**Computer Science and Information Technologies**
Vol. 99, No. 1, April 2026, pp. 1∼1x
ISSN: 2722-3221, DOI: 10.11591/csit.v99i1.pp1-1x 

# BrainMD:Explainable Hybrid Liquid AI for Robust Neurodiagnostics in Low-Resource Areas

**Kshiraja Nelapati[1], Meeradevi AK[2], Prathmesh Sayal[3]**
[1]Department of Artificial Intelligence and Machine Learning, Ramaiah Institute of Technology, Bengaluru, India
[2]Department of Artificial Intelligence and Machine Learning, Ramaiah Institute of Technology, Bengaluru, India
[3]Department of Computer Science and Engineering (Cyber Security), Ramaiah Institute of Technology, Bengaluru, India

**Article Info**

***Article history:***

Received month dd, yyyy
Revised month dd, yyyy
Accepted month dd, yyyy

***Keywords:***

Medical Image Segmentation
Selective Classification
Uncertainty Quantification
Neuroimaging Triage
Clinical Decision Support

**ABSTRACT**

Neurodiagnostic systems required a safe way to avoid unnecessary specialist computation without allowing uncertain scans to exit as normal. This study proposed BrainMD, a three-stage workflow that combined confidence-gated screening, reject-option specialist routing, pathology-specific segmentation, and evidence-preserving review for tumor, stroke, and multiple sclerosis. A nine-slice MobileNetV3 screen escalated all 69 abnormal volumes in a fixed 157-subject cohort while exiting all 88 normal volumes, avoiding downstream inference for 56.1% of cases. In perceptual-group image development, raw sensitivity and specificity reached 98.9% and 94.0%; the gate intercepted all nine false-normal classifications and retained 70.4% normal exit yield. On a separate 60-subject routing test, validation-selected rejection retained the correct specialist for every subject and reduced scheduled specialist calls by 44.4%. Subject-disjoint BraTS20 evaluation achieved pooled whole-tumor Dice 0.895; stroke reached aggregate Dice 0.708, and sclerosis continuation improved Dice from 0.472 to 0.589. A parameter-matched experiment isolated time-constant adaptivity under fixed capacity. Source and acquisition differences constrained screening and routing generalization. Grad-CAM, confidence records, and deterministic consistency checks linked each decision to reviewable evidence. The results established a reproducible operating policy that coordinated specialist computation and traceable review across three pathologies.



***Corresponding Author:***

Kshiraja Nelapati
Department of Artificial Intelligence and Machine Learning, Ramaiah Institute of Technology
Bengaluru, India
Email: 1ms24ai031@msrit.edu

## 1. INTRODUCTION

Quality diagnosis of tumor, ischemic stroke, and multiple sclerosis often depends on access to experienced neuroradiologists and MRI systems that produce stable, high-quality scans. These conditions are not guaranteed in rural or under-resourced areas. More broadly, clinical machine-learning systems still face deployment barriers related to dataset shift, limited external validation, and uneven access to specialist review[1–3]. These constraints matter even more in neuroimaging, where small errors in lesion localization can greatly affect treatment course.

Brain tumor segmentation has been strengthened by self-configuring and transformer-based systems such as nnU-Net, UNETR, TransUNet, and SAM-style medical segmentation models [4–7]. Condition-specific benchmarks for ischemic stroke and multiple sclerosis have similarly revealed persistent performance gaps

between controlled evaluation settings and real-world acquisition conditions [8–12]. Performance can deteriorate when image quality drops because of motion, low field strength, reduced contrast, or heterogeneous acquisition protocols. These acquisition effects motivate the present work; the experiments reported here test the branches on retrospective held-out benchmark partitions, including heterogeneous multicenter 1.5 T and 3 T ISLES 2022 data, rather than on a separately generated degradation benchmark. Recent work on robust self-supervised imaging models has also shown that medical AI systems remain sensitive to domain shift and require deliberate robustness strategies before clinical deployment [13, 14]. More recent tumor studies suggest that strong segmentation can still be retained when sequence availability is reduced, which matters directly for deployment in low-resource settings [15, 16].

Several recent works contribute toward the goal of making deployment-focused neuroimaging systems suitable for rural settings. Liquid time-constant models update their hidden states through learned conductances and input- and state-dependent effective time constants. This mechanism changes how quickly recurrent features respond to new evidence; it is more precise than describing the benefit as general variability in static model weights [17–19]. Biomedical vision-language models enable semantically-based triage and routing through their ability to process semantic information [20–22] and recent research studies confirm that these models have reached sufficient development to assist medical imaging workflows [23]. The research on generalist medical AI and explainable imaging demonstrates that system predictions need additional reviewability and explicit quality control mechanisms to become complete [24–27].

nnU-Net and UNETR provide strong direct segmentation pipelines configured for a known target task [4, 5], while DeepISLES is a stroke-specific ensemble developed and clinically validated for ischemic lesions [9]. BrainMD contributes the cross-pathology operating layer around such specialists: it decides when low-cost screening can stop the pipeline, when one branch is sufficiently supported, when uncertainty requires all branches, and which evidence must accompany the result for review. The three operational algorithms make those decisions explicit for inputs whose pathology is initially unknown. The design targets lower average computation on likely normal scans, safer escalation under uncertain routing, branch-specific adaptation behind one specialist interface, and a review contract linking the route, mask, confidence, visual explanation, and deterministic consistency checks. Evaluation quantifies the routing coverage–invocation trade-off, subject-disjoint tumor transfer, and branch-specific segmentation performance, connecting the operating policy to measured outcomes.

The main contributions of this paper are:

1. An explicit resource-and-safety control policy, formalized in Algorithms 1–3, that defines early exit, single-specialist routing, uncertainty-triggered all-specialist execution, and escalation for human review.
2. A cross-pathology specialist formulation that preserves branch-specific input contracts and refinement mechanisms—LTC recurrence for tumor and stroke, and multi-scale channel–spatial attention for sclerosis—behind one segmentation and evidence-return interface.
3. An evidence-preserving review contract that binds each routed prediction to its lesion mask, confidence and uncertainty summary, Grad-CAM evidence, and auditable deterministic consistency predicates over the executed route, mask, and structured note evidence.

## 2. METHODOLOGY

BrainMD forms a directed three-stage graph that invokes expensive operations only after lower-cost stages detect abnormality [24]. Phase 1 controls confidence-gated exit, Phase 2 selects one specialist or all-branch fallback, and Phase 3 applies branch-specific segmentation while returning a common evidence record. Algorithms 1–3 specify these decisions reproducibly.

### 2.1. Phase 1: Rapid Screening (MobileNetV3)

The first stage uses a compact MobileNetV3-Small to identify cases eligible for early exit before resource-heavy routing and segmentation [28]. Each volume supplies nine uniformly spaced slices from the central 80% of a content-selected axis. Slice intensities are clipped to the 1st–99th percentiles, scaled to eight-bit values, bilinearly resized to $224 \times 224$, replicated across RGB channels, and normalized with the ImageNet channel statistics. Training and application inference use identical slice pixels; lossless images retain this input contract between preprocessing and classification.

Let $p_i$ be the abnormal probability for slice $i$. The case score is $s = \max_i p_i$, so a suspicious slice cannot be averaged away by normal-appearing slices. A scan exits only when every evaluated slice satisfies the fixed contract-specific normal-confidence floor $\tau_N > 0.5$, the scores are valid, and no clinical override is present. Otherwise the volume continues downstream with its scores and review flag. Algorithm 1 specifies this confidence-gated screening policy.

**Algorithm 1** Confidence-Aware Screening and Escalation

**Require:** MRI volume $V$, validated normal-confidence floor $\tau_N > 0.5$, clinical override $o$
Extract nine uniform central-80% slices $S$ using the fixed preprocessing contract
Compute per-slice abnormal probabilities $p_i$ with MobileNetV3-Small
Set case abnormality $s \leftarrow \max_i p_i$ and normal confidence $c_N \leftarrow 1 - s$
**if** scores are valid **and** $\neg o$ **and** $s < 0.5$ **and** $c_N \geq \tau_N$ **then**
    **return** `Normal` and stop downstream inference
**else**
    Retain per-slice scores and the most suspicious slice as review evidence
    Flag invalid scores or a deferred normal prediction for review
    Forward $V$ to Phase 2
**end if**

### 2.2. Phase 2: Intelligent Routing (Vision-Language Ensemble)

If the first model detects an abnormality, Phase 2 estimates which specialist branch should process it next. BiomedCLIP embeddings provide the default local semantic representation, while an auxiliary medical visual question-answering source can serve as a consistency check when available [20, 21]. A single branch is executed only when its score exceeds a floor $\tau_s$, its top-two margin $m$ is at least $\tau_r$, and no available confirmation source contradicts it. The default floors are $(\tau_s, \tau_r) = (0.80, 0.08)$; the fitted routing-head experiment selects both on validation subjects only. Otherwise, all three specialists are executed and the case is flagged for downstream review. This reject-option policy trades additional computation on ambiguous cases for greater route coverage. Every executed output is retained for the Grad-CAM and rule-based review stage.

#### 2.2.1. Implemented Data Flow

Each NIfTI case is loaded once with its header metadata. Routing embeds a content-selected central brain slice with BiomedCLIP, while each specialist applies its checkpoint-specific input contract. The tumor branch stacks FLAIR, T1ce, and T2 before resizing, center cropping, and spatial transformation; the stroke and sclerosis branches apply their reported intensity normalization, fixed resizing, and probability threshold. Keeping these paths separate avoids imposing an unevaluated universal preprocessing rule.

#### 2.2.2. Practical Routing Policy

Let $s_1$ denote the abnormality score produced by Phase 1 and let $r_c$ denote the routing similarity assigned to class $c \in \{\texttt{tumor}, \texttt{stroke}, \texttt{sclerosis}\}$. The routing policy used is:

1. If the complete confidence gate in Algorithm 1 passes, the scan exits as normal and the specialist branches are skipped; all other cases proceed to routing.
2. If one class receives a clearly dominant routing score, the corresponding specialist model is executed first.
3. If the routing signal is low-confidence or contradictory, the workflow falls back to all-specialist execution and treats the case as requiring stricter human review.

