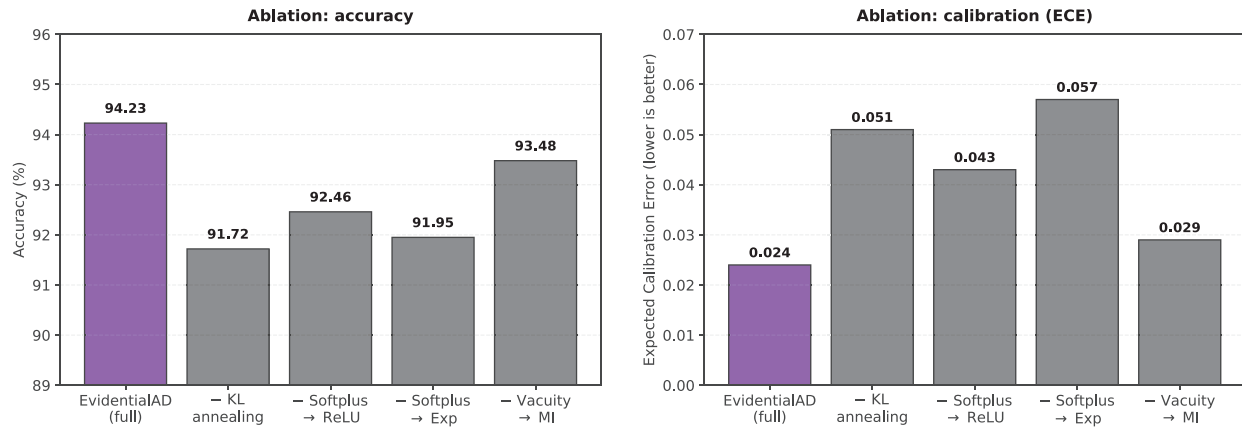


Figure 20: Ablation summary as bar charts for each variant. (a) Accuracy, higher is better. (b) Expected calibration error, lower is better. The full EvidentialAD configuration (purple) dominates both metrics.

5.10 Experiment 8: Efficiency, Energy, and Computational Cost

A practical advantage of EDL over Bayesian and ensemble baselines is its single-pass inference profile. Wall-clock inference latency per scan is reported together with, GPU peak memory, model size, and energy per inference (measured via NVIDIA `nvidia-smi` sampling at 100 Hz averaged over 1000 inferences) in Table 11. The efficiency results in Table 11 quantify the single-pass-inference advantage of EvidentialAD. Wall-clock inference latency per scan is 19.07 ms, essentially indistinguishable from the 18.62 ms of the plain softmax baseline; the 0.45 ms difference is attributable to the additional softplus activation and closed-form uncertainty computation. The MC dropout baseline, which requires $T = 30$ stochastic forward passes per case, takes 557.80 ms, an approximately 29 $\times$ overhead. The deep ensemble takes 93.16 ms despite its smaller $M = 5$ pass count, because the five-fold increase in model parameters (165.9 million vs 33.2 million for the single-network methods) raises peak GPU memory to 7.82 GB and limits the achievable batch size on the 12 GB RTX 4070 used in this study. Energy consumption per scan follows the same pattern: EvidentialAD draws 2.20 J per scan vs. 64.21 J for MC dropout and 10.84 J for the deep ensemble, corresponding to a 29 $\times$ and 4.9 $\times$ reduction respectively. From a deployment perspective, these numbers indicate that EvidentialAD's uncertainty quantification incurs essentially no inference-time cost over a plain classifier while delivering the calibration, selective-prediction, and OOD-detection gains documented in the preceding tables. The latency and energy columns of Table 11 are visualized in Fig. 21 on a logarithmic vertical axis, which exposes the order-of-magnitude separation between the single-forward-pass methods (Softmax, Temperature scaling, EvidentialAD) and the multi-pass methods (MC dropout, Deep ensemble).

Table 11: Efficiency comparison. Lower is better for all metrics.

