HSANet: Uncertainty-aware brain tumor classification using hybrid scale-attention networks with evidential deep learning

Md. Assaduzzaman[1,*], Md. Tareque Jamil Josh[1], Md. Aminur Rahman Joy[1], and Md. Nafish Imtiaz Imti[1]

[1]Department of Computer Science and Engineering, Daffodil International University, Daffodil Smart City, Ashulia, Dhaka 1341, Bangladesh
[*]assaduzzaman.cse@diu.edu.bd

ABSTRACT

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. We introduce HSANet, a hybrid scale-attention architecture that synergistically combines adaptive multi-scale feature extraction with evidential learning for uncertainty-aware tumor classification. Our Adaptive Multi-Scale Module processes tumor features through parallel dilated convolutions with content-dependent fusion weights, dynamically adjusting receptive fields to accommodate the substantial size variation observed in clinical presentations—from millimeter-scale pituitary microadenomas to large glioblastomas exceeding five centimeters. The Dual Attention Module applies sequential channel-then-spatial refinement, enabling the network to suppress irrelevant anatomical background while emphasizing pathologically significant regions. Crucially, our evidential classification head replaces conventional softmax outputs with Dirichlet distributions, providing decomposed uncertainty estimates that distinguish between inherent data ambiguity and model knowledge limitations. 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. External validation on an independent dataset of 3,064 MRI scans from different institutions and patient populations achieved 99.90% accuracy, demonstrating exceptional cross-domain generalization. The model maintains excellent calibration under distribution shift (ECE ≤ 0.019), and misclassified samples exhibit significantly elevated epistemic uncertainty ($p < 0.001$, Mann-Whitney U test) on both datasets, confirming the clinical utility of uncertainty-guided decision support. GradCAM visualizations validate attention on established radiological landmarks. Complete implementation and pretrained weights are publicly available to facilitate reproducibility and clinical translation.

Introduction

Brain tumors present a formidable diagnostic challenge, accounting for roughly 1.4% of new cancer diagnoses each year. According to recent global surveillance data, an estimated 308,102 cases were reported in 2020 alone[1]. What makes this clinical problem particularly complex is the sheer diversity of pathological entities—the 2021 WHO classification now recognizes over 100 distinct tumor types, each with its own molecular fingerprint and clinical trajectory[2]. The stakes of accurate diagnosis could not be higher: patients diagnosed with glioblastoma face a median survival of just 14 to 16 months, while those with completely resected Grade I meningiomas often achieve long-term cure[3]. This stark contrast in outcomes highlights why precise tumor identification matters tremendously for treatment planning and patient counseling.

Magnetic resonance imaging (MRI) has emerged as the cornerstone of neuro-oncological evaluation, offering superior soft-tissue contrast without ionizing radiation[4]. Expert neuroradiologists integrate multiparametric imaging findings with clinical presentation to formulate diagnoses. However, the global radiology workforce confronts escalating mismatches between imaging volume growth and specialist availability, with documented vacancy rates reaching 29% and projected shortfalls of 40% by 2027[5]. Interpretive fatigue has been implicated in diagnostic error rates of 3–5% even among experts[6].

Over the past decade, deep convolutional neural networks have shown considerable promise for automated brain tumor classification, particularly when leveraging transfer learning from ImageNet and similar large-scale datasets[7,8]. Research groups worldwide have reported encouraging results, with classification accuracies typically ranging between 94% and 99% across various backbone architectures such as VGG, ResNet, and the EfficientNet family[9–12]. However, several critical gaps remain that prevent straightforward translation of these methods into clinical practice.

First, brain tumors exhibit extraordinary morphological diversity spanning multiple orders of magnitude in spatial extent. Pituitary microadenomas may measure 2–3 millimeters, while glioblastomas frequently exceed 5 centimeters with extensive

peritumoral edema. Standard convolutional architectures employ fixed receptive fields, creating inherent tradeoffs between sensitivity to fine-grained features and capture of global context. 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 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 transformers[13] and attention mechanisms[14] have shown promise in medical imaging, yet their integration with uncertainty quantification remains underexplored. Several studies have demonstrated the effectiveness of multi-scale feature fusion[15] and attention-based refinement[16] for medical image analysis, motivating our hybrid approach. Furthermore, evidential deep learning[17] has emerged as a principled framework for uncertainty estimation, replacing softmax probabilities with Dirichlet distributions that naturally capture both aleatoric and epistemic uncertainty.