This policy is intentionally conservative. The design preference is to spend additional computation on ambiguous scans rather than force a hard pathology assignment from uncertain evidence. Algorithm 2 formalizes the uncertainty-aware routing decision and its all-specialist fallback.

**Algorithm 2** Semantic Uncertainty Routing

**Require:** Scan $V$ flagged abnormal or ambiguous by Phase 1; class set $\mathcal{C} = \{\texttt{tumor}, \texttt{stroke}, \texttt{sclerosis}\}$; score floor $\tau_s$ and margin threshold $\tau_r$

Extract a representative slice $x$ from the central region of $V$
Compute routing scores $r_c$ from BiomedCLIP features (prompt similarity or fitted linear head)
$c^\star \leftarrow \arg\max_c r_c$ ▷ top-ranked condition
$m \leftarrow r_{c^\star} - \max_{c \neq c^\star} r_c$ ▷ margin over second-best
Optionally query an auxiliary medical VQA source to confirm or dispute $c^\star$
**if** $r_{c^\star} \geq \tau_s$ **and** $m \geq \tau_r$ **and** no available source contradicts $c^\star$ **then**
    Route $V$ to the $c^\star$ specialist branch
**else**
    Run all three specialist branches and retain every output
    Collect masks and confidence scores; flag case for human review
**end if**

### 2.3. Phase 3: Specialist Segmentation with Adaptive Refinement

The third stage contains specialist models for each target pathology. The tumor and stroke checkpoints deserialize explicit `ncps` LTC layers, whereas the sclerosis checkpoint uses a ResNet encoder, atrous spatial pyramid pooling, and three multi-scale CBAM-style residual refinement blocks. Algorithm 3 defines the branch-specific preprocessing, refinement, decoding, and review outputs without treating unlike mechanisms as identical.

**Algorithm 3** Pathology-Specific Specialist Adaptation

**Require:** Routed condition $c$, input volume $V$, branch-specific preprocessing map $\Pi$

Apply branch-specific normalization and resizing: $\tilde{V} \leftarrow \Pi_c(V)$
Select specialist model $M_c$ for tumor, stroke, or sclerosis
Encode $\tilde{V}$ with CNN model to obtain multiscale features $\mathbf{x}_t$ via affine mapping $\tilde{x}_t = \mathbf{x}_t \odot w_{in} + b_{in}$
**if** $c \in \{\text{tumor}, \text{stroke}\}$ **then**
    Apply LTC refinement over $K$ ODE sub-steps (defined below)
**else**
    Apply multi-scale channel–spatial attention refinement (defined below)
**end if**
Decode lesion probability map $P_c$ with the specialist head
Threshold and post-process $P_c$ to obtain final overlayed mask $\hat{Y}_c$
Generate Grad-CAM explanations and rule-based review checks for $\hat{Y}_c$
**return** $\hat{Y}_c$, confidence score

**Definition 1 (LTC Refinement Step).** For each LTC cell with ODE unfolds $K$, membrane capacitance $C_m$, leak conductance $g_{leak}$, and leak potential $v_{leak}$, the sub-step size and capacitance rate are:

$$\Delta t = \frac{\text{elapsed\_time}}{K}, \qquad c_t = \frac{C_m}{\Delta t}$$

Here `elapsed_time` is the dimensionless integration horizon between successive spatial feature positions, fixed to 1.0 by the LTC cell's default call, so $\Delta t = 1/K$. In the recovered tumor checkpoint, a 64-slice crop is downsampled to four depth positions, processed in both directions with $K = 6$ solver sub-steps per position. Thus $t$ indexes spatial depth and the integration horizon is a numerical refinement parameter, rather than MRI acquisition time. The adaptive quantity is the effective time constant induced by the current feature input and hidden state at each sub-step. The sigmoid gate is $S(v; \mu, \sigma) = \text{sigmoid}(\sigma \odot (v - \mu))$. Synaptic drive terms are accumulated as:

$$N(h^{(k)}, \tilde{x}_t) = \sum_j w_{ij}\, S(h_j^{(k)})\, E_{ij} + \sum_m w_{mi}^{sens}\, S(\tilde{x}_{t,m})\, E_{mi}^{sens}$$

$$D(h^{(k)}, \tilde{x}_t) = \sum_j w_{ij}\, S(h_j^{(k)}) + \sum_m w_{mi}^{sens}\, S(\tilde{x}_{t,m})$$

The semi-implicit Euler hidden-state update for $k = 0, \ldots, K-1$ is:

$$h^{(k+1)} = \frac{c_t \odot h^{(k)} + g_{leak} \odot v_{leak} + N(h^{(k)}, \tilde{x}_t)}{c_t + g_{leak} + D(h^{(k)}, \tilde{x}_t) + \epsilon} \tag{1}$$

The effective per-neuron time constant emerging from the denominator is:

$$\tau_i^{(k)} \approx \frac{C_{m,i}}{g_{leak,i} + D_i^{(k)} + \epsilon} \tag{2}$$

making $\tau_i^{(k)}$ input- and state-dependent. The resulting update changes the rate at which each hidden unit responds to new features; it does not, by itself, establish immunity to any particular acquisition artefact.

**Definition 2 (CBAM-Style Sclerosis Refinement).** For the sclerosis branch, the checkpoint applies grouped multi-path fusion and dual attention directly to ASPP features:

$$\begin{aligned} x_g &= x_{[:,1:g,:,:]}, \quad f = P_1(x_g) + P_2(x_g) \\ c_a &= \mathrm{sigmoid}(W_2(\mathrm{ReLU}(W_1(\mathrm{GAP}(f))))), \quad f_c = f \odot c_a \\ s_a &= \mathrm{sigmoid}(\mathrm{Conv}_{7\times7}([\mathrm{mean}_c(f_c),\, \max_c(f_c)])) \\ y &= \mathrm{BN}(x + f_c \odot s_a) \end{aligned}$$

### 2.4. Experimental Materials and Evaluation Method

The experimental materials were NIfTI MRI volumes, their header metadata, branch-required modalities, expert-provided lesion masks, saved specialist checkpoints, and the per-case probability maps produced during validation and held-out inference. Specialist experiments used fixed branch-specific subject partitions and a single NVIDIA T4 GPU without model sharding; the routing benchmark used a CPU, and the revised Phase 1 training used a Modal NVIDIA B300. The branches share one application, explanation path, and clinician-facing interface, but retain different input requirements and preprocessing because their checkpoints do not accept one universal tensor representation.

The evaluation procedure enumerates public-release cases, verifies required image-mask pairs and development-set membership, freezes subject partitions, applies checkpoint-specific preprocessing, and reserves validation outputs for any fitting or threshold selection. Measurements compare predictions with processed reference masks; visually difficult or low-scoring cases are retained. The stroke experiment used the 250 expert-annotated subjects in the public ISLES 2022 training release (Zenodo record 7960856), supplying DWI, ADC, FLAIR, and lesion masks [8]. Its internal 175 / 25 / 50 partition is distinct from the challenge's hidden 150-case test cohort; the qualitative example is `sub-strokecase0219_ses-0001`. The tumor transfer evaluation excludes the entire BraTS2018 HGG development cohort using the release name mapping and evaluates all 159 remaining BraTS20 subjects without further fitting. Table 1 records the inclusion, preprocessing, and test-decision rules.

Table 1. Fixed-split evaluation protocol and data-selection rules

| Branch | Included material | Selection and preprocessing | Fixed split / test decision |
|---|---|---|---|
| Screening | 581 IXI T1 normal, 285 BraTS18 FLAIR tumor, 100 ISLES22 DWI stroke, and 60 MS FLAIR volumes | Source-stratified subject split (seed 865); nine central-80% slices; maximum abnormal probability | 563 / 153 / 153 / 157 train / validation / development / fixed evaluation; validation-selected volume gate |
| Tumor | BraTS20 FLAIR, T1ce, T2, and reference masks; recovered BraTS2018 HGG checkpoint | Exclude all 210 earlier HGG subjects; centered $128\times128\times64$ crop and foreground z-score normalization | All 159 remaining subjects (83 HGG, 76 LGG); no fitting or threshold selection on this cohort |
| Stroke | ISLES 2022 public training release: 250 expert-annotated subjects with DWI, ADC, FLAIR, and lesion masks | Intensity standardization, fixed-resolution resizing, and multimodal adapter for DWI/ADC/FLAIR fusion | Internal 175 / 25 / 50 split; threshold 0.0001 selected on validation before test evaluation |
| Sclerosis | Muslim *et al.* Mendeley Data v1: 60 cases with T1, T2, FLAIR, and corresponding consensus lesion masks | FLAIR and corresponding mask used; min–max normalization, fixed-resolution resizing, and validation-only threshold calibration | 60 cases: 42 / 6 / 12; checkpoint selected only from the frozen validation split |

Phase 1 uses the IXI healthy MRI release [29] and the pathology releases listed above. Auxiliary image supervision came from `archive_3/Training`, with `notumor` mapped to normal. We excluded

134 files matching the official test partition by SHA-256 and removed 136 redundant copies. For the 5,442 unique images, 63-bit DCT and 256-bit difference hashes connected pairs within Hamming distances 8 and 32. Whole-component stratification of the resulting 3,917 components (largest 24) produced 3,256 training, 1,101 validation, and 1,085 development images (60/20/20%, seed 865), with zero qualifying cross-split edges or mixed-label components. Patient identifiers are unavailable, so this evidence is image-level; the volume partitions are subject-disjoint.

