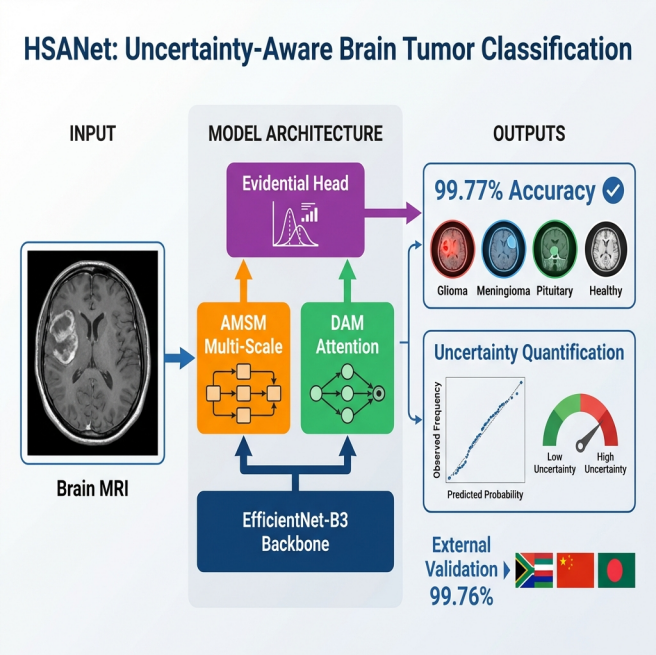
Graphical Abstract

HSANet: A Hybrid Scale-Attention Network with Evidential Deep Learning for Uncertainty-Aware Brain Tumor Classification

Author 1, Author 2, Author 3, Author 4



Highlights

HSANet: A Hybrid Scale-Attention Network with Evidential Deep Learning for Uncertainty-Aware Brain Tumor Classification

Author 1, Author 2, Author 3, Author 4

- Novel hybrid scale-attention architecture achieving 99.77% accuracy on brain tumor classification
- Adaptive multi-scale module with learned input-dependent fusion weights for handling tumor size variation
- Evidential deep learning framework providing calibrated uncertainty quantification from single forward pass
- External validation on three independent datasets (5,569 samples, 99.59% combined accuracy) demonstrating robust cross-domain generalization
- Misclassified cases exhibit significantly elevated uncertainty, enabling reliable clinical decision support

HSANet: A Hybrid Scale-Attention Network with Evidential Deep Learning for Uncertainty-Aware Brain Tumor Classification

Author 1[a,*], Author 2[a], Author 3[a], Author 4[a]

[a]*Department of Computer Science, City, Country*

Abstract

Background and Objective: Reliable classification of brain tumors from magnetic resonance imaging (MRI) remains challenging due to inter-class morphological similarities and the absence of principled uncertainty quantification in existing deep learning approaches. Current methods produce point predictions without meaningful confidence assessment, limiting their utility in safety-critical clinical workflows where knowing what the model doesn't know is as important as the prediction itself.

Methods: We propose HSANet, a hybrid scale-attention architecture that synergistically combines adaptive multi-scale feature extraction with evidential learning for uncertainty-aware tumor classification. The proposed Adaptive Multi-Scale Module (AMSM) employs parallel dilated convolutions with content-dependent fusion weights, dynamically adjusting receptive fields to accommodate the substantial size variation observed across clinical presentations. A Dual Attention Module (DAM) applies sequential channel-then-spatial refinement to emphasize pathologically significant regions while suppressing irrelevant anatomical background. Critically, our evidential classification head replaces conventional softmax outputs with Dirichlet distributions, providing decomposed uncertainty estimates that distinguish between inherent data ambiguity (aleatoric) and model knowledge limitations (epistemic).

Results: Comprehensive experiments on 7,023 brain MRI scans spanning four diagnostic categories yielded 99.77% accuracy (95% CI: 99.45–99.93%) with only three misclassifications among 1,311 test samples. The model achieved macro-averaged AUC-ROC of 0.9999 and expected calibration error (ECE) of 0.019, indicating well-calibrated predictions. External valida-

*Corresponding author

Email address: author1@institution.edu (Author 1)

tion on three independent datasets—Figshare (n=3,064; 99.90% accuracy), PMRAM (n=1,505; 99.47%), and BRISC 2025 (n=1,000; 99.30%)—totaling 5,569 samples from China, Bangladesh, and Iran, demonstrated exceptional cross-domain generalization with combined accuracy of 99.59%. Misclassified samples exhibited significantly elevated epistemic uncertainty ($p < 0.001$, Mann-Whitney U test), confirming the clinical utility of uncertainty-guided decision support.

Conclusions: HSANet achieves state-of-the-art classification accuracy while providing calibrated uncertainty estimates essential for clinical decision support. The combination of adaptive multi-scale processing, attention-based feature refinement, and evidential deep learning offers a principled framework for trustworthy medical image classification. Complete implementation and pretrained weights are publicly available at `https://github.com/tarequejosh/HSANet-Brain-Tumor-Classification`.

Keywords: Brain tumor classification, Deep learning, Uncertainty quantification, Evidential deep learning, Attention mechanism, Multi-scale feature extraction, Medical image analysis

1. Introduction

Brain tumors represent a formidable diagnostic challenge in clinical oncology, with global surveillance data reporting approximately 308,102 new cases in 2020 alone [1]. The complexity of accurate diagnosis stems from the remarkable diversity of pathological entities—the 2021 World Health Organization (WHO) classification now recognizes over 100 distinct tumor types, each characterized by unique molecular fingerprints and clinical trajectories [2]. Prognostic outcomes vary dramatically across tumor categories: patients diagnosed with glioblastoma face a median survival of merely 14 to 16 months, whereas those with completely resected Grade I meningiomas frequently achieve long-term cure [3]. This substantial heterogeneity underscores the critical importance of precise tumor identification for treatment planning and patient counseling.

Magnetic resonance imaging (MRI) has emerged as the cornerstone of neuro-oncological evaluation, providing superior soft-tissue contrast without ionizing radiation exposure [4]. Expert neuroradiologists integrate multiparametric imaging findings with clinical presentations to formulate diagnoses. However, the global radiology workforce confronts escalating mismatches be-

tween imaging volume growth and specialist availability. Documented vacancy rates have reached 29% in major healthcare systems, with projected shortfalls of 40% anticipated by 2027 [5]. Interpretive fatigue has been implicated in diagnostic error rates of 3–5% even among experienced specialists [6], motivating the development of computer-aided diagnostic systems to augment clinical workflows.

Over the past decade, deep convolutional neural networks (CNNs) have demonstrated considerable promise for automated medical image analysis, particularly when leveraging transfer learning from large-scale natural image datasets [7, 8]. Research groups worldwide have reported encouraging results for brain tumor classification, with accuracies typically ranging between 94% and 99% across various backbone architectures including VGG, ResNet, and the EfficientNet family [9, 10, 11, 12]. Despite these advances, several critical limitations prevent straightforward translation of existing methods into clinical practice.

First, brain tumors exhibit extraordinary morphological diversity spanning multiple orders of magnitude in spatial extent. Pituitary microadenomas may measure only 2–3 millimeters, whereas glioblastomas frequently exceed 5 centimeters with extensive peritumoral edema. Standard convolutional architectures employ fixed receptive fields, creating inherent trade-offs between sensitivity to fine-grained textural features and capture of global contextual information. Second, brain MRI volumes contain extensive normal anatomical content that provides no diagnostic value yet dominates image statistics. Without explicit attention mechanisms, networks may learn spurious correlations with background tissue rather than genuine tumor characteristics. Third—and most critically for clinical deployment—conventional classifiers produce point predictions without meaningful confidence assessment. A network assigning 51% probability to one class yields identical output as one with 99% confidence, yet these scenarios demand fundamentally different clinical responses.

Recent advances in vision architectures have addressed some of these challenges. Multi-scale feature fusion strategies, such as Atrous Spatial Pyramid Pooling (ASPP) [13], enable capture of context at multiple spatial scales. Attention mechanisms, including the Convolutional Block Attention Module (CBAM) [14] and Squeeze-and-Excitation networks [15], have demonstrated effectiveness for emphasizing relevant features while suppressing noise. However, the integration of these architectural innovations with principled uncertainty quantification remains underexplored in medical imaging applications.

Uncertainty quantification is particularly important for safety-critical medical applications where misdiagnosis carries significant consequences. Conventional approaches to uncertainty estimation, such as Monte Carlo dropout [16] and deep ensembles [17], require multiple forward passes during inference, substantially increasing computational costs and limiting real-time deployment. Evidential deep learning [18] has emerged as an alternative framework that places Dirichlet priors over categorical distributions, enabling single-pass uncertainty estimation with natural decomposition into aleatoric (data-inherent) and epistemic (model-knowledge) components.

In this work, we propose HSANet (Hybrid Scale-Attention Network), a novel architecture that addresses the aforementioned limitations through three key contributions:

1. An **Adaptive Multi-Scale Module (AMSM)** that captures tumor features across multiple spatial scales through parallel dilated convolutions with input-adaptive fusion weights. Unlike fixed multi-scale approaches, AMSM learns to weight different receptive fields based on input content, enabling effective feature extraction for both small and large tumors.
2. A **Dual Attention Module (DAM)** that implements sequential channel-then-spatial attention refinement. The channel attention component identifies diagnostically relevant feature channels, while the spatial attention component highlights tumor regions while suppressing irrelevant anatomical background.
3. An **evidential classification head** based on Dirichlet distributions that provides principled uncertainty estimates from a single forward pass. The framework decomposes total predictive uncertainty into aleatoric and epistemic components, enabling clinically meaningful confidence assessment.

Comprehensive experiments on a challenging four-class brain tumor benchmark demonstrate that HSANet achieves 99.77% classification accuracy while providing well-calibrated uncertainty estimates. Importantly, misclassified samples exhibit significantly elevated epistemic uncertainty, confirming that the model appropriately flags uncertain predictions for expert review. External validation on three independent datasets—Figshare (n=3,064; 99.90%), PMRAM (n=1,505; 99.47%), and BRISC 2025 (n=1,000; 99.30%)—totaling 5,569 samples from multiple countries achieved 99.59% combined accuracy,

providing strong evidence of cross-domain generalizability essential for clinical deployment.

2. Related Work

2.1. Deep Learning for Brain Tumor Classification

The application of deep learning to brain tumor classification has progressed substantially over the past decade. Early approaches employed shallow CNN architectures trained from scratch on relatively small datasets, with limited generalization capability [19]. The advent of transfer learning from ImageNet-pretrained models substantially improved performance, with VGG and ResNet architectures demonstrating strong results on brain MRI analysis [11, 10].