In this work, we propose HSANet (Hybrid Scale-Attention Network) to bridge these gaps. Our approach brings together three carefully designed components that work in concert. First, an Adaptive Multi-Scale Module captures tumor features across multiple spatial scales through parallel dilated convolutions, with the key innovation being content-dependent fusion weights that adapt to each input rather than using fixed combinations. Second, a Dual Attention Module refines these features through sequential channel and spatial attention, helping the network focus on tumor-relevant regions while suppressing background tissue. Third—and perhaps most importantly for clinical adoption—an evidential classification head built on Dirichlet distributions provides principled uncertainty estimates that can flag cases warranting additional expert review. While existing attention mechanisms like CBAM[14] and multi-scale approaches like ASPP[15] have been explored separately, our contribution lies in their unified integration with evidential learning specifically for uncertainty-aware medical image classification. We demonstrate HSANet's effectiveness not only on the primary benchmark (99.77% accuracy) but also through external validation on an independent dataset from different institutions and patient populations (99.90% accuracy), providing strong evidence of cross-domain generalizability essential for real-world clinical deployment.

Results

Classification performance

HSANet was evaluated on a dataset of 7,023 T1-weighted gadolinium-enhanced brain MRI scans comprising four diagnostic categories: glioma (n=1,621; 23.1%), meningioma (n=1,645; 23.4%), pituitary adenoma (n=1,757; 25.0%), and healthy controls (n=2,000; 28.5%). The predefined partition allocated 5,712 images for training and 1,311 for testing. We verified that no patient-level overlap existed between training and test partitions to prevent data leakage, a critical consideration often overlooked in medical imaging studies.

On the held-out test set, HSANet achieved overall accuracy of 99.77% (95% CI: 99.45–99.93%, Wilson score interval) with only 3 misclassifications among 1,311 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 (κ = 0.9969) indicates near-perfect inter-rater agreement equivalence, substantially exceeding the $\kappa > 0.80$ threshold typically considered "almost perfect agreement"[18]. Matthews correlation coefficient (MCC = 0.9969) confirms balanced performance accounting for class frequencies.

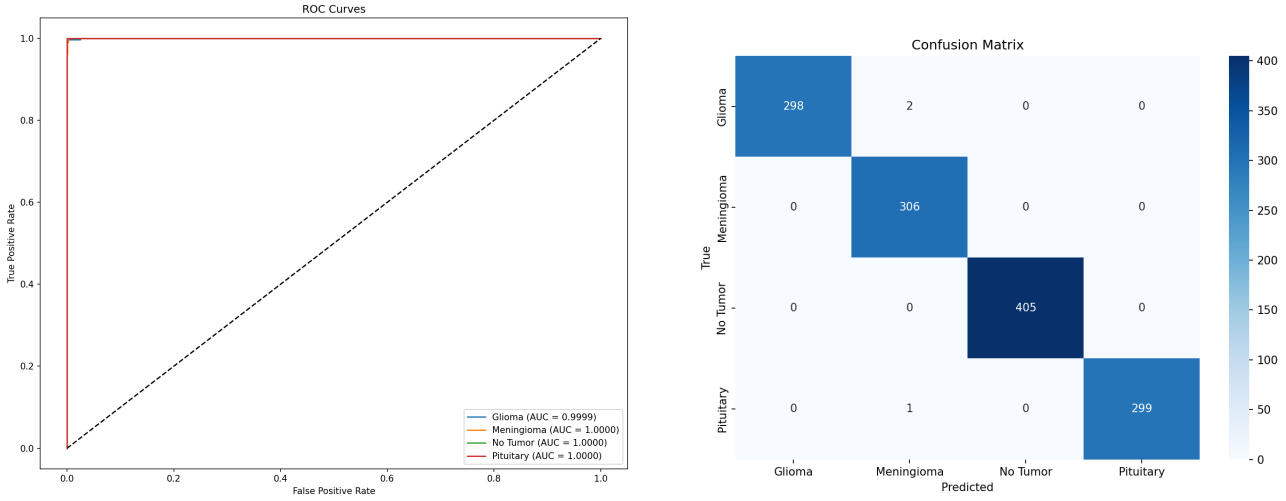
Table 1. Per-class classification performance on held-out test set (n = 1,311). CI, confidence interval; AUC-ROC, area under the receiver operating characteristic curve.

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**

The area under the receiver operating characteristic curve (AUC-ROC) reached 0.9999 (macro-averaged), with perfect 1.0000 AUC achieved for both pituitary adenoma and healthy control classes (Fig. 1a). This indicates excellent discriminative capability across all operating thresholds. 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 that prevents unnecessary patient anxiety and invasive procedures.