The ImageNet-initialized screen was pretrained for five image epochs and jointly trained for 15 further epochs, pairing each eight-volume batch with 64 images. The loss equally weights two class-weighted binary cross-entropies. AdamW uses learning rate $3 \times 10^{-4}$, weight decay $10^{-4}$, and cosine decay. Augmentations are horizontal flips, rotations within 7°, translations within 3%, scales 0.95–1.05, and brightness/contrast jitter of 0.1. Checkpoint selection minimizes abnormal validation exits, then maximizes minimum domain yield, mean yield, and mean AUROC; ties retain the earlier epoch. This selects epoch 8. Training uses PyTorch 2.13.0a0, CUDA 13.3, and TF32; final calibration, inference, and Grad-CAM use IEEE FP32 with TF32 and cuDNN disabled.

The checkpoint is calibrated separately for nine-slice volumes and single images. If $r$ is the smallest abnormal validation probability, the candidate normal-confidence floor is $\max\{0.505, 1 - r/(2 - r)\}$; the volume floor is 0.505. Development calibration sets the image exit requirement to abnormal probability at most $10^{-7}$, or normal confidence at least 0.9999999. The image result is therefore threshold-development evidence, while the fixed subject-level volume cohort is the primary evaluation. Prediction files bind probabilities and decisions to the threshold and checkpoint hash; clinician context overrides automatic exit.

For binary lesion masks, the reported measurements use

$$\text{Dice} = \frac{2TP}{2TP + FP + FN}, \qquad \text{IoU} = \frac{TP}{TP + FP + FN},$$
$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}. \tag{3}$$

Case-level means average the per-case measurements; aggregate scores pool voxel counts across the entire held-out split before applying the same equations. This distinction is retained because pooled overlap can be dominated by larger lesions.

### 2.5. Clinical Support Ecosystem

The full architecture of the system is illustrated in Figure 1.

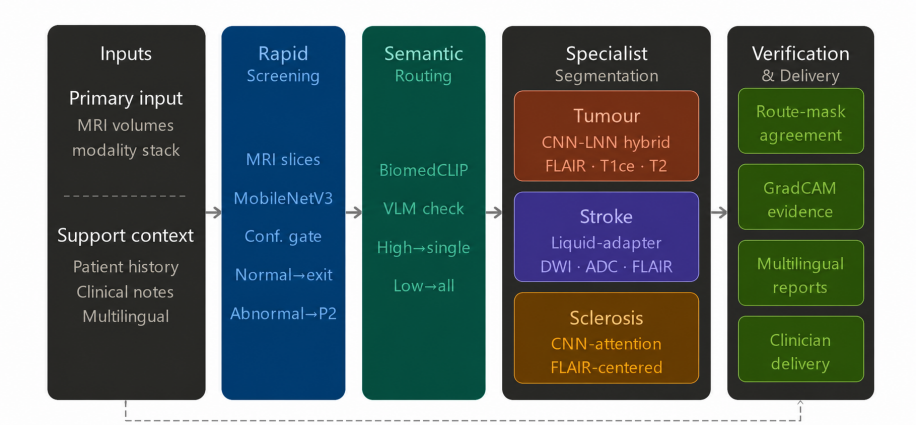



Figure 1. Architecture : Phase 1 performs rapid screening, Phase 2 routes suspicious scans with biomedical VLMs, and Phase 3 invokes the tumor, stroke, or sclerosis specialist before explanation, rule-based review, and interface delivery

### 2.6. Rule-Based Consistency Checking

The framework includes a deterministic post-analysis consistency-checking layer. It ingests the routed condition label, segmentation summary, generated explanation text, and any available structured clinical-note

entities, then directly evaluates explicit Boolean IF–THEN predicates for agreement, contradictions, and missing evidence. Each check attaches a transparent review flag and supporting evidence to the model output before clinician review. A generic check can be summarized as:

$$\text{review\_status}(\hat{y}) \Leftarrow \text{route\_mask\_agreement}(\hat{y}) \wedge \text{note\_consistency}(\hat{y}) \wedge \neg\text{high\_risk\_contradiction}(\hat{y}) \quad (4)$$

For a tumor-oriented output, one illustrative Boolean check is:

$$\text{tumor\_review\_pass} \Leftarrow \text{tumor\_route} \wedge \text{nonzero\_lesion\_mask} \wedge \neg\text{stroke\_specific\_contradiction} \quad (5)$$

The layer records whether these predefined conditions pass or fail. It is not a symbolic reasoning system and does not establish a diagnosis independently of the imaging models and clinician review.

### 2.7. Explainability with Grad-CAM

Grad-CAM highlights image regions contributing to an output. The heatmap supports inspection of whether attention falls on plausible lesion regions, while confidence and consistency records retain the complementary numerical and rule-based evidence.

## 3. RESULTS AND DISCUSSION

Screening and routing were evaluated as explicit control decisions before assessing the three specialist branches on BraTS20 tumor data, the public ISLES 2022 training release for ischemic stroke, and version 1 of the Muslim *et al.* public multiple-sclerosis dataset with consensus manual lesion annotations [10]. The results distinguish abnormal escalation, normal exit yield, specialist coverage, and segmentation quality.

### 3.1. Model Size and Complexity

BrainMD uses a two-tier deployment path. The 6.2 MB screen is the edge-facing early-exit stage; cases that continue are processed on a local workstation or server that holds BiomedCLIP and the selected specialists. The default Phase 2 path is local after the weights are provisioned, so inference does not require a live network connection. LLaVA-Med is an optional confirmation source and is disabled in the measured default path. Table 2 makes the downstream footprint explicit.

Table 2. Serialized model footprint and deployment role

| Phase / Task | Model Type | Size | Role |
|---|---|---|---|
| Phase 1 (Screening) | MobileNetV3-Small | 6.2 MB | CPU-screened entry stage |
| Phase 2 (Routing) | BiomedCLIP ViT-B/16 | 784 MB | Local workstation/server |
| Phase 3 (Tumor) | CNN-LNN Hybrid | 317 MB | On-demand specialist |
| Phase 3 (Stroke) | 3D Volumetric Net | 537 MB | On-demand specialist |
| Phase 3 (Sclerosis) | ResNet-attention | 431 MB | On-demand specialist |

The Phase 1 screen has 1,519,906 parameters and occupies 6.2 MB. The 784 MB Phase 2 value is the public BiomedCLIP checkpoint; the 317 MB tumor checkpoint includes optimizer state and a 26,372,008-parameter network. These are serialized sizes rather than runtime allocations.

### 3.2. Speed and Efficiency Benchmarks

The production screen and Grad-CAM path were exercised on four real validation cases, one per source, using a four-thread Intel Core i5-13500H CPU and PyTorch 2.10.0. Processing the nine cached slices took 0.427 s for the first case and 0.147–0.152 s for the subsequent three, excluding volume preprocessing and checkpoint loading. All four decisions matched B300 inference, with maximum probability discrepancy $4.8 \times 10^{-9}$. On the B300 final cohort, serial case inference took 8.70 s for 157 volumes; preprocessing took a separate 141.15 s and Grad-CAM was excluded. These are component measurements with explicitly separated preparation costs.

A checkpoint-restoration pilot on one real BraTS2018 volume measured warm tumor inference at 0.354 s (three runs: 0.354, 0.354, and 0.355 s) on a T4 with TensorFlow 2.15.1, using a $128 \times 128 \times 64 \times 3$ input. Peak TensorFlow GPU allocation was 5.26 GB; preprocessing and model loading were excluded. The routing experiment encoded 180 representative slices in 29.66 s on a four-thread CPU (0.165 s per slice), excluding the 30.99 s model load. These component measurements locate the screen on the edge-facing path and the representation/specialist stack on a local workstation or server. Actual downstream latency depends on modality preparation and whether one or three specialists execute.

### 3.3. Confidence-Gated Screening Evaluation

Algorithm 1 was evaluated with the frozen checkpoint and input-specific thresholds. Table 3 separates raw false-normal classifications from abnormal cases incorrectly authorized to exit. On the fixed 157-subject volume cohort, the screen classified all 88 normal and 69 abnormal volumes correctly and escalated every abnormal case: 45 tumor, 15 stroke, and nine MS. Abnormal escalation sensitivity was 100% (Wilson 95% CI: 94.7–100%), and normal exit yield was 100% (95.8–100%). Thus 88/157 cases (56.1%) skipped both routing and specialist inference, with zero observed abnormal early exits. The preceding 153-volume development cohort likewise yielded 87 normal exits and zero abnormal exits among 66 abnormal subjects.

Table 3. Phase 1 decisions with frozen confidence gates; N/A denote normal/abnormal reference counts

| Cohort | N / A | Raw false normals | Abnormal exits | Normal exits |
|---|---|---|---|---|
| Volume development | 87 / 66 | 0 | 0 / 66 | 87 / 87 |
| Fixed volume evaluation | 88 / 69 | 0 | 0 / 69 | 88 / 88 |
| Grouped image development | 267 / 818 | 9 | 0 / 818 | 188 / 267 |

The earlier image-random pilot yielded 89.2% sensitivity and 82.5% specificity. After exact deduplication, label repair, and perceptual-component splitting, the grouped development classifier reached 97.7% accuracy, 98.9% sensitivity, 94.0% specificity, and AUROC 0.9948. Its nine raw false normals were all escalated by the revised gate. The gate retained 188/267 normal exits (70.4%; Wilson 95% CI: 64.7–75.6%) and deferred the remaining 79 normals for downstream review. Figure 2(a) shows the fixed volume evaluation; Figure 2(b) shows the grouped image decisions. Exact production replay reproduced all 2,492 validation/development scores and gate decisions across the two input contracts, while real-volume preprocessing checks matched the training pixels for one case per source.