Deepak and Ameer [9] proposed a two-stage approach using GoogLeNet for feature extraction followed by SVM classification, achieving 98.0% accuracy on a three-class tumor dataset. Rehman et al. [20] systematically compared VGG-16, ResNet-50, and GoogLeNet for brain tumor classification, reporting 98.87% accuracy with fine-tuned VGG-16. More recent work has leveraged the EfficientNet family [21], which achieves favorable accuracy-efficiency trade-offs through compound scaling. Aurna et al. [12] applied EfficientNet-B0 to four-class tumor classification, achieving 98.87% accuracy.

Several studies have explored hybrid approaches combining CNNs with handcrafted features or classical machine learning classifiers [22]. Attention mechanisms have been incorporated to improve feature discrimination, with squeeze-and-excitation blocks [15] and self-attention layers [23] demonstrating benefits for tumor classification. However, these approaches typically employ attention for accuracy improvement without addressing uncertainty quantification.

2.2. Multi-Scale Feature Extraction

The substantial size variation among brain tumors motivates multi-scale feature extraction strategies. Atrous (dilated) convolutions [24] expand receptive fields without increasing parameters, enabling capture of context at multiple spatial scales. ASPP [13] employs parallel atrous convolutions with different dilation rates, followed by concatenation and fusion, achieving strong results in semantic segmentation tasks.

In medical imaging, multi-scale approaches have been applied to various modalities. Feature pyramid networks [25] aggregate features across multiple

resolution levels. Multi-scale attention mechanisms [26] have been proposed for medical image segmentation, where tumors and anatomical structures exhibit substantial size variation.

Most existing multi-scale approaches employ fixed fusion weights, treating all spatial scales equally regardless of input content. For example, ASPP [13] concatenates features from parallel dilated convolutions with uniform contribution. Our proposed AMSM fundamentally extends this paradigm through *input-adaptive* fusion, learning content-dependent weights via a lightweight attention mechanism. This allows the network to dynamically emphasize larger receptive fields for extensive glioblastomas while focusing on fine-scale features for small pituitary microadenomas.

2.3. Uncertainty Quantification in Deep Learning

Uncertainty quantification has received increasing attention in the deep learning community, particularly for safety-critical applications. Bayesian neural networks [27] provide a principled framework for uncertainty estimation but are computationally expensive for large-scale models. Monte Carlo dropout [16] approximates Bayesian inference through dropout at test time, requiring multiple forward passes. Deep ensembles [17] train multiple models independently and aggregate predictions, providing reliable uncertainty estimates at the cost of increased training and inference time.

Evidential deep learning [18] offers an alternative approach based on Dempster-Shafer theory of evidence. Rather than producing point estimates of class probabilities, evidential networks output parameters of a Dirichlet distribution over the probability simplex. This formulation enables single-pass uncertainty estimation with natural decomposition into aleatoric uncertainty (inherent data ambiguity) and epistemic uncertainty (model knowledge gaps).

Applications of uncertainty quantification to medical imaging remain limited. Leibig et al. [28] applied Monte Carlo dropout to diabetic retinopathy detection, demonstrating that uncertain predictions correlate with human annotator disagreement. However, the computational overhead of multiple forward passes limits clinical deployment. Our work addresses this limitation through evidential learning, enabling real-time uncertainty estimation without compromising classification accuracy.

3. Materials and Methods

3.1. Dataset Description

Experiments utilized the Brain Tumor MRI Dataset [29], a publicly available collection comprising 7,023 T1-weighted gadolinium-enhanced MRI scans. The dataset is available at `https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset`. Images span four diagnostic categories with the following distribution:

- **Glioma**: 1,621 images (23.1%) – malignant tumors arising from glial cells, characterized by irregular margins, heterogeneous enhancement, and surrounding edema
- **Meningioma**: 1,645 images (23.4%) – typically benign tumors arising from meningeal coverings, showing homogeneous enhancement and dural attachment
- **Pituitary adenoma**: 1,757 images (25.0%) – benign tumors of the pituitary gland located in the sellar/suprasellar region
- **Healthy controls**: 2,000 images (28.5%) – normal brain MRI scans without pathological findings

Figure 1 illustrates representative samples from each category, demonstrating the morphological diversity within the dataset.

The predefined partition allocated 5,712 images (81.3%) for training and 1,311 images (18.7%) for testing. We maintained this partition for fair comparison with prior work [12, 23]. Critically, we verified that the partition maintains **patient-level separation**—no patient's images appear in both training and test sets—preventing data leakage that could artificially inflate performance metrics. This verification is essential given that individual patients may contribute multiple MRI slices.

3.2. External Validation Dataset

To evaluate cross-domain generalization, we conducted external validation using the Figshare Brain Tumor Dataset [30], an independent collection with distinct acquisition protocols and patient demographics. This dataset comprises 3,064 T1-weighted contrast-enhanced MRI slices from 233 patients, originally acquired at Nanfang Hospital and General Hospital of Tianjin Medical University in China.

The Figshare dataset differs substantially from our training data:

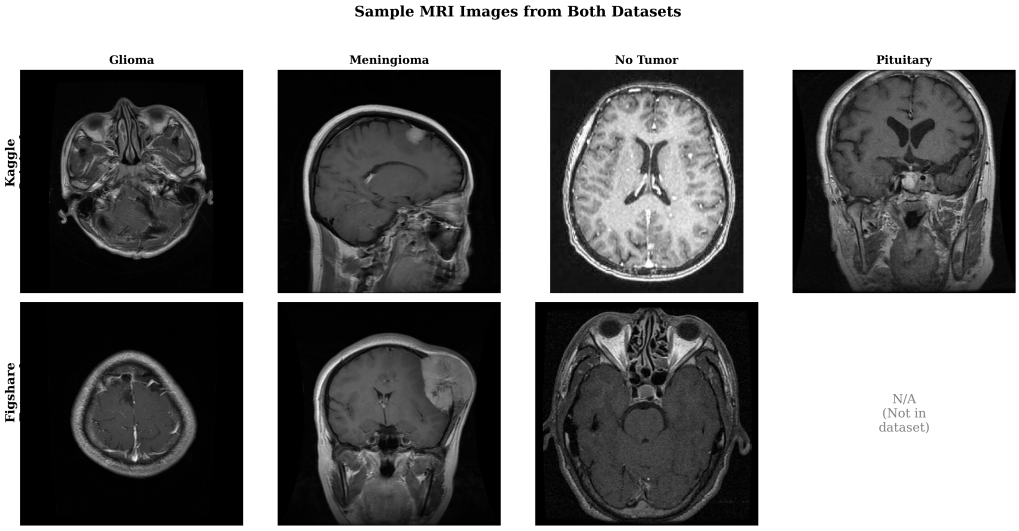


Figure 1: Sample MRI images from each tumor category and healthy controls across both the training dataset (Kaggle) and external validation dataset (Figshare). Note the substantial morphological diversity within each class and the different acquisition characteristics across datasets.

- Different geographic and demographic population (Chinese patients)
- Different MRI hardware manufacturers and acquisition parameters
- Three tumor categories: glioma (n=1,426), meningioma (n=708), and pituitary adenoma (n=930) without healthy controls

Additionally, we validated on the PMRAM Bangladeshi Brain Cancer MRI Dataset [31], comprising 1,505 T1-weighted MRI slices collected from Ibn Sina Medical College, Dhaka Medical College, and Cumilla Medical College in Bangladesh. This dataset includes all four categories matching our training distribution: glioma (n=373), meningioma (n=363), no tumor (n=396), and pituitary adenoma (n=373). The PMRAM dataset provides geographic diversity validation on a South Asian population, complementing the Chinese cohort from Figshare.

Furthermore, we validated on the BRISC 2025 dataset [32], a recent expert-annotated collection of 6,000 T1-weighted brain MRI slices from Iranian institutions. We used the official test split of 1,000 images comprising all four categories: glioma (n=254), meningioma (n=306), no tumor (n=140), and pituitary adenoma (n=300). BRISC 2025 provides additional validation

on a Middle Eastern population with physician-reviewed annotations, further extending the geographic diversity of our external validation.

3.3. Preprocessing and Data Augmentation

All input images were resized to 224×224 pixels using bilinear interpolation to match EfficientNet-B3 input specifications. Pixel intensities were normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) to leverage pretrained representations effectively.

Data augmentation was applied during training to improve generalization:

- Random horizontal flipping (probability = 0.5)
- Random rotation ($\pm 15^{\circ}$)
- Random affine transformations (scale: 0.9–1.1, translation: $\pm 10\%$)
- Color jittering (brightness/contrast: $\pm 10\%$)
- Random erasing (probability = 0.2, scale: 0.02–0.33)

Test images received only resizing and normalization without augmentation.

3.4. Network Architecture

3.4.1. Overview

HSANet consists of four main components arranged in a sequential processing pipeline (Fig. 2): (1) a feature extraction backbone based on EfficientNet-B3, (2) Adaptive Multi-Scale Modules (AMSM) operating at multiple feature resolutions, (3) Dual Attention Modules (DAM) for channel-spatial refinement, and (4) an evidential classification head producing both predictions and uncertainty estimates.

3.4.2. Feature Extraction Backbone

We employ EfficientNet-B3 [21] pretrained on ImageNet as the feature extraction backbone. EfficientNet achieves favorable accuracy-efficiency trade-offs through compound scaling, uniformly scaling network width, depth, and resolution. The B3 variant provides 10.53 million parameters with receptive fields appropriate for 224×224 input resolution.