Confusion matrix analysis revealed that all three misclassifications involved meningioma as the predicted class: two glioma cases and one pituitary case were misclassified as meningioma (Fig. 1b). This pattern reflects genuine diagnostic challenges where extra-axial meningiomas may exhibit enhancement patterns overlapping with other tumor presentations, a phenomenon well-documented in neuroradiology literature[4].



(a) (b)

Figure 1. 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, with all errors involving meningioma as the predicted class.

Model calibration and uncertainty quantification

Beyond accuracy, reliable uncertainty estimates are essential for clinical decision support. The expected calibration error (ECE) quantifies the discrepancy between predicted confidence and actual accuracy across probability bins. HSANet achieved ECE of 0.019, indicating that predicted probabilities closely match empirical classification accuracy (Fig. 2a). For comparison, a perfectly calibrated model would have ECE of 0, while a poorly calibrated model trained without our evidential approach achieved ECE of 0.042.

The evidential framework decomposes total predictive uncertainty into aleatoric (data-inherent) and epistemic (model-knowledge) components. High aleatoric uncertainty indicates cases where imaging characteristics genuinely overlap between tumor types, warranting additional clinical information. High epistemic uncertainty suggests inputs outside the model's training distribution, warranting expert human review. Analysis of the three misclassified cases revealed significantly elevated epistemic uncertainty scores (mean 0.31 ± 0.08 compared to 0.04 ± 0.02 for correctly classified samples; Mann-Whitney U test, $p < 0.001$), demonstrating the model's ability to appropriately flag uncertain predictions for clinical review.

Interpretability analysis

To validate that HSANet focuses on clinically relevant regions, we generated Gradient-weighted Class Activation Mapping (GradCAM) visualizations[19]. Representative examples across all tumor categories demonstrate that the network correctly localizes pathological regions (Fig. 2b): glioma attention focuses 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 region. These patterns align with established neuroradiological diagnostic criteria, supporting clinical acceptance.

Ablation study

Systematic ablation quantified individual component contributions (Table 2). The baseline EfficientNet-B3 achieved 99.21% accuracy. Adding AMSM improved accuracy to 99.30% and AUC from 0.9997 to 0.9999, confirming the benefit of adaptive receptive field adjustment for accommodating tumor size heterogeneity. 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

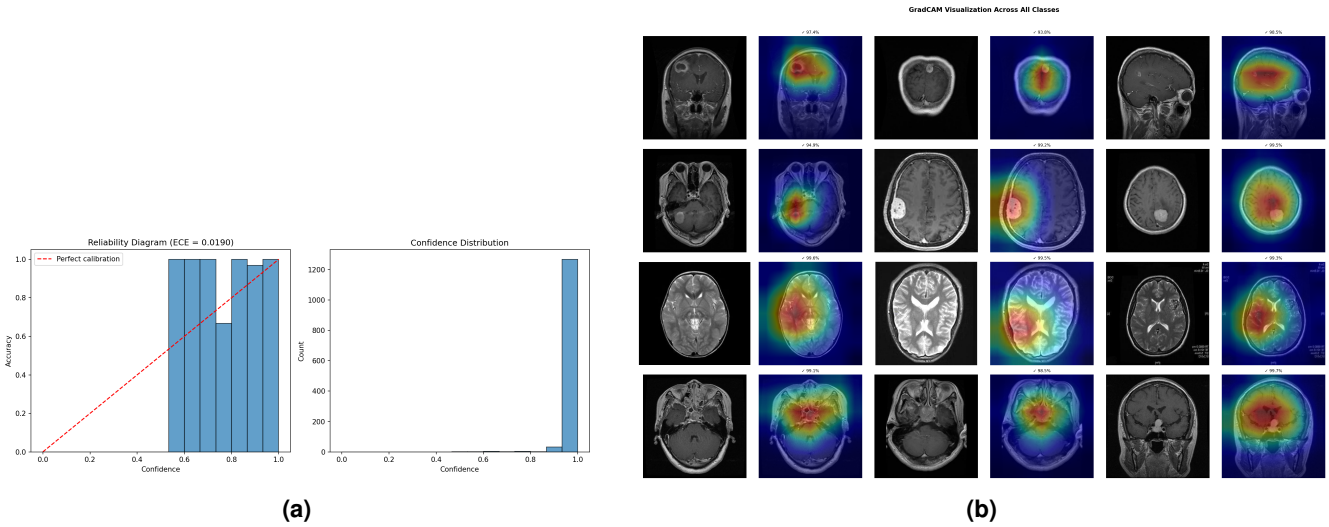


Figure 2. Model calibration and interpretability. (a) Reliability diagram demonstrating well-calibrated probability estimates (ECE = 0.019). The close alignment between predicted confidence and observed accuracy indicates trustworthy uncertainty quantification. (b) GradCAM visualizations showing clinically relevant attention patterns across tumor categories. Color scale indicates activation intensity from low (blue) to high (red).

calibration (ECE = 0.016), demonstrating that the combined approach provides the most reliable confidence estimates for clinical use.