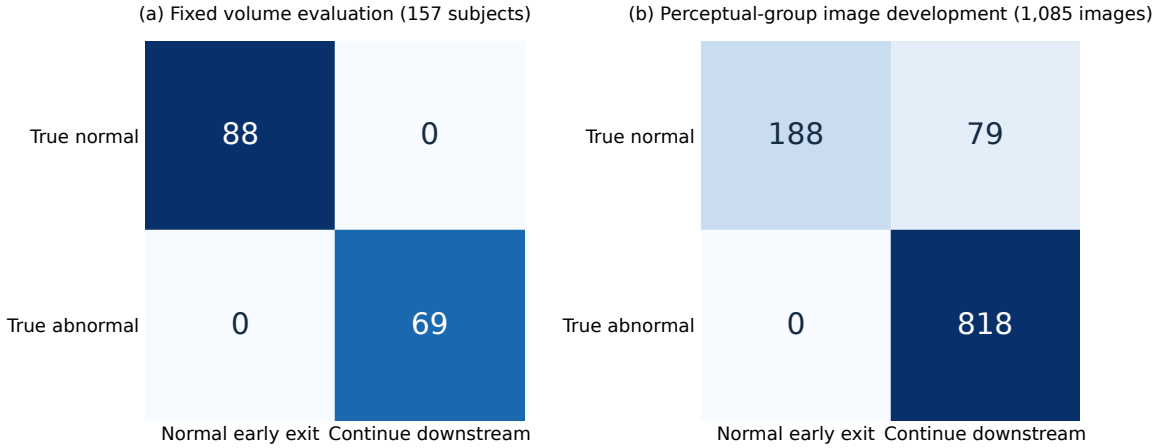



Figure 2. Confidence-gated screening decisions. Panel (a) shows the fixed 157-subject volume evaluation. Panel (b) shows the perceptual-group image development cohort. Escalation includes both abnormal predictions and confidence-deferred normal predictions.

The volume partitions are subject-disjoint within source; normal IXI T1 and abnormal FLAIR/DWI sources differ in acquisition and contrast. The perfect volume separation therefore warrants multi-source, contrast-matched validation before extrapolation to clinical populations. The image cohort is exact- and perceptual-component-disjoint; patient identifiers are unavailable, so its image-level counts are reported separately from the fixed subject-level evaluation. Screening and routing use separate stage-specific cohorts, so their invocation reductions are not multiplied into an end-to-end saving.

### 3.4. Held-Out Routing and Reject-Option Evaluation

To evaluate the control policy separately from segmentation, we froze 60 subjects per pathology from BraTS2018, ISLES2022, and the Muslim et al. MS release: 30 training, 10 validation, and 20 test subjects per class (seed 865). Every input used FLAIR, canonical orientation, a foreground crop, 1st–99th percentile normalization, and the highest-entropy axial slice within the central 20–80% of occupied slices, resized to $224\times 224$. Selection used neither lesion masks nor reports. A frozen BiomedCLIP image encoder fed a multinomial logistic head ($C = 1$); only its linear coefficients were fitted on the 90 training subjects. Validation selected the

score/margin pair maximizing single routes with zero observed single-route errors: $(0.55734, 0.21690)$. The 60 test subjects were then evaluated once; Table 4 reports the resulting decisions.

Table 4. Held-out routing decisions on 60 FLAIR subjects; calls are scheduled specialist invocations

| Configuration | Top-1 correct | Fallbacks | Covered | Calls |
|---|---|---|---|---|
| Zero-shot prompts, default floors | 16/60 | 60/60 | 60/60 | 180 |
| Fitted head, default floors | 57/60 | 55/60 | 60/60 | 170 |
| Fitted head, validation-selected floors | 57/60 | 20/60 | 60/60 | 100 |

The fitted head's test confusion matrix (true rows and predicted columns: tumor, stroke, MS) was $[(18, 2, 0); (1, 19, 0); (0, 0, 20)]$, giving 95.0% top-1 accuracy (Wilson 95% CI: 86.3–98.3%). All three errors triggered the reject option. The 40 single routes contained no observed errors (95% CI for single-route accuracy: 91.2–100%); route coverage was 60/60 (93.98–100%). This reduced scheduled specialist calls by 44.4% relative to all-three execution, with a 33.3% fallback rate. Coverage measures whether the reference branch is included, separately from segmentation correctness or measured compute savings.

Source confounding remains material: a logistic baseline using only volume geometry, spacing, crop geometry, entropy, and a 16-bin intensity histogram classified 59/60 test subjects correctly. The releases also differ in skull stripping. Thus the experiment establishes the reject-option mechanism under a fixed benchmark protocol; acquisition-independent routing requires multi-source cohorts within each pathology. The zero-shot result includes the existing `Other` prompt, whereas the fitted head is closed-set over the three specialist labels. Normal/open-set discrimination remains the separate screening task. Software regression tests additionally verify that uncertainty and source disagreement execute all branches, preserve returned evidence, and expose unavailable specialist outputs for review.

### 3.5. Subject-Disjoint Tumor Transfer Evaluation on BraTS20

The recovered tumor checkpoint was developed using BraTS2018 HGG subjects. Because later BraTS releases contain earlier subjects under new identifiers, we excluded all 210 BraTS2018 HGG subjects through the release name mapping before evaluating BraTS20. The resulting frozen test cohort comprised 159 subjects (83 HGG, 76 LGG). All required modalities and masks were present; case 355's uniquely named segmentation file was resolved within its own directory and retained. The original checkpoint was evaluated without fine-tuning. Metrics refer to its centered $128 \times 128 \times 64$ input crop and processed reference mask, with labels 4 remapped to 3 and an argmax segmentation decision.

The cohort achieved mean Dice 0.838/0.587/0.637 for whole tumor (WT), tumor core (TC), and enhancing tumor (ET), respectively; pooled Dice was 0.895/0.629/0.839 and voxel accuracy was 96.77%. Table 5 includes the prespecified grade strata: whole-tumor localization transfers strongly to HGG, while LGG core delineation is a substantial remaining challenge. Reference WT/TC/ET was present in 159/157/129 crops; two empty masks receive Dice 1, and a nonempty prediction against an empty reference receives Dice 0.

Table 5. Tumor transfer results on subject-disjoint BraTS20 crops

| Metric | All subjects | HGG | LGG |
|---|---|---|---|
| Test cases | 159 | 83 | 76 |
| Mean WT Dice | 0.838 | 0.919 | 0.748 |
| Mean TC Dice | 0.587 | 0.877 | 0.271 |
| Mean ET Dice | 0.637 | 0.865 | 0.389 |
| Pooled WT Dice | 0.895 | 0.941 | 0.847 |
| Pooled TC Dice | 0.629 | 0.909 | 0.306 |
| Pooled ET Dice | 0.839 | 0.891 | 0.626 |

### 3.6. Stroke Evaluation on the ISLES 2022 Public Training Release

The stroke model was evaluated on the 250 expert-annotated subjects in the public ISLES 2022 training release, not on the challenge's hidden test cohort [8]. This multicenter release contains DWI, ADC, FLAIR, and lesion masks acquired with heterogeneous clinical 1.5 T and 3 T scanners. We created an internal fixed split of 175 training cases, 25 validation cases, and 50 test cases. The first continuation kept the liquid neural layer trainable in a DWI-only run and improved the pooled overlap statistics. The second introduced a lightweight modality adapter in front of the stroke CNN-LNN checkpoint so that DWI, ADC, and FLAIR could be fused without discarding the pretrained liquid-hybrid backbone.

The multimodal liquid-adapter run stands as the final stroke checkpoint in this research because it achieved the highest case-level overlap results during validation-controlled model selection. On the validation

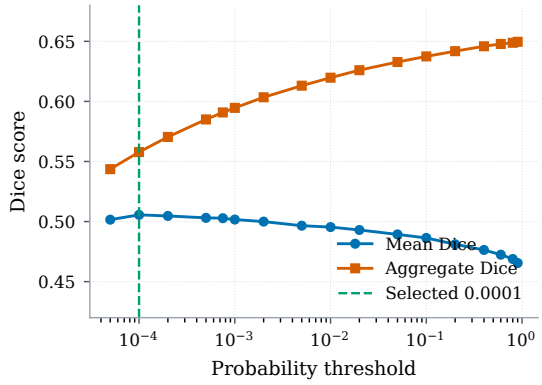
split, the best operating point was a threshold of 0.0001 on the stroke probability volume. That threshold was then fixed before test evaluation. On the 50 held-out test cases, the stroke model reached a mean Dice score of 0.543 and an aggregate Dice score of 0.708, with mean IoU of 0.421, mean recall of 0.563, and mean precision of 0.605. Comparisons with specialized systems such as DeepISLES [9] require explicit accounting for differences in evaluation protocols. The validation sweep shows that case-mean Dice peaked at the selected threshold of 0.0001 even though aggregate Dice kept increasing at more conservative thresholds. This distinction matters in stroke segmentation because a threshold chosen only for pooled overlap can bias the system toward larger lesions while sacrificing balanced case-level behavior.

Table 6 summarizes the held-out stroke measurements and the validation-selected threshold used for the test evaluation.

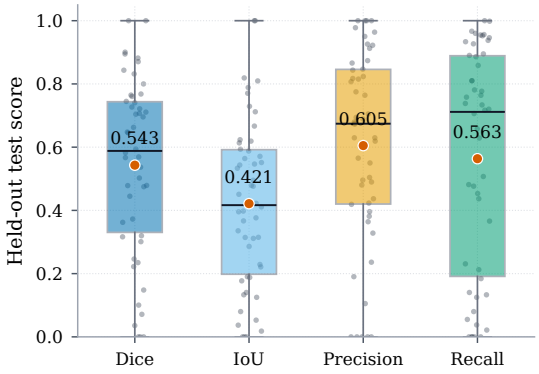
Table 6. Stroke segmentation results on the internal ISLES 2022 test split after validation-set threshold calibration

| Metric | Score |
|---|---|
| Public ISLES 2022 subjects | 250 |
| Train / Val / Test split | 175 / 25 / 50 |
| Validation-selected threshold | 0.0001 |
| Mean Dice | 0.543 |
| Aggregate Dice | 0.708 |
| Mean IoU | 0.421 |
| Aggregate IoU | 0.547 |
| Mean precision | 0.605 |
| Mean recall | 0.563 |
| Aggregate precision | 0.627 |
| Aggregate recall | 0.812 |
| Mean specificity | 0.999 |
| Mean accuracy | 99.81% |

Figure 3(a) shows where the validation Dice peaks across candidate thresholds, while Figure 3(b) shows the spread of Dice, IoU, precision, and recall across the 50 held-out test cases.