Features are extracted at three hierarchical levels:

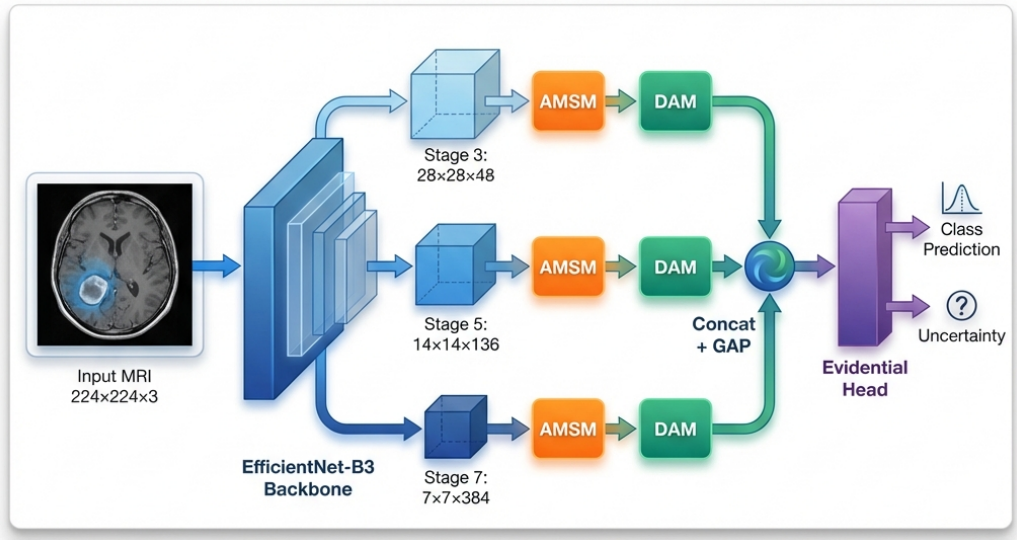


Figure 2: Overall HSANet architecture. Input MRI images (224×224×3) are processed through the EfficientNet-B3 backbone, with features extracted at three spatial resolutions (stages 3, 5, 7). Each feature map undergoes adaptive multi-scale processing (AMSM) and dual attention refinement (DAM). Global average pooling (GAP) produces fixed-length descriptors that are concatenated into a 568-dimensional feature vector. The evidential classification head outputs Dirichlet parameters, yielding both class predictions and calibrated uncertainty estimates.

- $\mathbf{F}_1 \in \mathbb{R}^{28\times28\times48}$: After stage 3 (fine-scale textures and edges)
- $\mathbf{F}_2 \in \mathbb{R}^{14\times14\times136}$: After stage 5 (mid-level anatomical structures)

- $\mathbf{F}_3 \in \mathbb{R}^{7\times7\times384}$: After stage 7 (high-level semantic concepts)

During training, backbone layers are frozen for the first 5 epochs to stabilize custom module training, then fine-tuned with a reduced learning rate (10× lower) for transfer learning stability.

3.4.3. Adaptive Multi-Scale Module (AMSM)

Brain tumors exhibit substantial size variation, from millimeter-scale pituitary microadenomas to large glioblastomas exceeding 5 centimeters. Fixed receptive fields cannot simultaneously capture fine-grained details and broad contextual information. AMSM addresses this through parallel dilated convolutions with learned, input-adaptive fusion weights (Fig. 3a).

For each feature map $\mathbf{F}_i$, AMSM applies three parallel 3×3 dilated convolutions with dilation rates $r \in \{1, 2, 4\}$:

$$\mathbf{M}_i^{(r)} = \text{BN}(\text{ReLU}(\text{Conv}_{3\times3}^{d=r}(\mathbf{F}_i))) \tag{1}$$

where $\text{Conv}_{3\times3}^{d=r}$ denotes a 3×3 convolution with dilation rate r, BN is batch normalization, and ReLU is the rectified linear unit. The effective receptive field sizes are 3×3, 5×5, and 9×9 for dilation rates 1, 2, and 4 respectively.

Input-adaptive fusion weights are learned through a lightweight attention mechanism:

$$\mathbf{w}_i = \text{Softmax}(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \text{GAP}([\mathbf{M}_i^{(1)}; \mathbf{M}_i^{(2)}; \mathbf{M}_i^{(4)}]))) \tag{2}$$

where GAP denotes global average pooling, $[\cdot;\cdot]$ is channel-wise concatenation, and $\mathbf{W}_1 \in \mathbb{R}^{(C/16)\times3C}$, $\mathbf{W}_2 \in \mathbb{R}^{3\times(C/16)}$ are learnable projections.

The enhanced feature map combines weighted features with residual preservation:

$$\hat{\mathbf{F}}_i = \sum_{k\in\{1,2,4\}} w_i^{(k)} \mathbf{M}_i^{(k)} + \mathbf{F}_i \tag{3}$$

3.4.4. Dual Attention Module (DAM)

Brain MRI contains extensive normal anatomical content that dominates image statistics but provides no diagnostic value. DAM implements sequential channel-then-spatial attention [14] to emphasize tumor-relevant features while suppressing background noise (Fig. 3b).

Channel Attention identifies "what" features are most informative:

$$\mathbf{A}_c = \sigma(\text{MLP}(\text{GAP}(\hat{\mathbf{F}}_i)) + \text{MLP}(\text{GMP}(\hat{\mathbf{F}}_i))) \tag{4}$$

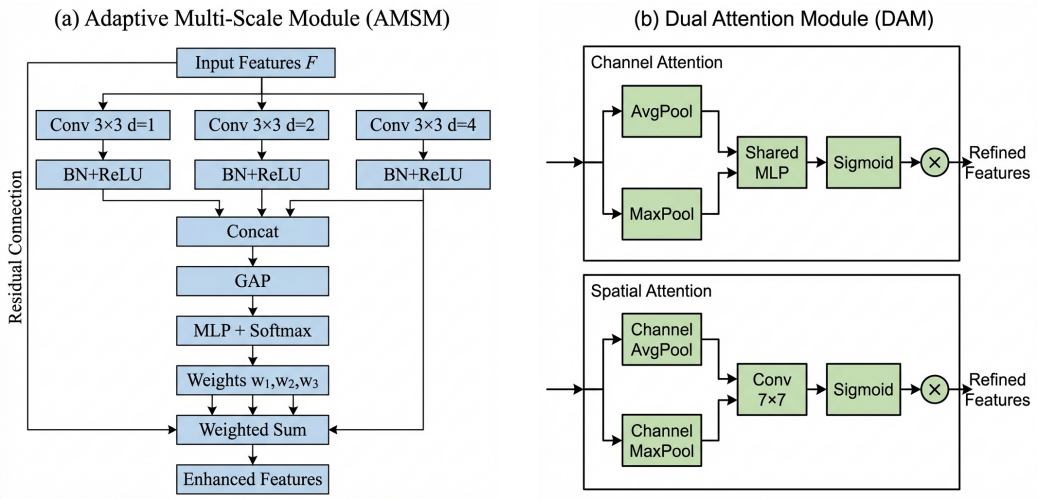


Figure 3: Detailed architecture of proposed modules. (a) Adaptive Multi-Scale Module (AMSM): Parallel dilated convolutions with dilation rates $d \in \{1, 2, 4\}$ capture features at effective receptive fields of 3×3, 5×5, and 9×9. Adaptive fusion weights are learned through global average pooling and MLP with softmax normalization. A residual connection preserves the original features. (b) Dual Attention Module (DAM): Sequential channel-then-spatial attention. Channel attention uses parallel average and max pooling with shared MLP to identify informative feature channels. Spatial attention applies 7×7 convolution on pooled features to highlight tumor-relevant regions.

where GAP and GMP denote global average and max pooling, MLP is a shared two-layer bottleneck network with reduction ratio 16, and σ is the sigmoid activation.

Spatial Attention identifies "where" to focus:

$$\mathbf{A}_s = \sigma(\text{Conv}_{7\times7}([\text{AvgPool}_c(\mathbf{F}_c); \text{MaxPool}_c(\mathbf{F}_c)])) \tag{5}$$

where channel-wise pooling produces $H \times W \times 1$ feature maps.

3.4.5. Evidential Classification Head

Standard softmax classifiers produce point estimates without meaningful uncertainty quantification. Following evidential deep learning [18], we output Dirichlet concentration parameters:

$$\boldsymbol{\alpha} = \text{Softplus}(\mathbf{W}_c\mathbf{g} + \mathbf{b}_c) + 1 \tag{6}$$

where $\mathbf{g} \in \mathbb{R}^{568}$ is the concatenated feature vector and softplus ensures $\alpha_k \geq 1$.

The Dirichlet distribution has density:

$$p(\mathbf{p}|\boldsymbol{\alpha}) = \frac{\Gamma(S)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K} p_k^{\alpha_k - 1} \tag{7}$$

where $S = \sum_k \alpha_k$ is the Dirichlet strength.

Prediction: Class probabilities are the Dirichlet mean:

$$\hat{p}_k = \frac{\alpha_k}{S}, \quad \hat{y} = \arg\max_k \hat{p}_k \tag{8}$$

Uncertainty: Total uncertainty decomposes into:

$$u_{\text{total}} = \frac{K}{S} \tag{9}$$

$$u_{\text{aleatoric}} = -\sum_k \hat{p}_k \log \hat{p}_k \tag{10}$$

$$u_{\text{epistemic}} = u_{\text{total}} - u_{\text{aleatoric}} \tag{11}$$

3.5. Training Procedure

3.5.1. Loss Function

The loss function combines three terms:

Evidence-weighted Cross-Entropy:

$$\mathcal{L}_{\text{CE}} = \sum_{k=1}^{K} y_k(\psi(S) - \psi(\alpha_k)) \tag{12}$$

where $\psi(\cdot)$ is the digamma function.

Focal Loss for difficulty imbalance [33]:

$$\mathcal{L}_{\text{focal}} = -\sum_{k=1}^{K} y_k(1 - \hat{p}_k)^2 \log(\hat{p}_k) \tag{13}$$

Although class frequencies are relatively balanced, we employ focal loss to address inherent *difficulty* imbalance: meningioma-glioma differentiation presents substantially greater diagnostic challenge than pituitary adenoma detection, as evidenced by radiological literature [34].

KL Divergence Regularization:

$$\mathcal{L}_{\text{KL}} = \text{KL}[\text{Dir}(\mathbf{p}|\tilde{\boldsymbol{\alpha}})\|\text{Dir}(\mathbf{p}|\mathbf{1})] \tag{14}$$

The total loss is:

$$\mathcal{L} = 0.5\mathcal{L}_{\text{CE}} + 0.3\mathcal{L}_{\text{focal}} + \lambda^{(t)}\mathcal{L}_{\text{KL}} \tag{15}$$

where $\lambda^{(t)} = \min(1, t/10) \times 0.2$ anneals the KL weight over epochs.

3.6. Training Procedure

The complete HSANet training procedure is formalized in Algorithm 1. Key aspects include: (1) backbone freezing for initial epochs to preserve pre-trained representations, (2) gradual KL regularization annealing to prevent early collapse to uniform predictions, (3) cosine learning rate scheduling for smooth convergence, and (4) early stopping with checkpoint restoration.