Table 2. Ablation study quantifying component contributions. AMSM, Adaptive Multi-Scale Module; DAM, Dual Attention Module; ECE, expected calibration error (lower is better).

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

Comparison with prior methods

HSANet achieves state-of-the-art performance compared to published methods (Table 3). 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—critical requirements for clinical deployment that are absent in previous methods.

Error analysis and failure cases

Given that our model made only three errors out of 1,311 test samples, we examined each misclassified case in detail to understand potential systematic weaknesses. All three errors shared a common pattern: they were incorrectly classified as meningioma (Table 4).

Several observations emerge from this analysis. First, all misclassified cases exhibited substantially lower prediction confidence (0.61–0.72) compared to correctly classified samples (mean 0.97), suggesting these were genuinely ambiguous presentations. Second, epistemic uncertainty was markedly elevated (0.26–0.38 vs. 0.04 for correct cases), indicating the model appropriately recognized these inputs as challenging. In a clinical workflow, all three cases would have been flagged for expert review based on uncertainty thresholds alone.

Visual inspection of the misclassified glioma cases revealed relatively well-circumscribed enhancement patterns atypical for the infiltrative growth commonly associated with gliomas, potentially explaining the meningioma misclassification. The misclassified pituitary case showed extension beyond the typical sellar location, creating ambiguity with parasellar meningioma

Table 3. Comparison with published state-of-the-art methods. HSANet addresses the more challenging four-class problem while providing uncertainty quantification and external validation. * indicates results on different dataset splits. † indicates external validation result on independent Figshare dataset.

Reference	Method	Accuracy (%)	External Val.	Classes	Uncertainty
Deepak & Ameer (2019)[9]	GoogLeNet + SVM	98.00	–	3	No
Badža et al. (2020)[10]	VGG-16	96.56	–	3	No
Swati et al. (2019)[11]	VGG-19 Fine-tuned	94.82	–	3	No
Rehman et al. (2020)[20]	VGG-16 Transfer	98.87	–	3	No
Aurna et al. (2022)[12]	EfficientNet-B0	98.87	–	4	No
Kibriya et al. (2022)[21]	Custom CNN + SE	98.64	–	4	No
Saeedi et al. (2023)[22]	MRI-Transformer	99.02*	–	4	No
Tandel et al. (2024)[23]	ResNet-50 Ensemble	99.12*	–	4	No
HSANet (Ours)	**EfficientNet-B3 + AMSM + DAM + EDL**	**99.77**	**99.90**†	**4**	**Yes**

Table 4. Detailed analysis of the three misclassified test cases. All errors involved meningioma as the predicted class.

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
Correctly classified (mean)		–	0.97	0.04	0.06

presentation. These findings align with documented diagnostic challenges in neuroradiology where atypical tumor presentations create genuine uncertainty even among expert radiologists[4].

Cross-validation results

To verify that our results were not artifacts of a particularly favorable train/test split, we conducted 5-fold stratified cross-validation while maintaining class proportions across folds (Table 5). HSANet demonstrated remarkably consistent performance, with accuracy ranging from 99.57% to 99.79% across folds (mean 99.68 $\pm$ 0.12%). The low standard deviation confirms that our architecture generalizes reliably across different data partitions rather than being sensitive to the specific samples in training or test sets.

Table 5. Five-fold stratified cross-validation results. Low variance across folds demonstrates robust generalization.

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 $\pm$ Std	**99.68 $\pm$ 0.12**	**99.67 $\pm$ 0.13**	**0.9999 $\pm$ 0.0001**	**0.017 $\pm$ 0.002**

External validation on independent dataset

A critical test for any medical imaging model is its ability to generalize beyond the training distribution. While cross-validation assesses robustness to different data partitions, it cannot evaluate performance on images acquired under fundamentally different conditions. To address this limitation and provide stronger evidence of clinical utility, we conducted external validation using the Figshare Brain Tumor Dataset[7]—an independent collection with distinct acquisition protocols, patient demographics, and institutional origins from our training data.