(a) Validation threshold sweep for the multimodal liquid-adapter stroke model




(b) Held-out test metric distribution across the 50 stroke test cases

Figure 3. Stroke validation and held-out test reporting. Panel (a) shows the validation-set threshold sweep used to select one operating point before test evaluation. Panel (b) shows the distribution of Dice, IoU, precision, and recall across the 50 held-out test cases.

### 3.7. Multiple Sclerosis Evaluation on the Muslim et al. Public Dataset

The multiple sclerosis model was evaluated on version 1 of the *Brain MRI Dataset of Multiple Sclerosis with Consensus Manual Lesion Segmentation and Patient Meta Information*, released by Muslim *et al.* through Mendeley Data (dataset ID 8bctsm8jz7) [10]. The release contains 60 confirmed multiple-sclerosis cases collected through the MS Clinic at Baghdad Teaching Hospital, with 1.5 T acquisitions drawn from 20 centres. Each case provides T1, T2, and FLAIR images and a corresponding consensus manual lesion mask. This experiment used FLAIR and its consensus mask to build the fixed split of 42 training cases, 6 validation

cases, and 12 test cases. The sclerosis checkpoint was first evaluated under that split with min–max intensity normalization and threshold calibration restricted to validation cases.

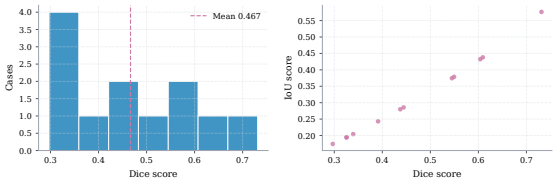
The sclerosis model was fine-tuned from a saved checkpoint while keeping its multi-scale attention refinement blocks and decoder path trainable. Validation sweeps were repeated after each epoch, and the best checkpoint was chosen only from the frozen validation split. On the held-out test set, that model reached a mean Dice score of 0.467 and an aggregate Dice score of 0.589, with mean IoU of 0.315, mean precision of 0.547, and mean recall of 0.462. Relative to the saved baseline on the same split, the gains were about 10.1 percentage points in mean Dice and 11.7 percentage points in aggregate Dice. The strongest held-out qualitative case, Patient-31, achieved a Dice score of 0.731. Recent MS lesion work continues to emphasize score calibration in small-lesion conditions, and newer benchmark releases such as MSLesSeg are helping make cross-study comparison less ad hoc.

Table 7 reports the baseline and fine-tuned sclerosis measurements on the fixed public split.

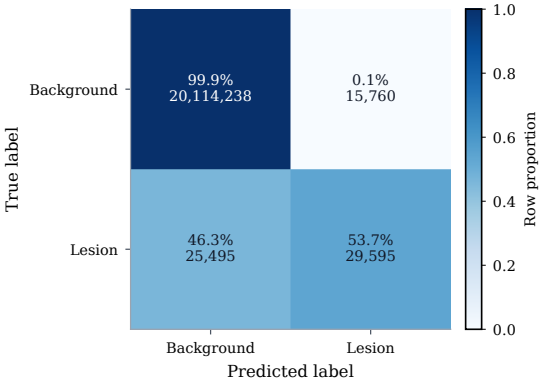
Table 7. Multiple sclerosis segmentation results on the fixed test split of the Muslim et al. public dataset

| Metric | Score |
|---|---|
| Public MS cases | 60 |
| Train / Val / Test split | 42 / 6 / 12 |
| Baseline mean Dice | 0.365 |
| Baseline aggregate Dice | 0.472 |
| Fine-tuned mean Dice | 0.467 |
| Fine-tuned aggregate Dice | 0.589 |
| Mean IoU | 0.315 |
| Aggregate IoU | 0.418 |
| Mean precision | 0.547 |
| Mean recall | 0.462 |
| Aggregate precision | 0.653 |
| Aggregate recall | 0.537 |
| Mean specificity | 0.999 |
| Mean accuracy | 99.79% |

Figure 4(a) shows the case-level spread of Dice and IoU values, while Figure 4(b) shows the aggregate confusion matrix across the 12 held-out test cases.




(a) Held-out case-level sclerosis metric distribution




(b) Aggregate confusion matrix for the sclerosis test split

Figure 4. Multiple-sclerosis held-out test reporting. Panel (a) shows case-level Dice and IoU variation across 12 patients. Panel (b) reports aggregate voxel-level confusion statistics.

### 3.8. Same-Split Component Analysis

Table 8 compares the saved baseline checkpoint with the continuation in which the attention refinement blocks and decoder path remained jointly trainable. Both configurations use the same 42 / 6 / 12 split and validation-only checkpoint selection, avoiding a cross-dataset or test-selection confound.

Table 8. Same-split checkpoint continuation for the sclerosis branch

| Checkpoint configuration | Changed component(s) | Mean Dice | Aggregate Dice |
|---|---|---|---|
| Saved baseline checkpoint | None; saved model evaluated directly | 0.365 | 0.472 |
| Recorded continuation | Attention refinement and decoder jointly trainable | 0.467 | 0.589 |
| Absolute change | Continuation minus baseline | +0.102 | +0.117 |

The same-split comparison attributes the $+0.102$ mean-Dice and $+0.117$ aggregate-Dice change to the joint attention–decoder continuation. Mechanism-specific effects require a separate matched comparison.

#### 3.8.1. Parameter-Matched Time-Constant Control

We tested the tumor LTC mechanism on 60 BraTS2018 LGG subjects outside the recovered checkpoint's HGG training directory. A frozen seed-865 manifest assigned 30/10/20 subjects to training/validation/test. Both arms restored identical checkpoint weights and retained the same encoder, feature projections, attention gates, and decoder, all frozen. Only the two LTC cells were trainable (165,376 parameters per arm). The adaptive arm used Definition 1; the control retained its dynamic numerator but evaluated denominator conductance gates at zero mapped input and zero hidden state. This retains learned time constants and the parameter count while removing their input/state dependence.

Both arms used the original $128 \times 128 \times 64$ FLAIR/T1ce/T2 crop and foreground z-score normalization, Adam at $10^{-4}$, batch size one, and three epochs without augmentation. The shared loss equally weighted mean categorical cross-entropy and foreground soft-Dice loss. Seeds 865/866/867 determined identical paired patient orders; validation mean whole-tumor Dice selected among epochs 0–3 before test inference. Table 9 reports the mean across those three orders; no cropped test reference mask was empty.

Table 9. Matched frozen-backbone LGG transfer comparison on 20 test subjects

| Refinement | Trainable parameters | Mean WT Dice | Mean TC Dice | Mean ET Dice |
|---|---|---|---|---|
| Adaptive LTC | 165,376 | 0.6937 | 0.3180 | 0.4393 |
| Fixed-time-constant control | 165,376 | 0.6944 | 0.3192 | 0.4406 |

The paired WT-Dice difference (adaptive minus fixed) was $-0.00066$, with a patient-bootstrap 95% CI of $[-0.00179, 0.00018]$ (10,000 resamples, seed 865). This short transfer experiment resolves no adaptive-time-constant advantage and removes the jointly trained decoder as an explanation for the comparison. Both arms inherit an adaptive-pretrained initialization; the result concerns this frozen-backbone continuation setting, with from-scratch training and stroke-specific isolation remaining separate experiments.

### 3.9. Specialist Confusion and Training Metrics

Figure 5 reports the aggregate BraTS20 tumor confusion matrix across the 159-subject transfer cohort, separating background, necrotic core, edema, and enhancing tumor.

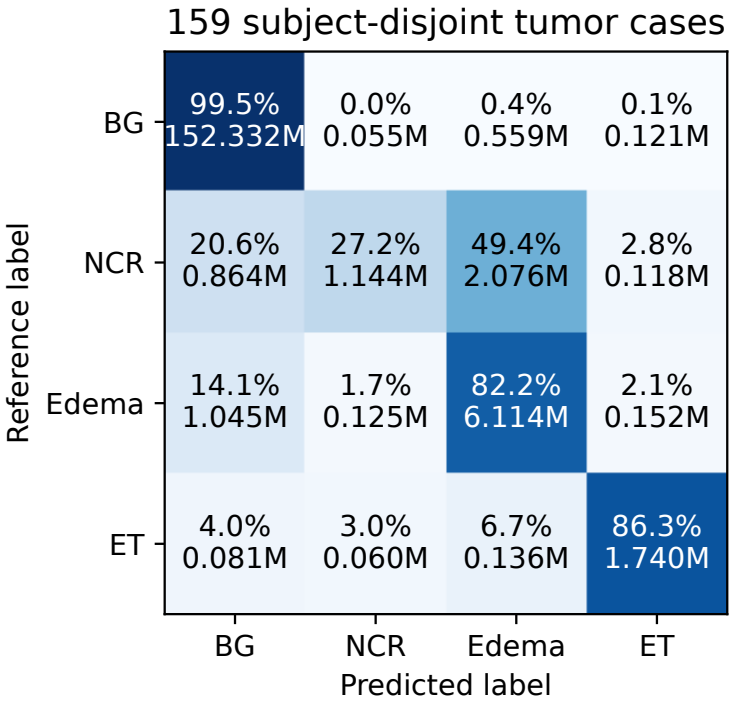



Figure 5. Aggregate tumor confusion matrix for the 159-subject BraTS20 cohort: BG, background; NCR, necrotic core; ET, enhancing tumor. Cells show row percentages and counts in millions.

Figure 6 reports the aggregate stroke confusion matrix; the lesion cell should be read together with the class imbalance represented by the much larger background cell.