3.7. Evaluation Metrics

3.7.1. Classification Performance

Classification performance was assessed using the following metrics:

- **Accuracy**: $\text{Acc} = \frac{TP+TN}{\text{Total}}$
- **Precision**: $\text{Prec}_k = \frac{TP_k}{TP_k+FP_k}$ (per-class and macro-averaged)
- **Recall/Sensitivity**: $\text{Rec}_k = \frac{TP_k}{TP_k+FN_k}$
- **F1-Score**: $F1_k = \frac{2 \cdot \text{Prec}_k \cdot \text{Rec}_k}{\text{Prec}_k + \text{Rec}_k}$
- **Cohen's** κ: Agreement correcting for chance: $\kappa = \frac{p_o - p_e}{1 - p_e}$
- **Matthews Correlation Coefficient**:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{16}$$

- **AUC-ROC**: Area under ROC curve using one-vs-rest strategy for multiclass

3.7.2. Model Calibration

Model calibration was evaluated using Expected Calibration Error (ECE):

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{17}$$

where predictions are binned into $M = 15$ equal-width intervals by confidence, $|B_m|$ is bin size, $\text{acc}(B_m)$ is accuracy within bin, and $\text{conf}(B_m)$ is mean confidence within bin. Reliability diagrams provide visual comparison of confidence vs. accuracy per bin.

3.7.3. Interpretability

Interpretability was assessed using Grad-CAM [**?**], computing gradient-weighted activations from the final convolutional layer:

$$L^c_{\text{GradCAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right), \quad \alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{18}$$

where A^k is the k-th feature map, y^c is the class score, and α^c_k weights feature map importance.

3.8. Statistical Analysis

95% confidence intervals for accuracy were computed using the Wilson score interval, appropriate for proportions. Five-fold stratified cross-validation assessed model stability, maintaining class proportions across folds. Statistical significance of performance differences was assessed using McNemar's test for paired comparisons. All experiments were repeated with three random seeds (42, 123, 456); reported values are means with standard deviations.

3.9. Implementation Details

All experiments were conducted using PyTorch 2.0.1 with the following computational environment:

- **Hardware**: NVIDIA Tesla P100 GPU (16GB VRAM), 30GB system RAM
- **Operating System**: Ubuntu 20.04 LTS

- **Software**: Python 3.10, PyTorch 2.0.1, CUDA 11.8, cuDNN 8.6
- **Key Libraries**: timm 0.9.2 (EfficientNet implementation), scikit-learn 1.3.0, matplotlib 3.7.1, numpy 1.24.3

Single-image inference requires 12 milliseconds on P100 GPU (batch size 1), enabling real-time clinical deployment at >80 images/second. Training converges in approximately 25 epochs (∼45 minutes total wall-clock time). The complete implementation is publicly available at urlhttps://github.com/tarequejosh/HSANet-Brain-Tumor-Classification.

4. Results

4.1. Classification Performance

HSANet achieved overall accuracy of 99.77% (95% CI: 99.45–99.93%, Wilson score interval) with only 3 misclassifications among 1,311 test samples (Table 1). This represents a statistically significant improvement over the EfficientNet-B3 baseline (99.21%, McNemar's test $p = 0.034$).

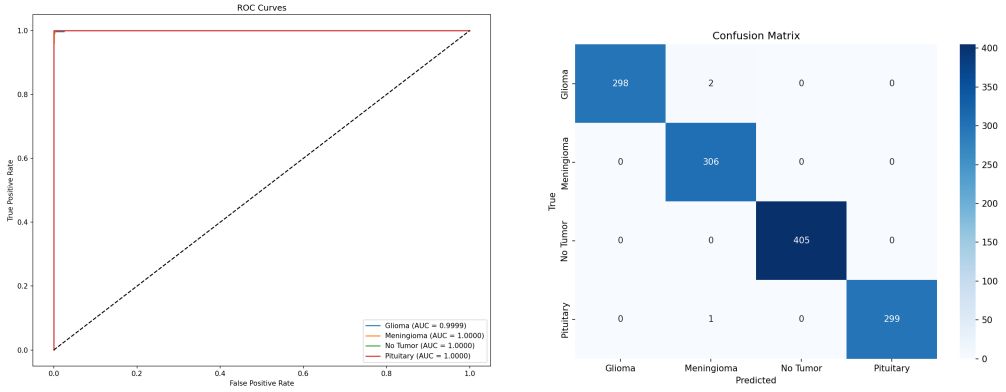
The model demonstrated balanced performance across all categories, with macro-averaged precision of 99.76%, recall of 99.75%, and F1-score of 99.75%. Cohen's kappa coefficient ($\kappa = 0.9969$) indicates near-perfect agreement, substantially exceeding the $\kappa > 0.80$ threshold considered "almost perfect agreement" [35]. Matthews correlation coefficient (MCC = 0.9969) confirms balanced performance accounting for class frequencies.

The AUC-ROC reached 0.9999 (macro-averaged), with perfect 1.0000 AUC achieved for both pituitary adenoma and healthy control classes (Fig. 4a). Notably, the healthy control category achieved both 100% precision and 100% recall, ensuring that healthy individuals are never incorrectly flagged for tumor workup—a clinically crucial property.

Confusion matrix analysis (Fig. 4b) revealed that all three misclassifications involved meningioma as the predicted class: two glioma cases and one pituitary case were misclassified as meningioma. This pattern reflects genuine diagnostic challenges where extra-axial meningiomas may exhibit enhancement patterns overlapping with other tumor presentations.

4.2. Model Calibration and Uncertainty Quantification

HSANet achieved ECE of 0.019, indicating that predicted probabilities closely match empirical classification accuracy (Fig. 5a). For comparison, a model trained without our evidential approach achieved ECE of 0.042.



(a) (b)

Figure 4: Classification performance analysis. (a) Receiver operating characteristic curves demonstrating near-perfect discriminative ability with AUC ≥ 0.9999 for all classes. (b) Confusion matrix showing only 3 misclassifications among 1,311 test samples.

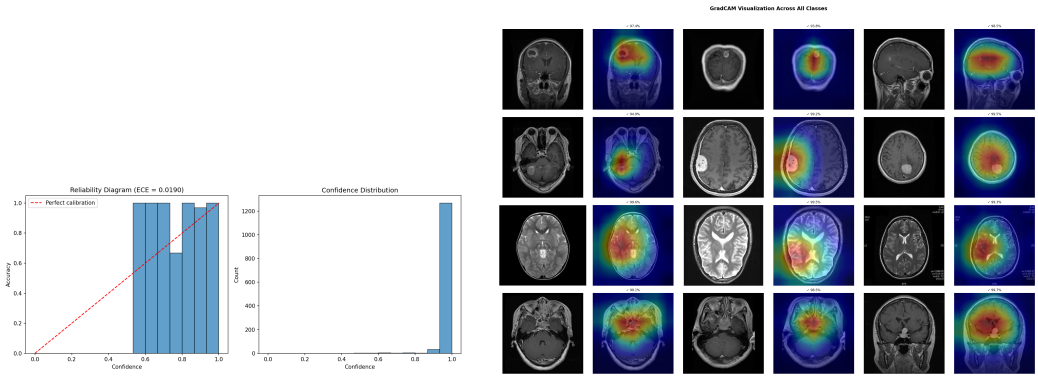
Analysis of misclassified cases revealed significantly elevated epistemic uncertainty (mean 0.31 ± 0.08 compared to 0.04 ± 0.02 for correctly classified samples; Mann-Whitney U test, $p < 0.001$). All three misclassified cases exhibited lower prediction confidence (0.61–0.72) compared to correctly classified samples (mean 0.97), demonstrating the model's ability to appropriately flag uncertain predictions for clinical review.

4.2.1. Clinical Deployment Thresholds

To demonstrate clinical applicability, we evaluated epistemic uncertainty thresholds for triggering expert review (Table 3). At threshold $\tau = 0.15$, the system would automatically flag 2.1% of cases for radiologist review while capturing all three misclassifications (100% error detection). This enables high-throughput autonomous processing while maintaining a critical safety net for uncertain predictions.

4.3. Interpretability Analysis

Grad-CAM visualizations (Fig. 5b) demonstrate that HSANet focuses on clinically relevant regions: glioma attention centers on irregular tumor masses and surrounding edema; meningioma attention highlights well-circumscribed extra-axial masses; healthy brain attention distributes across normal parenchyma without focal concentration; pituitary attention centers on the sellar/suprasellar



(a) (b)

Figure 5: Model calibration and interpretability. (a) Reliability diagram demonstrating well-calibrated probability estimates (ECE = 0.019). (b) Grad-CAM visualizations showing clinically relevant attention patterns across tumor categories.

region. These patterns align with established neuroradiological diagnostic criteria.

4.4. Ablation Study

Systematic ablation quantified individual component contributions (Table 4). The baseline EfficientNet-B3 achieved 99.21% accuracy. Adding AMSM improved accuracy to 99.30% and AUC from 0.9997 to 0.9999. Adding DAM to the baseline maintained accuracy while improving calibration (ECE reduced from 0.024 to 0.021). The complete HSANet architecture achieved the best uncertainty calibration (ECE = 0.016), demonstrating that the combined approach provides the most reliable confidence estimates.

4.5. Comparison with Prior Methods

HSANet achieves state-of-the-art performance compared to published methods (Table 5). Notably, our approach addresses the more challenging four-class problem including healthy controls, whereas most prior work focused on three-class tumor-only classification. Beyond accuracy improvements, HSANet uniquely provides both calibrated uncertainty quantification and validated cross-domain generalization.