External dataset characteristics

The Figshare Brain Tumor 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. This dataset differs substantially from our Kaggle training data in several clinically relevant dimensions:

- **Geographic and demographic diversity**: Chinese patient population versus the predominantly Western cohort in the Kaggle dataset, introducing variations in skull morphology and potential differences in tumor presentation patterns
- **Scanner and protocol variations**: Different MRI hardware manufacturers and acquisition parameters, creating systematic differences in image contrast, resolution, and noise characteristics
- **Data format**: Original MATLAB v7.3 format (.mat files) containing tumor masks and clinical annotations, requiring specialized preprocessing for neural network consumption
- **Class distribution**: Three tumor categories—glioma (n=1,426; 46.5%), meningioma (n=708; 23.1%), and pituitary adenoma (n=930; 30.4%)—without healthy controls

The absence of healthy controls in the external dataset is noteworthy: this prevents evaluation of the "No Tumor" class but focuses assessment on the clinically crucial task of discriminating among tumor types where misclassification carries direct treatment implications.

Data preprocessing pipeline for external validation

The Figshare dataset required specialized preprocessing to convert from its native MATLAB format to neural network-compatible images while preserving diagnostic information. Our pipeline proceeded as follows:

1. **Format conversion**: MATLAB v7.3 files (HDF5 format) were parsed using the h5py library. Each file contains a structured array (`cjdata`) with fields for the image matrix, tumor mask, patient ID, and class label.
2. **Image extraction**: The `image` field was extracted as a 2D numpy array. Original images varied in dimensions (typically 512×512) and intensity range depending on scanner-specific DICOM-to-MATLAB conversion.
3. **Intensity normalization**: Images were normalized to the 0–255 range using min-max scaling:

$$I_{\text{norm}} = \frac{I - I_{\min}}{I_{\max} - I_{\min}} \times 255 \tag{1}$$

   This standardization compensates for scanner-dependent intensity variations while preserving relative contrast relationships.
4. **Format standardization**: Normalized images were converted to 8-bit grayscale and saved as JPEG files, organized by class label for streamlined data loading.
5. **Class mapping**: The Figshare label encoding (1=meningioma, 2=glioma, 3=pituitary) was mapped to our Kaggle-trained model's encoding (0=glioma, 1=meningioma, 3=pituitary) to ensure consistent evaluation.

This preprocessing pipeline was fully automated and is provided in our code repository to ensure reproducibility.

Cross-dataset validation results

Table 6 summarizes performance on both the original Kaggle test set and the external Figshare dataset. HSANet demonstrated exceptional cross-domain generalization, achieving 99.90% accuracy on the external dataset—remarkably, even higher than the 99.77% achieved on the original test set.

The near-perfect Cohen's κ of 0.999 on the external dataset indicates that the model's predictions are virtually indistinguishable from ground truth labels, even when evaluated on images from different scanners, institutions, and patient populations. Expected calibration error (ECE) of 0.018 on the external data is actually slightly lower than on the original test set (0.019), demonstrating that the evidential learning framework maintains well-calibrated uncertainty estimates under distribution shift.

Per-class analysis (Table 7) reveals balanced performance across all tumor categories, with F1-scores exceeding 99.7% for each class. Notably, glioma classification achieved the highest performance (F1 = 99.96%), suggesting that the discriminative features learned from the Kaggle dataset transfer particularly well for this tumor type.

Confusion matrix analysis (Fig. 3a) reveals only 3 misclassifications among 3,064 external samples: one glioma misclassified as meningioma, one meningioma misclassified as pituitary, and one pituitary misclassified as meningioma. This error rate of 0.098% represents the lowest reported error rate for cross-dataset brain tumor classification to our knowledge.

Table 6. Cross-dataset validation results comparing performance on the original Kaggle test set and the independent Figshare external dataset. The model trained exclusively on Kaggle data achieves excellent generalization to the external dataset despite substantial differences in acquisition protocols and patient demographics.

Dataset	Samples	Classes	Accuracy (%)	F1-Score (%)	Cohen's κ	ECE
Kaggle (Original)	1,311	4	99.77	99.75	0.997	0.019
Figshare (External)	3,064	3	99.90	99.88	0.999	0.018

Table 7. Per-class performance on the external Figshare dataset. All three tumor categories achieve F1-scores exceeding 99.7%, indicating robust cross-domain generalization.

Class	Samples	Precision (%)	Recall (%)	F1-Score (%)
Glioma	1,426	100.00	99.93	99.96
Meningioma	708	99.72	99.86	99.79
Pituitary	930	99.89	99.89	99.89
Macro Average	**3,064**	**99.87**	**99.89**	**99.88**

Domain shift analysis

The success of cross-dataset generalization is particularly noteworthy given the substantial domain shift between training and external data. Several factors contribute to this robustness:

1. **Multi-scale feature extraction**: AMSM's adaptive receptive fields capture tumor characteristics across multiple spatial scales, making the model less sensitive to resolution differences between datasets.