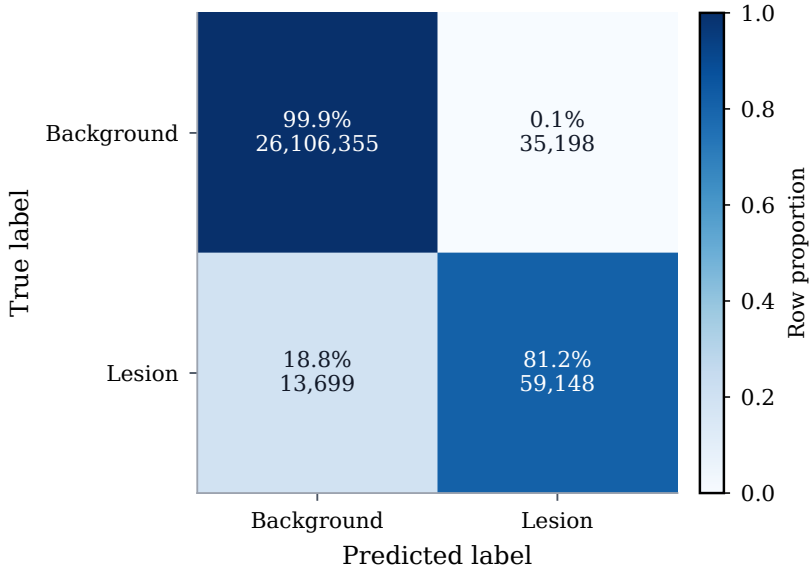



Figure 6. Aggregate stroke test-split confusion matrix separating background and lesion voxels.

Figure 7 presents the parameter-matched tumor refinement comparison: the left panel shows mean training loss, and the right panel shows validation WT Dice across the three paired patient orders.

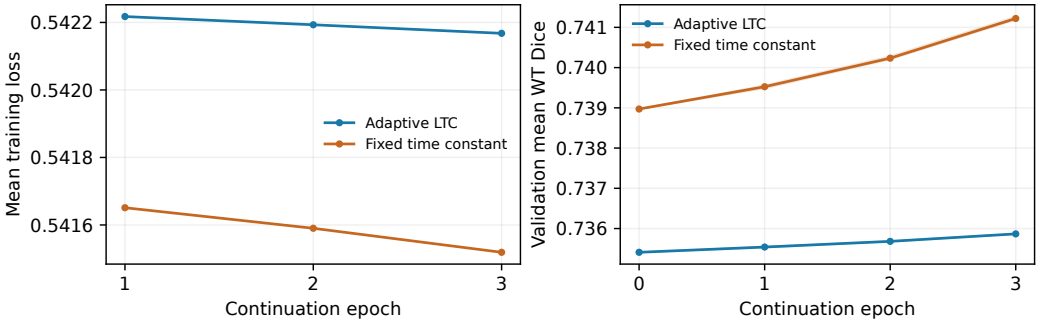



Figure 7. Matched tumor refinement continuation: mean training loss and validation WT Dice across three paired patient orders; validation shading denotes the across-order range.

Figure 8 presents stroke optimization diagnostics: the left panel shows training and validation loss, and the right panel shows training, validation, and threshold-swept Dice trajectories.

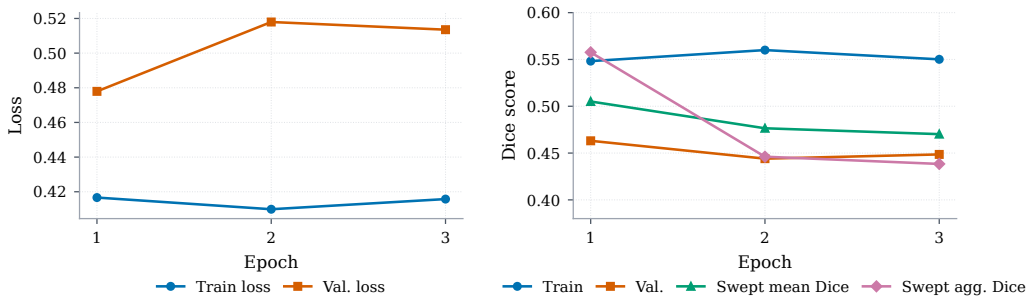



Figure 8. Training and validation curves for the stroke CNN-LNN hybrid branch, including loss and Dice measurements.

Figure 9 presents sclerosis optimization diagnostics: the left panel shows training and validation loss, the middle panel shows validation Dice and IoU, and the right panel shows the learning-rate schedule.

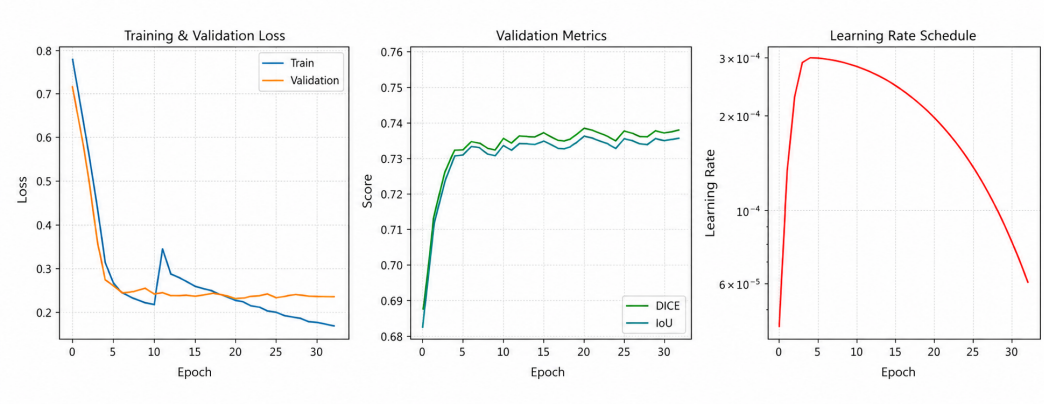
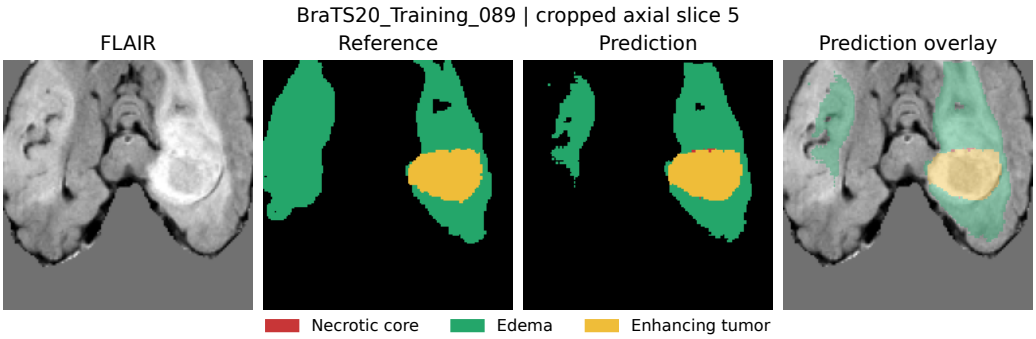



Figure 9. Optimization diagnostics for the multiple-sclerosis ResNet-attention branch, showing training and validation loss, validation Dice and IoU, and the learning-rate schedule.

### 3.10. Qualitative Segmentation Results

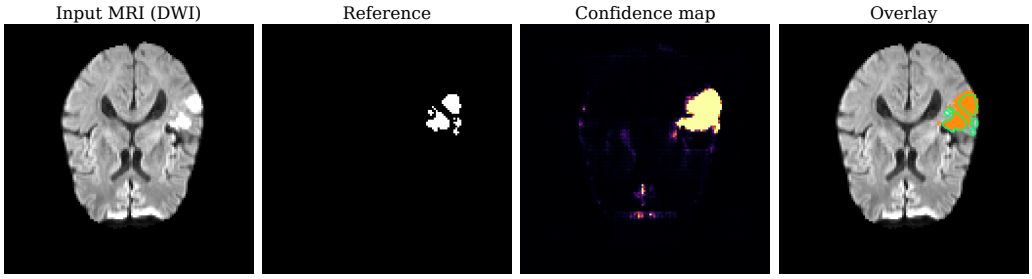
Figure 10(a) shows the held-out BraTS20 tumor case as a qualitative complement to the aggregate measurements.




(a) Subject-disjoint BraTS20 tumor case (BraTS20_Training_089)

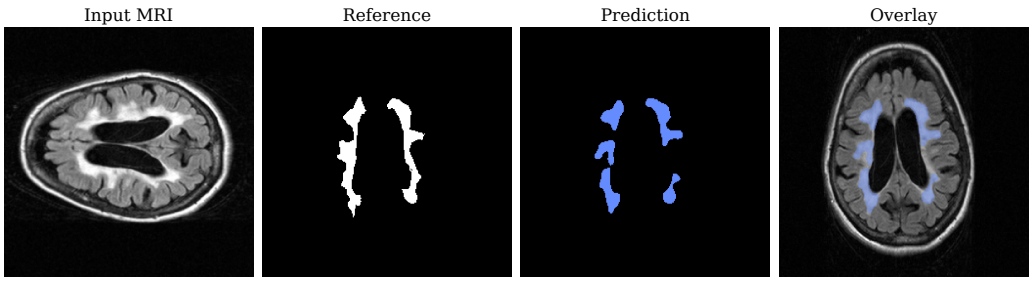
Figure 10. Held-out qualitative segmentation cases. Panel (a) shows the first case in the frozen tumor manifest, using the cropped slice with greatest reference lesion area for visualization only: input, reference, prediction, and overlay.

Figures 10(b) and 10(c) show the held-out ISLES 2022 stroke case sub-strokecase0219_ses-0001 and the held-out sclerosis case Patient-31, respectively. Together, the three examples complement, rather than replace, the aggregate measurements.




(b) Held-out ISLES stroke test case (sub-strokecase0219_ses-0001)

Figure 10. Held-out qualitative segmentation cases (continued). Panel (b) shows the stroke input, reference annotation, prediction, and overlay for the same qualitative inspection.