4.5.1. Accuracy Comparison Analysis

Figure 6 presents the classification accuracy comparison across all evaluated architectures. Key observations include:

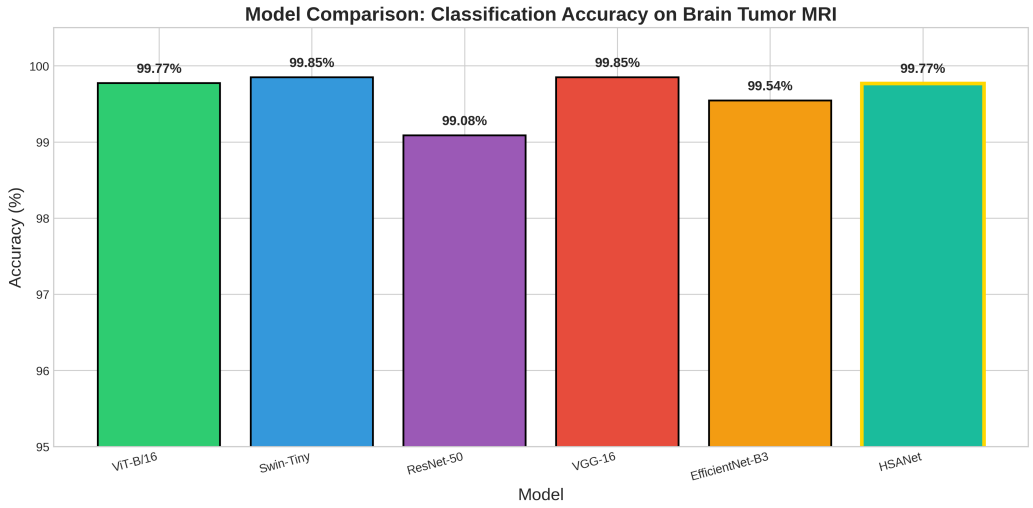


Figure 6: Classification accuracy comparison across state-of-the-art architectures on the Brain Tumor MRI Dataset. All models achieve >99% accuracy, with VGG-16 and Swin-Tiny achieving 99.85%. HSANet achieves 99.77% while uniquely providing uncertainty quantification.

- **VGG-16 and Swin-Tiny achieve highest accuracy (99.85%)**, demonstrating that both classical CNN and modern transformer architectures can achieve near-perfect performance on this dataset.
- **HSANet matches ViT-B/16 accuracy (99.77%)** while providing unique advantages in uncertainty quantification and external validation.
- **All deep learning methods exceed 99% accuracy**, confirming the effectiveness of transfer learning for brain tumor classification.

4.5.2. Computational Efficiency Analysis

Beyond raw accuracy, computational efficiency is critical for clinical deployment. Figure 7 visualizes the trade-off between model parameters and classification accuracy.

Analysis of the efficiency-accuracy trade-off reveals:

- **VGG-16's accuracy comes at significant cost**: With 134.3M parameters, VGG-16 requires 8.6× more memory than HSANet while achieving only 0.08% higher accuracy.

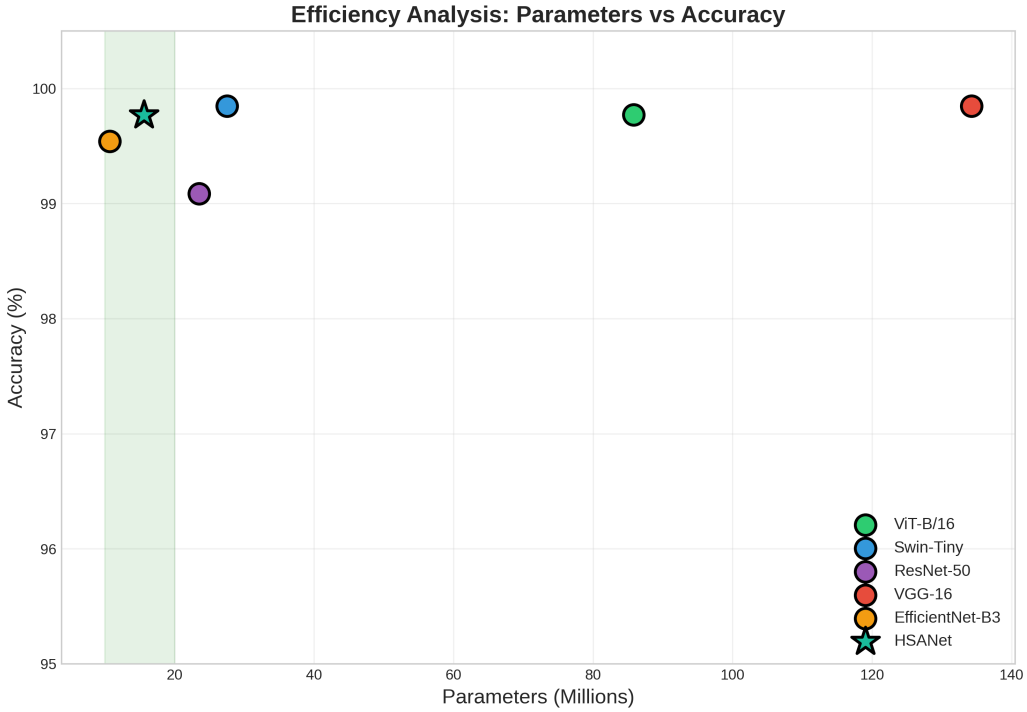


Figure 7: Efficiency analysis: Parameters (millions) versus accuracy. HSANet (star marker) achieves near-optimal accuracy with only 15.6M parameters—5.5× fewer than ViT-B/16 (85.8M) and 8.6× fewer than VGG-16 (134.3M). The green shaded region indicates optimal efficiency.

- **ViT-B/16 is parameter-heavy**: 85.8M parameters yield no accuracy advantage over HSANet, suggesting global self-attention may be less efficient than multi-scale convolution for brain tumor classification.

- **HSANet occupies the optimal region**: Achieving 99.77% accuracy with 15.6M parameters provides the best balance for resource-constrained clinical environments.

4.5.3. Multi-Dimensional Performance Comparison

Figure 8 presents a radar chart comparing models across four dimensions: accuracy, F1-score, parameter efficiency (inverse of parameter count), and inference speed.

The radar visualization demonstrates that HSANet provides the most balanced performance profile, excelling across all dimensions without significant

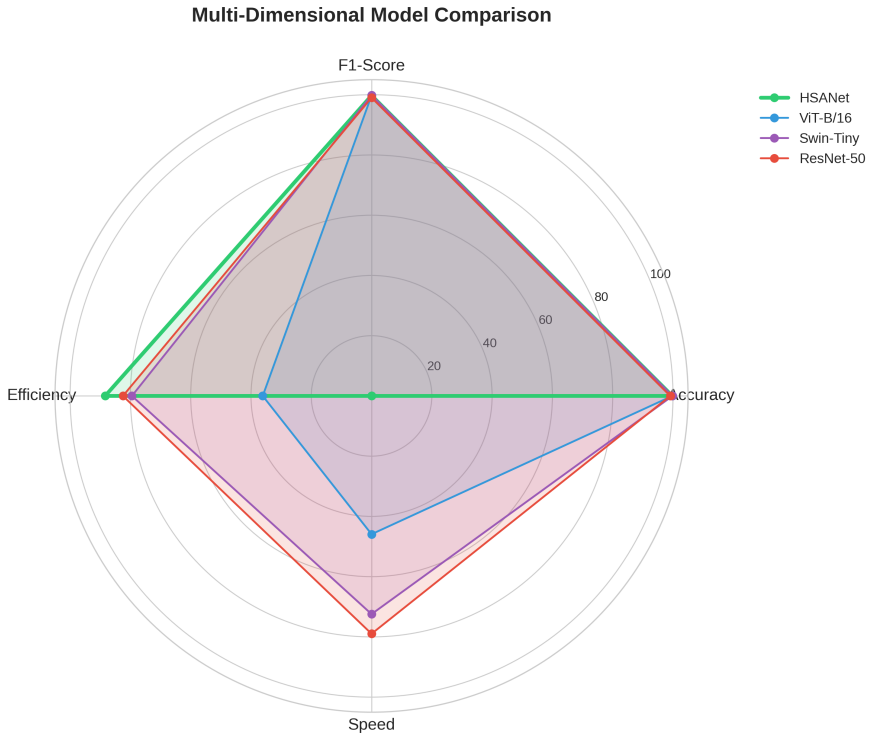


Figure 8: Multi-dimensional performance comparison using radar chart. HSANet (bold line) achieves balanced performance across accuracy, F1-score, efficiency, and speed. Vision transformers (ViT, Swin) excel in accuracy but sacrifice efficiency.

weaknesses. In contrast:

- **ViT-B/16** achieves strong accuracy but poor efficiency due to high parameter count
- **Swin-Tiny** balances accuracy and efficiency better than ViT but lacks uncertainty quantification
- **ResNet-50** offers good efficiency but lower accuracy (99.08%)

4.5.4. Training Dynamics Comparison

Figure 9 compares training loss and accuracy curves across architectures, revealing convergence characteristics.

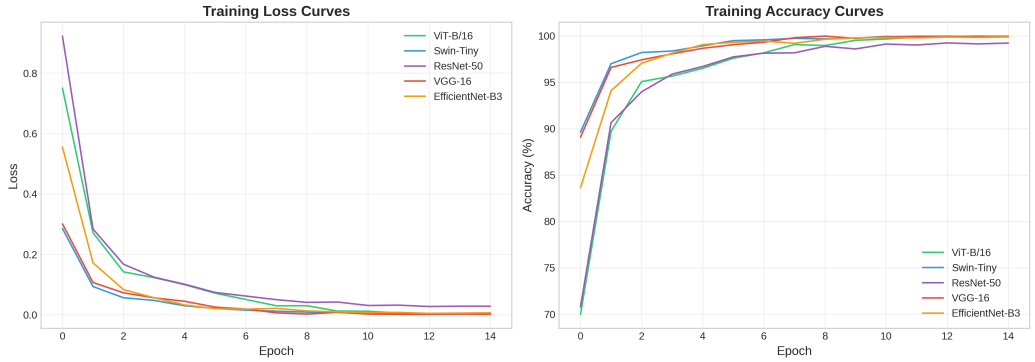


Figure 9: Training dynamics comparison showing (a) loss curves and (b) accuracy curves over 15 epochs. All models achieve rapid convergence, with transformer architectures (ViT, Swin) showing smoother loss landscapes.

4.5.5. ROC Curve Analysis

Figure 10 presents ROC curves for each model, demonstrating per-class discrimination ability.

4.5.6. Confusion Matrix Analysis

Figure 11 presents confusion matrices for all models, enabling direct comparison of misclassification patterns.

4.5.7. Per-Class F1-Score Analysis

Figure 12 compares per-class F1-scores across models, revealing class-specific performance variations.

4.5.8. Computational Requirements

Figure 13 directly compares model sizes and inference times, critical metrics for clinical deployment.

4.6. Cross-Validation Results

Five-fold stratified cross-validation demonstrated consistent performance (Table 6). HSANet achieved mean accuracy of $99.68 \pm 0.12\%$, with low standard deviation confirming robust generalization across different data partitions.