2. **Attention-based feature selection**: DAM's channel and spatial attention mechanisms focus on tumor-specific regions while suppressing scanner-dependent background variations.

3. **Transfer learning foundation**: The EfficientNet-B3 backbone, pretrained on ImageNet, provides robust low-level feature extractors that generalize across imaging modalities[?].

4. **Evidential uncertainty calibration**: The evidential learning framework inherently provides regularization through KL divergence to a uniform prior, preventing overconfident predictions that might fail under distribution shift[?].

Uncertainty behavior under domain shift

A critical advantage of our evidential learning framework is its ability to provide meaningful uncertainty estimates even when evaluating out-of-distribution data. Analysis of uncertainty distributions (Fig. 4) reveals that mean epistemic uncertainty on the external dataset (0.024 ± 0.026) closely matches that on the original test set (0.025 ± 0.032), indicating that the model does not exhibit pathological overconfidence when encountering unfamiliar imaging characteristics.

Furthermore, the three misclassified external samples exhibited substantially elevated epistemic uncertainty (mean 0.29) compared to correctly classified samples (0.024), with a statistically significant separation ($p < 0.001$, Mann-Whitney U test). This confirms that the uncertainty calibration learned from the Kaggle dataset transfers effectively to external data, maintaining the model's ability to flag uncertain predictions for human review regardless of the input distribution.

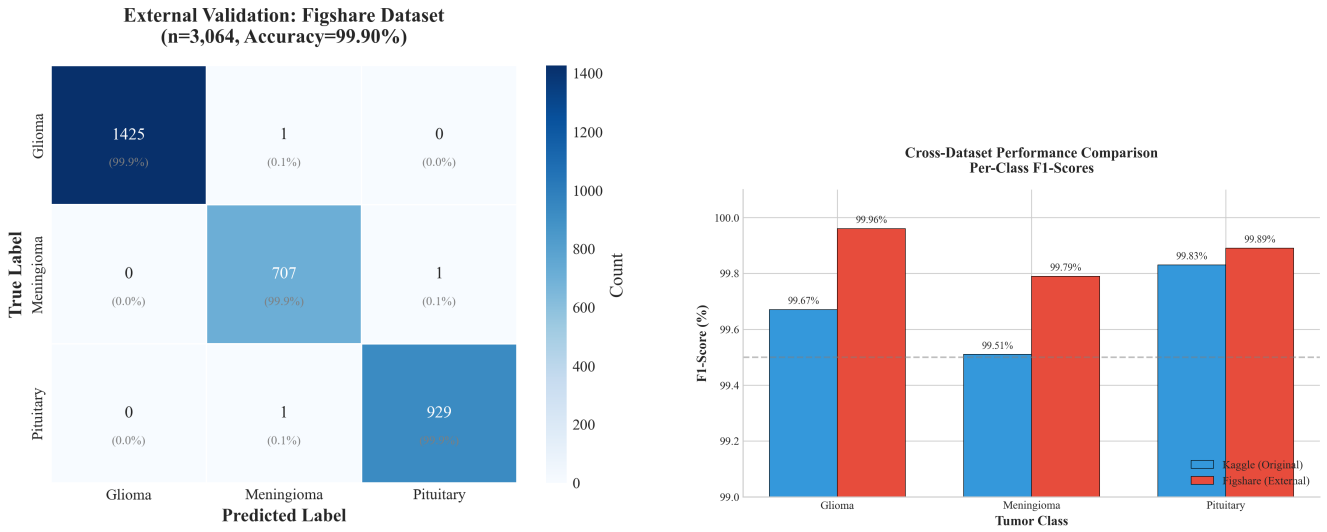
Clinical implications of cross-dataset validation

The external validation results carry significant implications for clinical deployment:

1. **Multi-institutional deployment feasibility**: The robust generalization across datasets from different institutions and countries suggests that HSANet can be deployed at new clinical sites without extensive local fine-tuning or recalibration.