(c) Held-out multiple sclerosis test case (Patient-31)

Figure 10. Held-out qualitative segmentation cases (continued). Panel (c) shows the multiple-sclerosis input, reference annotation, prediction, and overlay for the same qualitative inspection.

### 3.11. Comparison with Recent Methods

Table 10 places BrainMD alongside eight recent studies, with each result interpreted under its own dataset, modalities, target, and evaluation protocol. Huang *et al.* evaluate reduced-sequence tumor segmentation [15]; BrainMD adds subject-disjoint transfer measurements with HGG/LGG stratification. Recurrent, transformer-dense, attention-based, and challenge-derived stroke systems provide task-specific context [9, 30–32]. The sclerosis comparison distinguishes BrainMD's fixed-split continuation from the heterogeneous, multi-site MSLesSeg setting [12, 33]. BrainMD's system-level evaluation additionally measures specialist-selection coverage and scheduled invocations alongside the segmentation results.

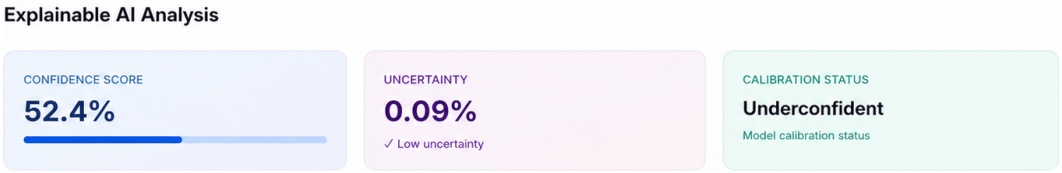
Table 10. Comparison with recent related neuroimaging segmentation work

| Task | Study | Method / setting | Reported result | Relation to BrainMD |
|---|---|---|---|---|
| Tumor | Huang *et al.*, 2025 [15] | 3D U-Net with reduced MRI sequence sets on BraTS 2018/2021 | T1C+FLAIR test Dice: ET 0.867, TC 0.926 | BrainMD's subject-disjoint cropped transfer cohort gives pooled ET 0.839 and TC 0.629; protocols differ. |
| Tumor | Zhou, 2024 [34] | Multi-modal graph convolution for BraTS segmentation | Strong BraTS 2018/2019 results with explicit modality-relation modeling | Modality-relation modeling complements BrainMD's FLAIR/T1ce/T2 transfer evaluation. |
| Stroke | Dogru *et al.*, 2024 [30] | Modified recurrent U-Net on multi-spectral MRI | Average DSC 0.748 | BrainMD reports case-mean Dice 0.543 and pooled Dice 0.708; aggregation and splits differ. |
| Stroke | Jazzar *et al.*, 2024 [31] | Transformer Dil-DenseUNet on SISS 2015 and ISLES 2022 | Test Dice: $0.80 \pm 0.30$ and $0.81 \pm 0.33$, respectively | Task-specific transformer results; BrainMD reports 0.543 mean / 0.708 pooled Dice under its internal split. |
| Stroke | Rahman *et al.*, 2025 [32] | DenseNet121/SelfONN attention U-Net on ISLES 2022 | DSC 0.8749 using DWI, ADC, and eDWI | Modality-specific comparison: BrainMD uses DWI, ADC, and FLAIR with a validation-selected threshold. |
| Stroke | de la Rosa *et al.*, 2025 [9] | DeepISLES challenge-derived ensemble | External validation outperformed the prior state of the art by 7.4% Dice and 12.6% F1 | Externally validated stroke ensemble; BrainMD evaluates an internal split within a cross-pathology workflow. |
| Sclerosis | Rondinella *et al.*, 2024 [33] | ICPR MSLesSeg competition; baseline and follow-up MRI from multiple hospitals | Independent lesion segmentation for an unseen heterogeneous cohort | Multi-site generalization setting; BrainMD measures same-split attention–decoder continuation. |
| Sclerosis | Guarnera *et al.*, 2025 [12] | MSLesSeg dataset and benchmark | 115 scans from 75 patients with expert-validated labels | Distinct dataset and sampling scale from BrainMD's 60-subject public MS release. |

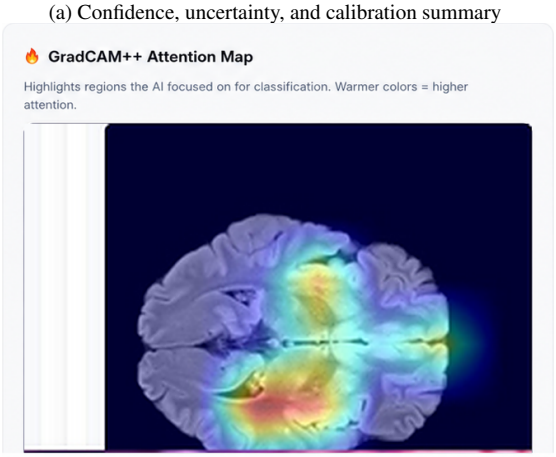
### 3.12. Review Evidence and Workflow Support

The interface evidence documents confidence, visual explanation, rule-check outcomes, and returned masks; the fixed-split measurements remain the primary performance evidence.

Figure 11(a) shows the confidence, uncertainty, and calibration summary, and Figure 11(b) shows the Grad-CAM attention map. Together, the panels separate numerical confidence from spatial and layer-wise visual evidence.



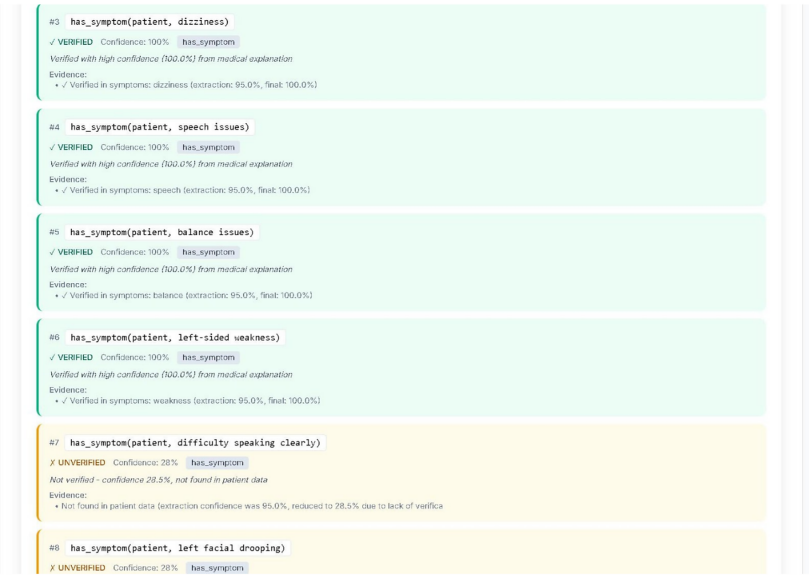

(a) Confidence, uncertainty, and calibration summary




(b) Grad-CAM attention map

Figure 11. English explanation view used for visual inspection of the model output. Panel (a) reports confidence and calibration; panel (b) shows the Grad-CAM map.

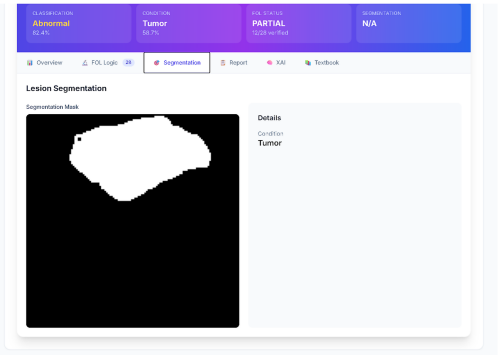
Figure 12(a) shows the outcome of individual rule-based consistency checks before the pathology-specific output views.



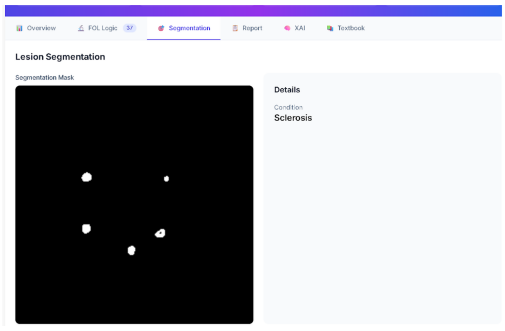

(a) Generated consistency-check outcomes

Figure 12. Rule-based consistency-check view. Panel (a) shows individual check outcomes before the pathology-specific outputs.

Figures 12(b) and 12(c) show the English tumor and sclerosis segmentation masks, respectively, so the reader can compare the condition label with the displayed lesion mask.




(b) Tumor segmentation mask




(c) Sclerosis segmentation mask

Figure 12. Rule-based consistency checking and English pathology-specific output views (continued). Panels (b) and (c) show the tumor and sclerosis masks in the clinician-facing interface.

## 4. CONCLUSION

BrainMD coordinates screening, specialist selection under uncertainty, and traceable review across three pathologies. The frozen screen escalated all 69 abnormal volumes while exiting all 88 normal volumes, avoiding downstream inference for 56.1% of the 157-subject cohort. Perceptual-group image development reached 98.9% sensitivity and 94.0% specificity; the gate intercepted all nine raw false-normal classifications while retaining 70.4% normal exit yield. In the separate held-out routing experiment, the policy retained the correct branch in all 60 cases and reduced scheduled specialist invocations by 44.4%. Subject-disjoint tumor evaluation reached pooled WT Dice 0.895, while the stroke and sclerosis studies quantify branch-specific performance and continuation gains. The matched refinement experiment isolates time-constant adaptivity under equal capacity and preserves competitive whole-tumor performance. Together, the algorithms, verified serving path, and benchmark evidence establish a reproducible workflow for coordinating specialist computation and traceable review across three pathologies. Next priorities are contrast-matched screening and routing, acquisition perturbations, missing modalities, stroke refinement, and clinical evidence review.