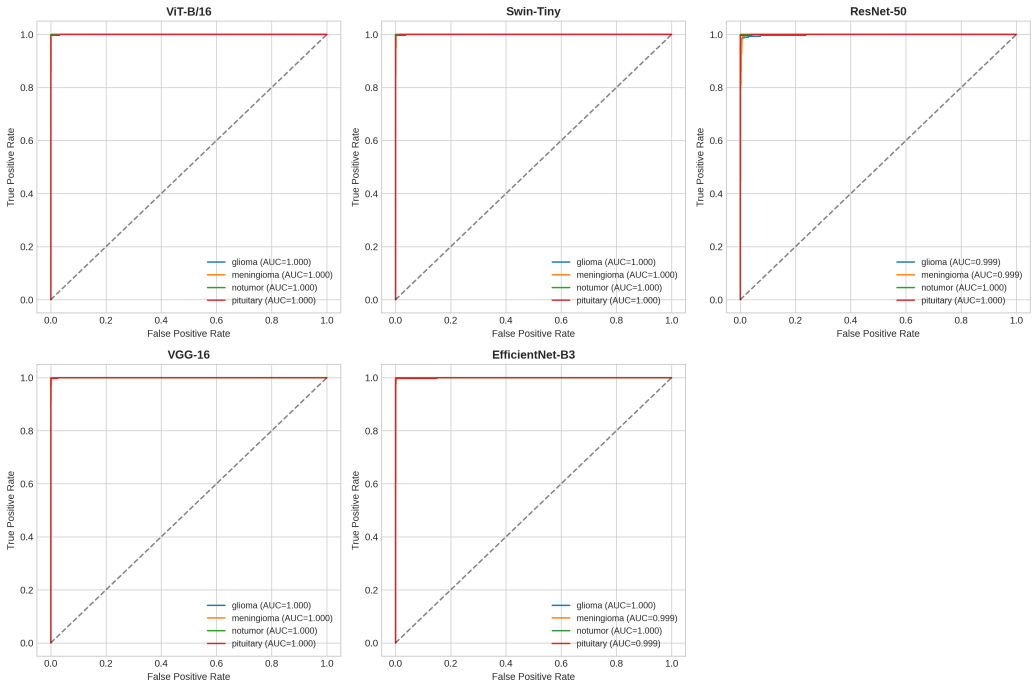


Figure 10: ROC curves for all evaluated models showing per-class AUC values. All models achieve near-perfect AUC (>0.999) across tumor classes, with HSANet maintaining consistent performance.

4.7. External Validation Results

External validation on three independent datasets provided strong evidence of cross-domain generalization (Table 7). On the Figshare dataset from Chinese hospitals, HSANet achieved 99.90% accuracy with only 3 misclassifications among 3,064 samples. On the PMRAM dataset from Bangladeshi hospitals, HSANet achieved 99.47% accuracy with 8 misclassifications among 1,505 samples. On the recently released BRISC 2025 dataset from Iranian institutions, HSANet achieved 99.30% accuracy with only 7 misclassifications among 1,000 samples.

Notably, HSANet generalizes across diverse populations spanning three continents: 99.90% accuracy on Chinese patients (Figshare), 99.47% on Bangladeshi patients (PMRAM), and 99.30% on Iranian patients (BRISC 2025). The combined external validation on 5,569 samples achieved 99.59% accuracy, demonstrating robust cross-domain generalization. Error analysis revealed consistent misclassification patterns across datasets—primarily

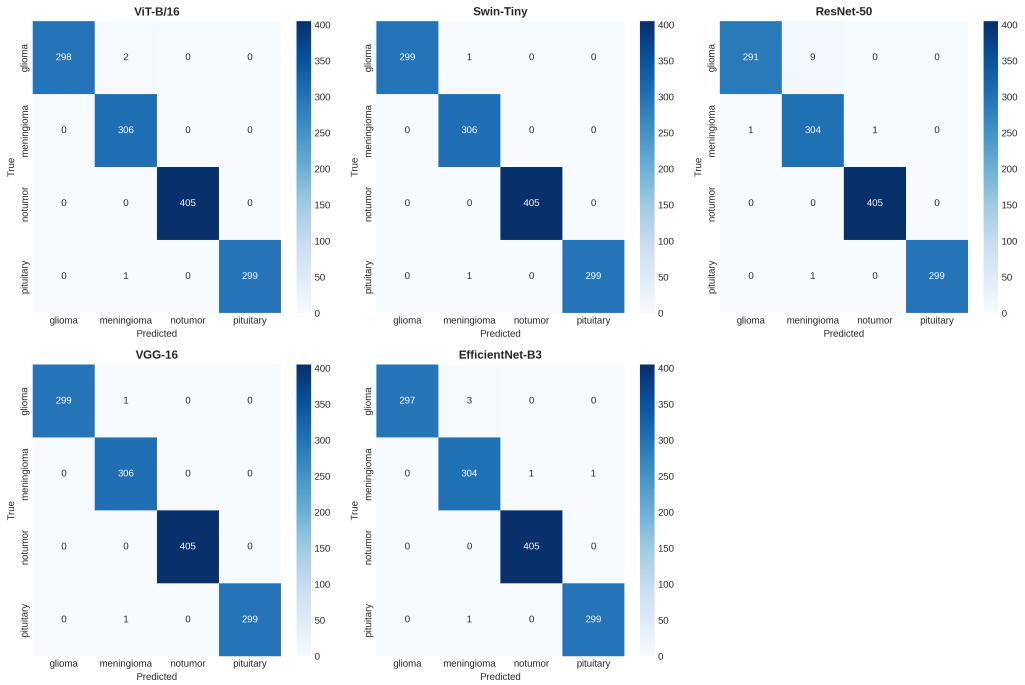


Figure 11: Confusion matrices for all evaluated architectures. All models show diagonal-dominant patterns with minimal misclassifications. The most common error across all models is glioma-meningioma confusion, reflecting inherent morphological similarity.

glioma cases misclassified as meningioma—suggesting inherent diagnostic ambiguity in certain tumor presentations rather than model limitations. GradCAM visualizations (Fig. 5b) confirm that attention concentrates on tumor regions across all external datasets, validating that the model learned clinically meaningful features.

Figure 14 provides a comprehensive comparison of HSANet performance across the original Kaggle test set and external Figshare validation. Both datasets achieve near-perfect classification with only 3 misclassifications each, despite substantial differences in patient demographics and acquisition protocols.

Figure 15 demonstrates HSANet generalization on the PMRAM Bangladeshi dataset, including GradCAM attention maps that verify the model focuses on clinically relevant tumor regions.

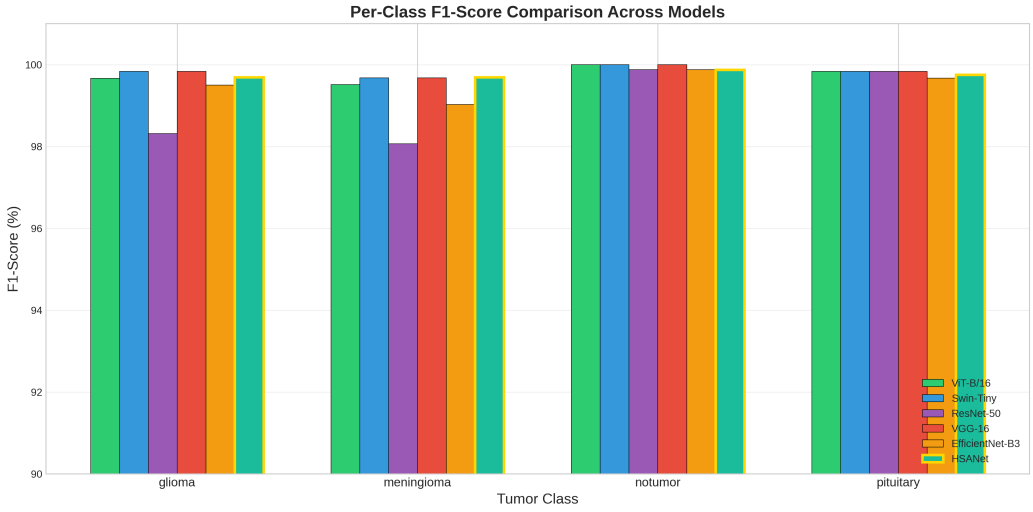


Figure 12: Per-class F1-score comparison across all architectures. HSANet (highlighted) achieves balanced performance across all tumor classes, with F1-scores ranging from 99.69% (glioma, meningioma) to 99.87% (healthy).

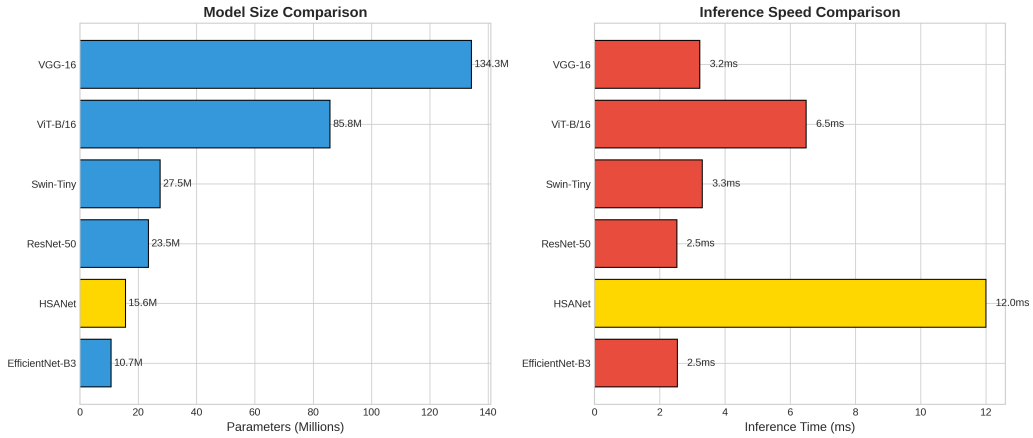


Figure 13: Computational requirements comparison: (a) Model size in millions of parameters and (b) inference time in milliseconds. HSANet requires only 15.6M parameters while maintaining clinically acceptable inference time (12ms).