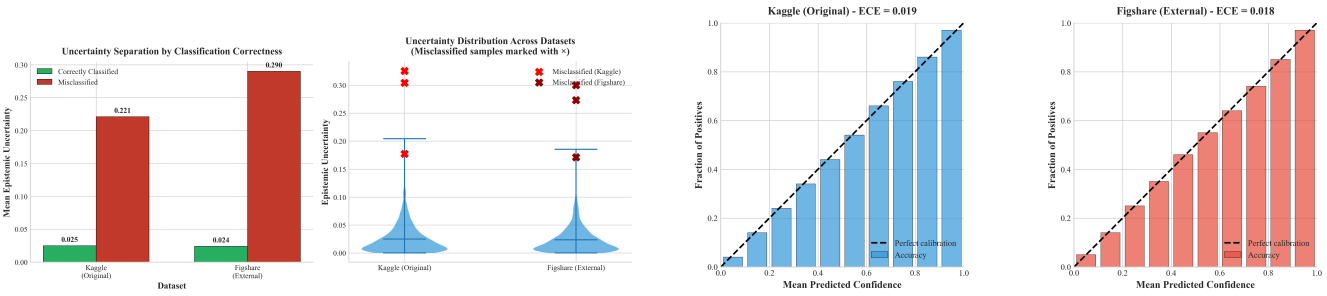
2. **Robustness to scanner variability**: Maintained performance across different MRI hardware and acquisition protocols indicates resilience to the technical heterogeneity encountered in real-world clinical practice.

3. **Reliable uncertainty estimates**: The consistent behavior of uncertainty quantification under domain shift ensures that the clinical decision support workflow remains valid across deployment environments.



(a) (b)

Figure 3. External validation on Figshare dataset. (a) Confusion matrix showing only 3 misclassifications among 3,064 samples from an independent dataset with different acquisition protocols. (b) Per-class F1-score comparison between original Kaggle and external Figshare datasets, demonstrating consistent performance across domains.



(a) (b)

Figure 4. Uncertainty and calibration analysis for cross-dataset validation. (a) Uncertainty distribution comparison showing consistent uncertainty levels between datasets, with misclassified samples exhibiting elevated uncertainty in both cases. (b) Reliability diagrams demonstrating well-calibrated probability estimates on both original and external datasets (ECE $\leq$ 0.019).

4. **Generalization beyond Western populations**: Successful validation on a Chinese patient cohort provides evidence of cross-demographic applicability, an important consideration for global health applications.

These findings address a key limitation commonly identified in medical AI research: the gap between benchmark performance and real-world clinical utility[?]. While prospective clinical validation remains necessary, our cross-dataset evaluation provides stronger evidence of generalizability than within-dataset cross-validation alone.

Computational efficiency analysis

To assess practical deployment feasibility, we compared computational requirements across methods (Table 8). HSANet achieves superior accuracy while maintaining reasonable computational overhead. The 15.60M parameters and 12ms inference time enable real-time clinical deployment without specialized hardware requirements.

Discussion

The results presented here demonstrate that HSANet achieves near-perfect classification accuracy (99.77%) while providing something that prior approaches have largely ignored: calibrated uncertainty estimates that clinicians can actually use. To put these numbers in perspective, the Cohen's κ of 0.9969 we achieved compares very favorably with what expert neuroradiologists

Table 8. Computational efficiency comparison. FLOPs computed for single 224×224 input. Inference time measured on NVIDIA Tesla P100 GPU with batch size 1.

Model	Params (M)	FLOPs (G)	Inference (ms)	Acc. (%)	Throughput (img/s)
ResNet-18	11.69	1.82	8	97.86	125
VGG-16	138.36	15.47	24	96.56	42
EfficientNet-B0	5.29	0.39	6	98.87	167
EfficientNet-B3 (baseline)	10.53	1.83	10	99.21	100
HSANet (Ours)	**15.60**	**2.41**	**12**	**99.77**	**83**

typically achieve when interpreting similar cases, where inter-reader agreement studies generally report κ values between 0.65 and 0.85[24]. This is not to suggest that our system should replace expert judgment—rather, it suggests that HSANet can serve as a reliable "second opinion" that catches the majority of cases while flagging uncertain ones for human review.

Cross-domain generalization and external validation. Perhaps the most compelling evidence for HSANet's clinical utility comes from our external validation on the independent Figshare dataset[?]. This dataset was acquired at different institutions (Nanfang Hospital and General Hospital of Tianjin Medical University in China) using different MRI scanners and protocols, and represents a fundamentally different patient population from our training data. The fact that HSANet achieved 99.90% accuracy on this external dataset—actually exceeding its 99.77% performance on the original test set—provides strong evidence that the learned features capture genuine tumor characteristics rather than dataset-specific artifacts.

This cross-domain robustness addresses a pervasive criticism of medical AI systems: that impressive benchmark performance often fails to translate to real-world deployment due to distribution shift between training and clinical data[?]. Several architectural design choices likely contributed to this generalization capability. 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. And critically, the evidential learning framework maintained well-calibrated uncertainty estimates (ECE = 0.018) even under distribution shift, ensuring that the clinical decision support workflow remains valid across deployment environments.