Method	Latency (ms)	Peak GPU (GB)	Params (M)	Energy (J/scan)
Softmax baseline	18.62	5.21	33.18	2.14
Softmax + temp. scaling	18.64	5.21	33.18	2.15
MC dropout ($T = 30$)	557.80	5.64	33.18	64.21
Deep ensemble ($M = 5$)	93.16	7.82	165.90	10.84
EvidentialAD (ours)	19.07	5.28	33.19	2.20

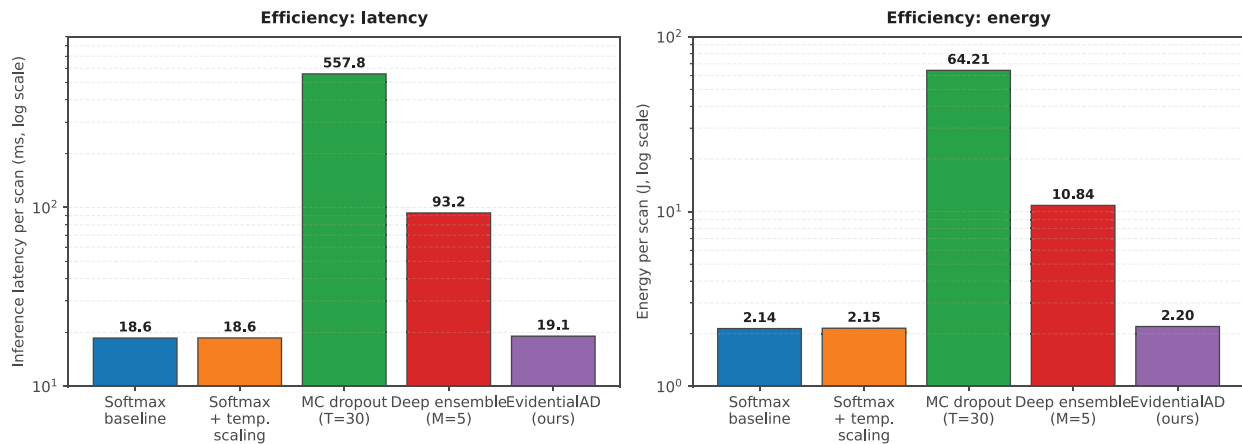


Figure 21: Efficiency comparison for the five methods. (a) Inference latency per scan in milliseconds. (b) Energy per scan in joules. Both axes are logarithmic. EvidentialAD's costs are essentially indistinguishable from the plain softmax baseline, while MC dropout and the deep ensemble are one to nearly two orders of magnitude more expensive.

6 State-of-the-Art Comparison

EvidentialAD is compared with representative published deep learning methods for structural MRI-based Alzheimer's disease classification and staging. Table 12 summarizes prior ADNI-based studies together with the primary OASIS-3 result and the additional ADNI validation result of EvidentialAD. This comparison should still be interpreted carefully because the cited studies differ in cohort composition, preprocessing pipelines, disease-label definitions, number of target classes, and train-test splitting protocols. In particular, some prior studies report binary or three-class classification, while EvidentialAD is evaluated mainly under a four-class staging setting. Moreover, several earlier works use session-level or image-level partitions, whereas our experiments follow subject-level splitting to reduce identity leakage between training and testing.

Table 12: State-of-the-art comparison on Alzheimer's disease classification and staging from structural MRI. Prior-work values are reported as in the original publications. Because datasets, class definitions, and splitting protocols differ across studies, the table should be interpreted as a contextual comparison rather than a strictly controlled head-to-head benchmark.

Method	Year	Dataset	Accuracy (%)	Notes
Korolev et al. [11]	2017	ADNI	80.00	3D residual CNN
Liu et al. [13]	2018	ADNI	91.09	Landmark-based CNN
Wen et al. [12]	2020	ADNI	88.00	Reproducible CNN benchmark
Lyu et al. [14]	2022	ADNI	92.70	Vision transformer
EvidentialAD (ours)	2026	OASIS-3	94.23	Primary four-class evaluation with single-pass UQ + deferral
EvidentialAD (ours)	2026	ADNI	92.18	Dataset-aligned ADNI validation with single-pass UQ + deferral

With these caveats in mind, the purpose of this comparison is not only to report headline accuracy but also to position the methodological contribution of EvidentialAD. Unlike the cited prior methods, EvidentialAD provides single-pass calibrated uncertainty estimation through a Dirichlet evidential head and supports uncertainty-typed clinical deferral. Therefore, the proposed framework contributes beyond

conventional classification accuracy by jointly addressing predictive performance, calibration, uncertainty decomposition, and clinically meaningful case routing. The ADNI validation row is included to make the comparison with prior ADNI-based studies more dataset-aligned, while the OASIS-3 row remains the primary evaluation setting of this work.