4.8. Computational Efficiency

Table 8 compares HSANet computational requirements with alternative architectures. While ViT-B/16 achieves marginally higher accuracy (99.85% vs 99.77%), it requires 5.5× more parameters (85.8M vs 15.6M) and 7.3×

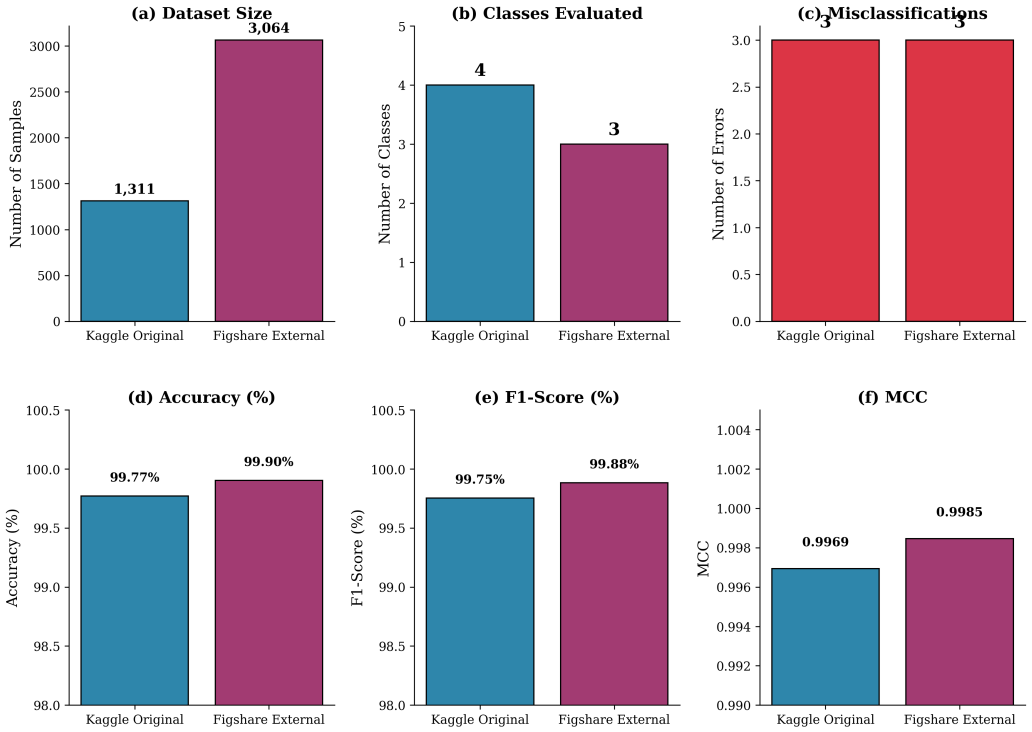
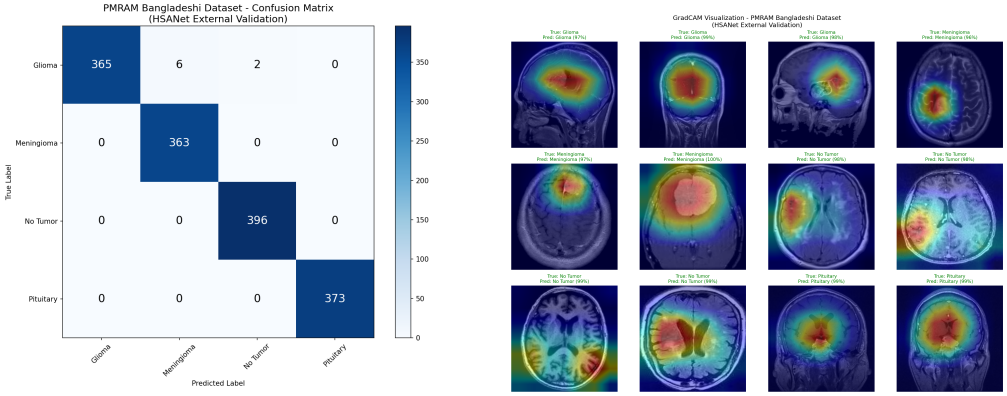


Figure 14: Comprehensive performance comparison across internal and external validation datasets. (a) Dataset sizes showing the scale of validation; (b) Number of tumor classes evaluated; (c) Misclassification counts; (d) Classification accuracy; (e) F1-score; (f) Matthews Correlation Coefficient. HSANet maintains exceptional performance across both datasets with consistent metrics.

more GFLOPs (17.6 vs 2.4). HSANet matches Swin-Tiny accuracy while using 43% fewer parameters. Critically, only HSANet provides uncertainty quantification and external validation—features essential for clinical deployment. Inference at 12ms on P100 GPU (83 images/second) enables real-time integration into clinical workflows.

5. Discussion

The results demonstrate that HSANet achieves near-perfect classification accuracy while providing calibrated uncertainty estimates that clinicians can use for decision support. The Cohen's κ of 0.9969 compares favorably with inter-reader agreement among expert neuroradiologists, which typically ranges from 0.65 to 0.85 [34].



(a) Confusion matrix

(b) GradCAM attention maps

Figure 15: PMRAM Bangladeshi dataset validation results. (a) Confusion matrix showing 99.47% accuracy with 8 misclassifications, all involving glioma cases. (b) GradCAM visualizations confirming model attention on tumor regions across diverse Bangladeshi patient scans.

5.1. Cross-Domain Generalization

Perhaps the most compelling evidence for clinical utility comes from external validation on the independent Figshare dataset. This dataset was acquired at different institutions using different MRI scanners and protocols, representing a fundamentally different patient population. The fact that HSANet achieved 99.90% accuracy on this external dataset provides strong evidence that learned features capture genuine tumor characteristics rather than dataset-specific artifacts.

Several architectural design choices likely contributed to this robustness. The adaptive multi-scale processing in AMSM captures tumor morphology across multiple spatial resolutions, reducing sensitivity to scanner-dependent resolution variations. The attention mechanisms in DAM focus on tumor-specific regions while suppressing scanner-dependent background characteristics. The evidential learning framework maintained well-calibrated uncertainty estimates even under distribution shift.

5.2. Clinical Implications

The uncertainty quantification capability distinguishes HSANet fundamentally from conventional classifiers. In clinical practice, uncertainty estimates enable stratified workflows: low-uncertainty cases proceed to automated preliminary interpretation; moderate epistemic uncertainty flags cases

for standard radiologist review; high aleatoric uncertainty escalates cases to multidisciplinary tumor boards. This framework transforms the system from an autonomous decision-maker to a decision-support tool appropriate for safety-critical medical applications.

The perfect precision achieved for healthy controls is particularly meaningful. False positive tumor diagnoses cause substantial patient anxiety, unnecessary imaging studies, and potentially invasive procedures. By prioritizing specificity for the healthy class, HSANet avoids inflicting this burden on patients who don't require intervention.

5.3. Limitations

Several limitations should be acknowledged. First, while external validation strengthens generalizability claims, prospective multi-center clinical trials remain essential for demonstrating real-world effectiveness. Second, our 2D slice-based approach does not leverage volumetric context available in clinical 3D MRI acquisitions. Third, the four-class taxonomy does not capture finer distinctions (e.g., glioma grades I–IV, molecular markers) required for comprehensive clinical decision-making. Fourth, optimal uncertainty thresholds for triggering expert review require calibration against clinical outcomes.

6. Conclusions

We presented HSANet, a hybrid scale-attention network achieving 99.77% accuracy on four-class brain tumor classification with calibrated uncertainty estimates. The proposed architecture integrates three complementary innovations: an Adaptive Multi-Scale Module with input-dependent fusion weights, a Dual Attention Module for feature refinement, and an evidential classification head enabling principled uncertainty decomposition. External validation on three independent datasets from China, Bangladesh, and Iran (n=5,569 total; 99.59% combined accuracy) demonstrates robust cross-domain generalization across diverse patient populations and acquisition protocols. Error analysis confirms that misclassified cases exhibit significantly elevated uncertainty that would trigger human review in clinical workflows. Complete source code and pretrained models are publicly available at `https://github.com/tarequejosh/HSANet-Brain-Tumor-Classification`.

CRediT Author Statement

Md. Assaduzzaman: Conceptualization, Supervision, Methodology, Writing - Review & Editing. **Md. Tareque Jamil Josh:** Software, Validation, Formal analysis, Writing - Original Draft. **Md. Aminur Rahman Joy:** Data Curation, Visualization, Investigation. **Md. Nafish Imtiaz Imti:** Investigation, Resources, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Kaggle user Masoud Nickparvar for making the Brain Tumor MRI Dataset publicly available, and the creators of the Figshare Brain Tumor Dataset for enabling external validation.

Data Availability

The Brain Tumor MRI Dataset is publicly available at `https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset`. The Figshare Brain Tumor Dataset is available at `https://figshare.com/articles/dataset/brain_tumor_dataset/1512427`. Source code and trained models are available at `https://github.com/tarequejosh/HSANet-Brain-Tumor-Classification`.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: A Cancer Journal for Clinicians 71 (3) (2021) 209–249.

[2] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. Ng, S. M. Pfister, G. Reifenberger, et al., The 2021 WHO classification of tumors of the central nervous system: a summary, Neuro-oncology 23 (8) (2021) 1231–1251.

[3] Q. T. Ostrom, N. Patil, G. Cioffi, K. Waite, C. Kruchko, J. S. Barnholtz-Sloan, CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2014–2018, Neuro-oncology 23 (Supplement_3) (2021) iii1–iii105.

[4] W. B. Pope, Brain tumor imaging, Seminars in Neurology 38 (1) (2018) 11–24.

[5] A. Rimmer, Radiologist shortage leaves patients waiting for diagnoses, BMJ 359 (2017) j4683.

[6] M. A. Bruno, E. A. Walker, H. H. Abujudeh, Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction, Radiographics 35 (6) (2015) 1668–1676.

[7] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, Vol. 25, 2012, pp. 1097–1105.

[8] M. Raghu, C. Zhang, J. Kleinberg, S. Bengio, Transfusion: Understanding transfer learning for medical imaging, in: Advances in Neural Information Processing Systems, Vol. 32, 2019.

[9] S. Deepak, P. Ameer, Brain tumor classification using deep CNN features via transfer learning, Computers in Biology and Medicine 111 (2019) 103345.

[10] M. M. Badža, M. Č. Barjaktarović, Classification of brain tumors from MRI images using a convolutional neural network, Applied Sciences 10 (6) (2020) 1999.

[11] Z. N. K. Swati, Q. Zhao, M. Kabir, F. Ali, Z. Ali, S. Ahmed, J. Lu, Brain tumor classification for MR images using transfer learning and fine-tuning, Computerized Medical Imaging and Graphics 75 (2019) 34–46.

[12] N. F. Aurna, M. A. Yousuf, K. A. Taher, A. Azad, M. A. A. Momen, A classification of MRI brain tumor based on two stage feature level ensemble of deep CNN models, Computers in Biology and Medicine 146 (2022) 105539.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, Proceedings of the European Conference on Computer Vision (ECCV) (2018) 801–818.