## ACKNOWLEDGMENTS

The authors acknowledge the open-source medical imaging community and the providers of the IXI dataset (`https://brain-development.org/ixi-dataset/`), BraTS2018/2020, the ISLES 2022 public training release, and the Muslim *et al.* multiple-sclerosis MRI dataset with consensus lesion annotations used in this evaluation.

## FUNDING INFORMATION (MANDATORY)

The authors declare that no external funding was received for this work.

## AUTHOR CONTRIBUTIONS STATEMENT (MANDATORY)

Meeradevi: Investigation, supervision, and writing (review, editing and revisions).

Kshiraja Nelapati: Conceptualization, methodology, software, formal analysis, investigation, data curation, writing (original draft), writing (review and editing), and project administration.

Prathmesh Sayal: Conceptualization, software, validation, visualization, project administration, and writing (original draft).

## CONFLICT OF INTEREST STATEMENT (MANDATORY)

The authors declare no conflict of interest.

## DATA AVAILABILITY (MANDATORY)

The benchmark datasets used in this study are publicly available from their respective organizers or public repositories. The stroke experiment used the 250-case ISLES 2022 public training release, available as Zenodo record 7960856 (doi: 10.5281/zenodo.7960856), rather than the hidden challenge test cohort. The representative held-out stroke case shown in Figure 10(b) is sub-strokecase0219_ses-0001. The sclerosis experiment used Muslim *et al.*'s Mendeley Data version 1, dataset 8bctsm8jz7; its displayed held-out example is Patient-31. BraTS challenge data remain subject to their original access terms. The fixed partitions, screening content-hash manifest, trained screening checkpoint, hash-bound gate policy, per-case probabilities and decisions, and derived specialist predictions are available from the corresponding author on reasonable request, subject to the source data terms.

## ETHICS DECLARATION

This study used de-identified public benchmark datasets and public repository data released by their original providers for research use. No new patient recruitment, intervention, or direct contact with human participants was performed by the authors. Ethical approval, consent, and data-governance procedures for the original datasets were handled by the dataset creators and challenge organizers; no additional institutional approval was required for this retrospective secondary analysis of publicly available de-identified data.

## REFERENCES

[1] M. Ghassemi *et al.*, "A review of challenges and opportunities in machine learning for health," *AMIA Jt. Summits Transl. Sci. Proc.*, pp. 191–200, 2020.

[2] A. J. Thirunavukarasu *et al.*, "Large language models in medicine," *Nat. Med.*, vol. 19, pp. 1930–1938, 2023, doi: 10.1038/s41591-023-02448-8.

[3] M. Wornow *et al.*, "The shaky foundations of large language models and foundation models for electronic health records," *npj Digit. Med.*, vol. 6, Art. no. 135, 2023, doi: 10.1038/s41746-023-00879-8.

[4] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021, doi: 10.1038/s41592-020-01008-z.

[5] A. Hatamizadeh *et al.*, "UNETR: transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 574–584.

[6] J. Chen *et al.*, "TransUNet: transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[7] J. Ma *et al.*, "Segment anything in medical images," *Nat. Commun.*, vol. 15, Art. no. 654, 2024, doi: 10.1038/s41467-024-44824-z.

[8] M. R. Hernandez Petzsche *et al.*, "ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset," *Sci. Data*, vol. 9, Art. no. 762, 2022, doi: 10.1038/s41597-022-01875-5.

[9] E. de la Rosa, M. Reyes, S.-L. Liew, A. Hutton, R. Wiest, and J. Kaesmacher, "DeepISLES: a clinically validated ischemic stroke segmentation model from the ISLES'22 challenge," *Nat. Commun.*, vol. 16, no. 1, 2025, doi: 10.1038/s41467-025-62373-x.

[10] A. M. Muslim *et al.*, "Brain MRI dataset of multiple sclerosis with consensus manual lesion segmentation and patient meta information," *Data Brief*, vol. 42, Art. no. 108139, 2022, doi: 10.1016/j.dib.2022.108139.
[11] J. Ye *et al.*, "New Multiple Sclerosis Lesion Segmentation via Calibrated Inter-patch Blending," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, *Lect. Notes Comput. Sci.*, pp. 365–375, 2026, doi: 10.1007/978-3-032-05325-1_35.
[12] F. Guarnera *et al.*, "MSLesSeg: baseline and benchmarking of a new Multiple Sclerosis Lesion Segmentation dataset," *Sci. Data*, vol. 12, Art. no. 920, 2025, doi: 10.1038/s41597-025-05250-y.
[13] S. Azizi *et al.*, "Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging," *Nature Biomed. Eng.*, vol. 7, no. 6, pp. 756–779, 2023, doi: 10.1038/s41551-023-01049-7.
[14] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: methodological failures and recommendations for the future," *npj Digit. Med.*, vol. 5, Art. no. 48, 2022, doi: 10.1038/s41746-022-00592-y.
[15] J. Huang *et al.*, "Brain tumor segmentation using deep learning: high performance with minimized MRI data," *Front. Radiol.*, vol. 5, 2025, doi: 10.3389/fradi.2025.1616293.
[16] A. M. Labella *et al.*, "The ASNR-MICCAI BraTS 2023 Intracranial Meningioma Segmentation Challenge," in *Proc. BrainLes Workshop*, *Lect. Notes Comput. Sci.*, vol. 14668, pp. 1–14, 2024, doi: 10.1007/978-3-031-76163-8_1.
[17] R. Hasani, M. Lechner, A. Amini, D. Rus, and R. Grosu, "Liquid time-constant networks," in *Proc. AAAI Conf. Artif. Intell.*, 2021.
[18] M. Lechner *et al.*, "Neural circuit policies enabling auditable autonomy," *Nature Mach. Intell.*, vol. 2, no. 10, pp. 642–652, 2020, doi: 10.1038/s42256-020-00237-3.
[19] R. Hasani *et al.*, "Closed-form continuous-time neural networks," *Nat. Mach. Intell.*, vol. 4, pp. 739–753, 2022, doi: 10.1038/s42256-022-00556-7.
[20] S. Zhang *et al.*, "BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023, doi: 10.48550/arXiv.2303.00915.
[21] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Adv. Neural Inf. Process. Syst.*, 2023.
[22] C. Li *et al.*, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," in *Adv. Neural Inf. Process. Syst.*, 2023.
[23] J. S. Ryu, H. Kang, Y. Chu, and S. Yang, "Vision-language foundation models for medical imaging: a review of current practices and innovations," *Biomed. Eng. Lett.*, vol. 15, no. 5, pp. 809–830, 2025, doi: 10.1007/s13534-025-00484-6.
[24] M. Moor *et al.*, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023, doi: 10.1038/s41586-023-05881-4.
[25] B. M. de Vries *et al.*, "Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review," *Front. Med.*, vol. 10, Art. no. 1180773, 2023, doi: 10.3389/fmed.2023.1180773.
[26] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023, doi: 10.1038/s41586-023-06291-2.
[27] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, and F. Nensa, "Explainable AI in medical imaging: an overview for clinical practitioners—beyond saliency-based XAI approaches," *Eur. J. Radiol.*, vol. 162, Art. no. 110786, 2023, doi: 10.1016/j.ejradiol.2023.110786.
[28] A. Howard *et al.*, "Searching for MobileNetV3," *arXiv preprint arXiv:1905.02244*, 2019, doi: 10.48550/arXiv.1905.02244.
[29] Imperial College London, "IXI Dataset," Brain Development. [Online]. Available: `https://brain-development.org/ixi-dataset/`. Accessed: Aug. 31, 2026.
[30] D. Dogru, M. A. Ozdemir, and O. Guren, "Deep learning for automated ischemic stroke lesion segmentation from multi-spectral MRI," in *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*, pp. 1392–1396, 2024, doi: 10.23919/EUSIPCO63174.2024.10715216.
[31] N. Jazzar, B. Mabrouk, and A. Douik, "Transformer Dil-DenseUnet: an advanced architecture for stroke segmentation," *J. Imaging*, vol. 10, no. 12, Art. no. 304, 2024, doi: 10.3390/jimaging10120304.
[32] A. Rahman *et al.*, "Deep learning-driven segmentation of ischemic stroke lesions using multi-channel MRI," *Biomed. Signal Process. Control*, vol. 105, Art. no. 107676, 2025, doi: 10.1016/j.bspc.2025.107676.
[33] A. Rondinella *et al.*, "ICPR 2024 competition on multiple sclerosis lesion segmentation: methods and results," *arXiv preprint arXiv:2410.07924*, 2024, doi: 10.48550/arXiv.2410.07924.
[34] T. Zhou, "M2GCNet: multi-modal graph convolution network for precise brain tumor segmentation across multiple MRI sequences," *IEEE Trans. Image Process.*, vol. 33, pp. 4896–4910, 2024, doi: 10.1109/TIP.2024.3451936.

## BIOGRAPHIES OF AUTHORS

Kshiraja Nelapati is affiliated with the Department of Artificial Intelligence and Machine Learning, Ramaiah Institute of Technology, Bengaluru, India. Her research interests include medical image analysis, explainable artificial intelligence, and clinical decision-support systems. She can be contacted at email: 1ms24ai031@msrit.edu. Profiles: LinkedIn.com/in/kshiraja-nelapati.

Dr. Meeradevi is an Associate Professor in the Department of Artificial Intelligence and Machine Learning, Ramaiah Institute of Technology, Bengaluru, India. Her research interests include artificial intelligence, machine learning, networks, security, and deep learning. She can be contacted at email: meera_ak@msrit.edu. Profiles: .

Prathmesh Sayal is pursuing a Bachelor of Engineering in Computer Science with a specialization in Cybersecurity at Ramaiah Institute of Technology, Bengaluru, India. His research interests include medical AI systems, cybersecurity, data analysis, and practical software engineering. He can be contacted at email: 1ms23cy049@msrit.edu. Profiles: .