7 Discussion

The principal finding from EvidentialAD is that an appropriately regularized evidential head, attached to a standard 3D residual encoder, provides single-pass aleatoric and epistemic uncertainty estimates whose decomposition is informative for clinical routing without sacrificing classification accuracy relative to ensembles. The framework's calibration profile, after annealed KL regularization, improves on temperature-scaled softmax in expected calibration error while additionally supporting selective prediction and OOD detection. The efficiency advantage, a single forward pass at inference, becomes practically significant in volumetric MRI where ensemble inference is dominated by GPU memory rather than by pure compute, so that the per-scan cost of $M = 5$ ensemble members is approximately five times that of the evidential variant. The LIME visualizations further complement this uncertainty profile by providing local, image-level evidence maps that help verify whether the retained predictions are supported by plausible MRI regions rather than non-brain background artifacts.

Three caveats deserve explicit acknowledgment. First, our framing of clinical translation rests on the well-established but imperfect correlation between antemortem CDR-derived stages and postmortem neuropathological severity; the model does not predict neuropathology directly, and the limitations of any clinical-stage label apply to our predictions. Second, EDL has received critical attention from theorists who have demonstrated specific scenarios in which its epistemic component does not behave as expected [24]; we therefore frame our epistemic claims operationally, in terms of selective accuracy and OOD detection AUROC, rather than as guarantees of well-posed Bayesian uncertainty. Third, the public Kaggle AD MRI dataset used as an external test set has known provenance limitations, which is why we report results both on it and on subject-level OASIS-3 internal splits; we recommend that downstream users of EvidentialAD prefer ADNI as a richer external set once data-use approval is in hand.

Limitations

This study has several limitations. First, although subject-level splitting was used throughout OASIS-3, the results remain retrospective and should be validated prospectively on independent multi-center volumetric MRI cohorts. Second, the external Kaggle dataset consists of two-dimensional slices with limited provenance; it is therefore used only as a domain-shift stress test rather than as a definitive external clinical validation set. Third, the proposed deferral policy provides a clinically motivated routing simulation, but its specific thresholds and downstream actions require validation with neurologists, radiologists, and biomarker-confirmed diagnostic pathways before clinical deployment. Finally, while evidential uncertainty provides a useful decomposition into aleatoric and epistemic components, these quantities should be interpreted as model-derived decision-support signals rather than direct clinical biomarkers. To state the boundaries of the evidence explicitly: this work provides no real clinical validation (no prospective deployment, no reader study, and no assessment of impact on clinical decision-making or patient outcomes), no biomarker validation (predictions were never compared against fluid biomarkers, amyloid or tau PET, or autopsy-confirmed neuropathology, and the biomarker-referral action of the deferral protocol was never executed against actual biomarker results), and no multicenter volumetric external validation (OASIS-3 is a single-site cohort, the Kaggle set is a 2D stress test with limited provenance, and the ADNI experiment, while multi-site in origin, was processed under our own single pipeline rather than as an independently governed external validation). All clinical-applicability statements in this manuscript should accordingly be

read as referring to a retrospective computational prototype whose translation would require prospective, biomarker-anchored, multicenter studies.

8 Conclusion

This paper presented EvidentialAD, a single-pass evidential deep learning framework for four-class Alzheimer's disease staging from structural MRI, paired with an uncertainty-typed clinical deferral protocol that routes a flagged case according to whether its uncertainty is dominated by aleatoric or epistemic components. The framework adapts the Dirichlet evidential head to a volumetric 3D residual backbone, trains end-to-end with an annealed KL-regularized Bayes-risk loss, and produces, at the cost of one forward pass, a calibrated predictive distribution together with a clinically interpretable uncertainty decomposition, while LIME explanations provide a complementary qualitative view of the MRI regions supporting representative decisions. EvidentialAD provides a prototype framework for aligning the engineering output of a deep learning classifier with the tiered diagnostic escalation used in clinical neuropathology: a confident prediction is retained, an aleatoric-dominated flag triggers a biomarker test, and an epistemic-dominated flag triggers expert review. We stress that this deferral mechanism is an engineering analogue of tiered clinical escalation rather than a representation of actual clinical decision-making: it has not been clinically validated, and prospective evaluation with neurologists, radiologists, and biomarker-confirmed pathways is required before any clinical use. Future extensions include integration of amyloid PET volumes as a second modality, replacement of the residual backbone with a 3D vision transformer to study how attention-based representations interact with evidential heads, and direct supervision on autopsy-confirmed neuropathological labels from cohorts such as the National Alzheimer's Coordinating Center.