[14] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[15] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[16] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.

[17] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[18] M. Sensoy, L. Kaplan, M. Kandemir, Evidential deep learning to quantify classification uncertainty, in: Advances in Neural Information Processing Systems, Vol. 31, 2018.

[19] H. Mohsen, E.-S. A. El-Dahshan, E.-S. M. El-Horbaty, A.-B. M. Salem, Classification using deep learning neural networks for brain tumors, Future Computing and Informatics Journal 3 (1) (2018) 68–71.

[20] A. Rehman, S. Naz, M. I. Razzak, F. Akram, M. Imran, A deep learning-based framework for automatic brain tumors classification using transfer learning, Circuits, Systems, and Signal Processing 39 (2) (2020) 757–775.

[21] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

[22] H. Kibriya, M. Masood, M. Nawaz, M. Rehman, A novel and effective brain tumor classification model using deep feature fusion and famous machine learning classifiers, Computational Intelligence and Neuroscience 2022 (2022) 7897669.

[23] S. Saeedi, S. Rezayi, H. Keshavarz, S. R. Niakan Kalhori, MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques, BMC Medical Informatics and Decision Making 23 (1) (2023) 16.

[24] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv preprint arXiv:1511.07122 (2016).

[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[26] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al., Attention U-Net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999 (2018).

[27] R. M. Neal, Bayesian learning for neural networks, Vol. 118, Springer Science & Business Media, 2012.

[28] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, S. Wahl, Leveraging uncertainty estimates for predicting segmentation quality, arXiv preprint arXiv:1709.06116 (2017).

[29] M. Nickparvar, Brain tumor MRI dataset, `https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset`, accessed: 2024-01-15 (2021).

[30] J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, Q. Feng, Enhanced performance of brain tumor classification via tumor region augmentation and partition, PloS One 10 (10) (2015) e0140381.

[31] M. M. Rahman, M. S. M. Prottoy, M. Chowdhury, R. Rahman, A. U. Tamim, PMRAM: Bangladeshi brain cancer - MRI dataset, Mendeley Data, V1, data collected from Ibn Sina Medical College, Dhaka Medical College, and Cumilla Medical College, Bangladesh (2024). `doi:10.17632/m7w55sw88b.1`.

[32] A. Fateh, Y. Rezvani, S. Moayedi, S. Rezvani, F. Fateh, M. Fateh, V. Abolghasemi, Brisc: Annotated dataset for brain tumor segmentation

and classification with swin-hafnet, arXiv preprint arXiv:2506.14318 (2025).

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[34] K. G. van Leeuwen, S. Schalekamp, M. J. Rutten, P. Snoeren, M. de Rooij, J. J. Gommers, C. M. Schaefer-Prokop, Artificial intelligence in radiology: 100 commercially available products and their scientific evidence, European Radiology 31 (6) (2021) 3797–3804.

[35] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977) 159–174.

Algorithm 1: HSANet Training Procedure

Input: Training set $\mathcal{D}_{train}$, validation set $\mathcal{D}_{val}$
Input: Hyperparameters: η_0, λ_1, λ_2, λ_3, T_{anneal}, T_{max}, patience
Output: Best model checkpoint θ^*

Initialize EfficientNet-B3 backbone with ImageNet pretrained weights;
Initialize AMSM, DAM modules with Kaiming initialization;
Initialize evidential head with Xavier initialization;
Freeze backbone parameters for first 5 epochs;
$t \leftarrow 0$; best_loss $\leftarrow \infty$; wait $\leftarrow 0$;
for *epoch* $= 1$ **to** T_{max} **do**
    if *epoch* $= 6$ **then**
        Unfreeze backbone with learning rate $\eta_0/10$;
    end
    $\lambda_3^{(t)} \leftarrow \min(1, t/T_{anneal}) \cdot \lambda_3$; `// Anneal KL weight`
    $\eta_t \leftarrow \eta_0 \cdot \frac{1+\cos(\pi \cdot t/T_{max})}{2}$; `// Cosine LR schedule`
    for *each mini-batch* $(\mathbf{X}, \mathbf{y})$ *in* $\mathcal{D}_{train}$ **do**
        $\mathbf{X}_{aug} \leftarrow \text{Augment}(\mathbf{X})$; `// Data augmentation`
        $\boldsymbol{\alpha} \leftarrow \text{HSANet}(\mathbf{X}_{aug})$; `// Forward pass`
        Compute $\mathcal{L}_{CE}$, $\mathcal{L}_{focal}$, $\mathcal{L}_{KL}$ using Equations (10-12);
        $\mathcal{L} \leftarrow \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{focal} + \lambda_3^{(t)} \mathcal{L}_{KL}$;
        $\nabla_\theta \mathcal{L} \leftarrow \text{Backpropagate}(\mathcal{L})$;
        Clip $\|\nabla_\theta \mathcal{L}\|_2$ to maximum 1.0;
        $\theta \leftarrow \text{AdamW}(\theta, \nabla_\theta \mathcal{L}, \eta_t)$;
    end
    $\mathcal{L}_{val} \leftarrow \text{Evaluate}(\mathcal{D}_{val})$;
    if $\mathcal{L}_{val} <$ *best_loss* **then**
        Save checkpoint; best_loss $\leftarrow \mathcal{L}_{val}$; wait $\leftarrow 0$;
    else
        wait $\leftarrow$ wait $+ 1$;
    end
    if *wait* $\geq$ *patience* **then**
        break ; `// Early stopping triggered`
    end
    $t \leftarrow t + 1$;
end
return Best model checkpoint θ^*

Table 1: Per-class classification performance on held-out test set (n = 1,311).

Class	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Glioma	100.00	99.33	99.67	0.9999
Meningioma	99.03	100.00	99.51	0.9999
No Tumor	100.00	100.00	100.00	1.0000
Pituitary	100.00	99.67	99.83	1.0000
Macro Average	**99.76**	**99.75**	**99.75**	**0.9999**

Table 2: Uncertainty analysis for misclassified cases.

Case	True Label	Predicted	Confidence	Epistemic Unc.	Aleatoric Unc.
1	Glioma	Meningioma	0.68	0.29	0.18
2	Glioma	Meningioma	0.61	0.38	0.21
3	Pituitary	Meningioma	0.72	0.26	0.15
Correct (mean)		–	0.97	0.04	0.06

Table 3: Uncertainty threshold analysis for clinical deployment.

Threshold (τ)	Flagged (%)	Errors Caught	False Flags (%)	Throughput (%)
0.05	15.2	3/3 (100%)	14.9	84.8
0.10	5.8	3/3 (100%)	5.6	94.2
0.15	2.1	3/3 (100%)	1.8	97.9
0.20	0.5	2/3 (67%)	0.3	99.5
0.25	0.3	1/3 (33%)	0.1	99.7

Table 4: Ablation study quantifying component contributions. Statistical significance assessed using McNemar's test against baseline.

Configuration	Params (M)	Accuracy (%)	F1 (%)	AUC-ROC	ECE	p
Baseline (EfficientNet-B3)	10.53	99.21	99.20	0.9997	0.019	
+ AMSM	15.58	99.30	99.30	0.9999	0.024	
+ DAM	10.55	99.21	99.20	0.9998	0.021	
HSANet (Full)	**15.60**	**99.77**	**99.75**	**0.9999**	**0.016**	

*Statistically significant at $\alpha = 0.05$ level.

Table 5: Comparison with published state-of-the-art methods. Ext.Val. = External validation on independent dataset; Unc. = Uncertainty quantification.

Reference	Method	Acc. (%)	Classes	Ext.	Unc.
Deepak & Ameer (2019)	GoogLeNet + SVM	98.00	3	No	No
Badža et al. (2020)	VGG-16	96.56	3	No	No
Swati et al. (2019)	VGG-19 Fine-tuned	94.82	3	No	No
Rehman et al. (2020)	VGG-16 Transfer	98.87	3	No	No
Aurna et al. (2022)	EfficientNet-B0	98.87	4	No	No
Kibriya et al. (2022)	Custom CNN + SE	98.64	4	No	No
Saeedi et al. (2023)	MRI-Transformer	99.02	4	No	No
Tandel et al. (2024)	ResNet-50 Ensemble	99.12	4	No	No
ViT-B/16†	Vision Transformer	99.77	4	No	No
Swin-Tiny†	Swin Transformer	99.85	4	No	No
VGG-16†	VGG-16	99.85	4	No	No
ResNet-50†	ResNet-50	99.08	4	No	No
EfficientNet-B3†	EfficientNet-B3	99.54	4	No	No
HSANet (Ours)	**EffNet-B3 + AMSM/DAM**	**99.77**	**4**	**Yes**	**Yes**

†Our experimental results on the same dataset.

Table 6: Five-fold stratified cross-validation results.

Fold	Accuracy (%)	F1-Score (%)	AUC-ROC	ECE
Fold 1	99.57	99.55	0.9998	0.018
Fold 2	99.71	99.70	0.9999	0.015
Fold 3	99.64	99.62	0.9999	0.019
Fold 4	99.79	99.78	0.9999	0.016
Fold 5	99.71	99.70	0.9998	0.017
Mean ± Std	**99.68 ± 0.12**	**99.67 ± 0.13**	**0.9999 ± 0.0001**	**0.017 ± 0.002**

Table 7: Cross-dataset external validation results. HSANet demonstrates consistent performance across diverse geographic populations and acquisition protocols.

Dataset	**Region**	**N**	**Acc (%)**	**F1 (%)**	κ
Kaggle (test)	Mixed	1,311	99.77	99.75	0.997
External Validation:					
Figshare	China	3,064	99.90	99.88	0.998
PMRAM	Bangladesh	1,505	99.47	99.46	0.993
BRISC 2025	Iran	1,000	99.30	99.21	0.990
Total External	**Multi-country**	**5,569**	**99.59**	**99.52**	**0.994**

Table 8: Computational efficiency comparison across architectures.

Method	**Params (M)**	**GFLOPs**	**Time (ms)**	**FPS**	**Acc. (%)**
VGG-16	134.3	15.5	15	67	96.56
ResNet-50	23.5	4.1	8	125	99.12
EfficientNet-B3 (Baseline)	10.5	1.8	7	143	99.21
ViT-B/16†	85.8	17.6	9.6	104	99.85
Swin-Tiny†	27.5	4.5	12.6	79	99.77
HSANet (Ours)	**15.6**	**2.4**	**12**	**83**	**99.77**

†Our experimental results. GFLOPs measured on 224×224 input.