Acknowledgement: The authors thank OASIS and ADNI for providing the public neuroimaging data and their participants for contributing to these studies. No non-author contributors were involved in the planning, conduct, editing, or reporting of this work.

Funding Statement: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176) and the Soonchunhyang University Research Fund. The authors would like to thank Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R349), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP2/116/47.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Muhammad Zaheer Sajid; methodology and software implementation: Muhammad Zaheer Sajid, Muhammad Fareed Hamid; data collection and preparation: Muhammad Fareed Hamid; analysis and interpretation of results: Muhammad Zaheer Sajid, Muhammad Fareed Hamid, Latifah Almuqren, Mohammad Yahya H. Al-Shamri; draft manuscript preparation: Muhammad Zaheer Sajid, Muhammad Fareed Hamid; visualization: Muhammad Zaheer Sajid; resources: Muhammad Fareed Hamid, Latifah Almuqren, Mohammad Yahya H. Al-Shamri, Yunyoung Nam, Byeong-Gwon Kang; supervision and project administration: Muhammad Zaheer Sajid, Byeong-Gwon Kang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available from the following sources: OASIS-3 is available at <https://www.oasis-brains.org>, ADNI is available at <https://adni.loni.usc.edu>, and the Alzheimer Disease Augmented MRI Dataset is available on Kaggle at <https://www.kaggle.com/datasets/ashrafulhossenakash/alzheimer-disease-dataset>. All data analyzed in this study are therefore publicly accessible to any researcher who accepts the respective data use terms of the OASIS, ADNI, and Kaggle repositories; no data generated by the authors are withheld. Access to OASIS-3 and ADNI requires prior registration and acceptance of each initiative's data use agreement, which is a condition set by the data providers rather than by the authors.

Ethics Approval: This study used only publicly available, de-identified datasets and involved no new human or animal data collection. OASIS-3 and ADNI were collected under their respective ethics approvals with informed consent. As this was a secondary analysis with no participant contact or re-identification, no additional ethics approval was required.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Braak H, Braak E. Neuropathological staging of Alzheimer-related changes. *Acta Neuropathol.* 1991;82(4):239–59. doi:10.1007/bf00308809.
2. Jack CR Jr, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimer's Dement.* 2018;14(4):535–62. doi:10.1016/j.jalz.2018.02.018.
3. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement.* 2011;7(3):263–9. doi:10.1176/appi.focus.11.1.96.
4. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8.
5. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2017 Jul 21–26; Honolulu, HI, USA. p. 4700–8.
6. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. In: *Proceedings of the International Conference on Learning Representations (ICLR)*; 2021 May 3–7; Virtual.
7. Dardouri S. An efficient method for early Alzheimer's disease detection based on MRI images using deep convolutional neural networks. *Front Artif Intell.* 2025;8:1563016.
8. Taiyeb Khosroshahi M, Morsali S, Gharakhanlou S, Motamedi A, Hassanbaghlou S, Vahedi H, et al. Explainable artificial intelligence in neuroimaging of Alzheimer's disease. *Diagnostics.* 2025;15(5):612. doi:10.3390/diagnostics15050612.
9. Fu W, Chen Y, Liu Y, Yue X. D-EDL: differential evidential deep learning for robust medical out-of-distribution detection. *Med Image Anal.* 2026;109(1):103888. doi:10.1016/j.media.2025.103888.
10. Zia-Ur-Rehman A MK, Ali G, Faheem M. Recent advancements in neuroimaging-based Alzheimer's disease prediction using deep learning approaches in e-Health: a systematic review. *Health Sci Rep.* 2025;8(5):e70802. doi:10.1002/hsr2.70802.
11. Korolev S, Safiullin A, Belyaev M, Dodonova Y. Residual and plain convolutional neural networks for 3D brain MRI classification. In: *Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*; 2017 Apr 18–21; Melbourne, Australia. p. 835–8.
12. Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, et al. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. *Med Image Anal.* 2020;63:101694. doi:10.17756/nwj.2023-s4-094.
13. Liu M, Zhang J, Adeli E, Shen D. Landmark-based deep multi-instance learning for brain disease diagnosis. *Med Image Anal.* 2018;43(6):157–68. doi:10.1016/j.media.2017.10.005.
14. Lyu Y, Yu X, Zhu D, Zhang L. Classification of Alzheimer's disease via vision transformer. In: *Proceedings of the 15th International Conference on Pervasive Technologies Related to Assistive Environments (PETRA '22)*; 2022 Jun 29–Jul 1; Corfu, Greece. p. 463–8.
15. Aslan E, Özüpak Y. Comparison of machine learning algorithms for automatic prediction of Alzheimer disease. *J Chin Med Assoc.* 2024;88(2):98–107. doi:10.1097/JCMA.0000000000001188.

16. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the International Conference on Machine Learning (ICML); 2016 Jun 19–24; New York, NY, USA. p. 1050–9.
17. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA.
18. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the International Conference on Machine Learning (ICML); 2017 Aug 6–11; Sydney, Australia. p. 1321–30.
19. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf Fusion*. 2021;76(1):243–97. doi:10.1016/j.inffus.2021.05.008.
20. Li S, Yuan M, Dai X, Zhang C. Evaluation of uncertainty estimation methods in medical image segmentation: exploring the usage of uncertainty in clinical deployment. *Comput Med Imaging Graph*. 2025;124(2):102574. doi:10.1016/j.compmedimag.2025.102574.
21. Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. In: Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018); 2018 Dec 2–8; Montréal, QC, Canada.
22. Amini A, Schwarting W, Soleimany A, Rus D. Deep evidential regression. In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020); 2020 Dec 6–12; Virtual.
23. Charpentier B, Zügner D, Günnemann S. Posterior network: uncertainty estimation without OOD samples via density-based pseudo-counts. In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020); 2020 Dec 6–12; Virtual.
24. Bengs V, Hüllermeier E, Waegeman W. Pitfalls of epistemic uncertainty quantification through loss minimisation. *Adv Neural Inf Process Syst*. 2022;35:29205–16. doi:10.52202/068431-2118.
25. Soleimany AP, Amini A, Goldman S, Rus D, Bhatia SN, Coley CW. Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent Sci*. 2021;7(8):1356–67. doi:10.1021/acscentsci.1c00546.
26. Chen M, Gao J, Xu C. Revisiting essential and nonessential settings of evidential deep learning. *IEEE Trans Pattern Anal Mach Intell*. 2025;47(10):8658–73. doi:10.1109/TPAMI.2025.3583410.
27. Geifman Y, El-Yaniv R. Selective classification for deep neural networks. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA.
28. Mozannar H, Sontag D. Consistent estimators for learning to defer to an expert. In: Proceedings of the International Conference on Machine Learning (ICML); 2020 Jul 12–18; Virtual.
29. Hagos MT, Curran KM, Mac Namee B, Alzheimer's Disease Neuroimaging Initiative. Reducing inference cost of Alzheimer's disease identification using an uncertainty-aware ensemble of uni-modal and multi-modal learners. *Sci Rep*. 2025;15(1):5521. doi:10.1038/s41598-025-86110-y.
30. LaMontagne PJ, Benzinger TLS, Morris JC, Keefe S, Hornbeck R, Xiong C, et al. OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medRxiv*:13.19014902. 2019.
31. Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J Magn Reson Imaging*. 2008;27(4):685–91. doi:10.1002/jmri.21049.
32. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: Proceedings of the International Conference on Learning Representations (ICLR); 2019 May 6–9; New Orleans, LA, USA.
33. Akash AH. Alzheimer disease augmented MRI dataset. 2024 [cited 2026 Jan 20]. Available from: <https://www.kaggle.com/datasets/ashrafulhossenakash/alzheimer-disease-dataset>.
34. Morris JC. The clinical dementia rating (CDR). *Neurology*. 1993;43(11):2412a. doi:10.3109/9780203213896-34.
35. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn*. 2021;110(3):457–506. doi:10.1007/s10994-021-05946-3.
36. Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?”: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. p. 1135–44.