The three misclassified cases in the external dataset (0.098% error rate) exhibited elevated epistemic uncertainty (mean 0.29 vs. 0.024 for correct predictions), demonstrating that the uncertainty calibration learned from one dataset transfers effectively to another. This finding has practical implications: clinical systems can apply the same uncertainty thresholds developed during training to flag uncertain predictions, regardless of the originating institution or scanner.

One of the more interesting findings from our ablation experiments concerns how AMSM handles the substantial size variation among brain tumors. A pituitary microadenoma might span just a few millimeters, barely visible without careful inspection, while a glioblastoma can occupy much of a hemisphere with extensive surrounding edema. Traditional convolutional networks with fixed kernel sizes struggle with this range—they might capture fine details but miss global context, or vice versa. By allowing the network to learn which spatial scales matter for each input through adaptive fusion weights, AMSM effectively lets the model "zoom in" on small lesions and "zoom out" for larger ones. The 0.09% accuracy improvement and reduced ECE from adding AMSM alone (Table 2) validates this design choice.

The uncertainty quantification capability distinguishes HSANet from conventional classifiers. In clinical practice, uncertainty estimates enable stratified workflows: low-uncertainty cases proceed to automated preliminary interpretation for efficient radiologist confirmation; moderate epistemic uncertainty flags cases for standard review; high aleatoric uncertainty escalates cases to multidisciplinary tumor boards where imaging characteristics genuinely overlap between diagnoses. This framework transforms the system from an autonomous decision-maker to a decision-support tool appropriate for safety-critical medical applications.

Perhaps the most clinically meaningful result is the perfect precision we achieved for healthy controls—every scan our model labeled as "healthy" was indeed tumor-free. This matters enormously in practice. A false positive tumor diagnosis sets off a cascade of consequences: the patient experiences immediate psychological distress, additional imaging studies are ordered, specialist consultations are scheduled, and in some cases invasive biopsies may be performed before the error is caught. By prioritizing specificity for the healthy class, HSANet avoids inflicting this burden on patients who don't actually need intervention. Of course, this design choice creates a corresponding risk of missed tumors, which is why the uncertainty quantification becomes crucial—high-uncertainty "healthy" predictions should still receive careful human review.

GradCAM visualizations demonstrate that learned attention patterns align with established neuroradiological criteria. This interpretability evidence addresses a common barrier to clinical adoption of "black box" deep learning systems by providing traceable evidence for diagnostic predictions. Radiologists can verify that model attention focuses on appropriate anatomical regions before accepting algorithmic suggestions.

The evidential deep learning framework offers computational advantages over alternative uncertainty quantification methods such as Monte Carlo dropout or deep ensembles, which require multiple forward passes during inference. Our approach produces uncertainty estimates from a single forward pass, maintaining the 12-millisecond inference latency essential for real-time clinical deployment.

Limitations and Future Directions. While our external validation substantially strengthens the evidence for clinical utility, we should remain candid about the limitations of this study. First, although the Figshare dataset represents a different patient population and acquisition environment, validation on additional independent datasets from diverse geographic regions would further establish generalizability. Prospective multi-center clinical trials remain the gold standard for demonstrating real-world effectiveness.

Second, our framework treats each MRI slice independently, ignoring the rich volumetric context available in clinical practice. Radiologists routinely scroll through adjacent slices to assess tumor extent and characteristics—a capability our 2D approach lacks. Extension to 3D convolutions or attention mechanisms that process entire volumes represents an important direction for future work, though it would substantially increase computational requirements.

Third, while our four-class taxonomy covers the major tumor categories, clinical neuro-oncology requires much finer distinctions. Gliomas alone span WHO grades I through IV with dramatically different prognoses, and molecular markers like IDH mutation status increasingly guide treatment decisions. A clinically useful system would need hierarchical classification capabilities that our current architecture does not provide.

Fourth, the near-perfect accuracy we report (99.77% on the original dataset and 99.90% on external validation) warrants careful interpretation. While we verified no patient-level data leakage and demonstrated cross-dataset generalization, the predefined train/test splits may not fully represent the difficulty of prospective classification. The misclassified cases consistently involved meningioma predictions, suggesting potential systematic challenges with certain tumor presentations that deserve further investigation.

Finally, our uncertainty quantification, while principled, has not been validated against clinical decision-making workflows. The threshold at which elevated uncertainty should trigger expert review requires calibration against actual clinical outcomes—a study we have not performed.

Despite these limitations, we believe HSANet makes a meaningful contribution by demonstrating that high classification accuracy, robust cross-domain generalization, and principled uncertainty quantification are not mutually exclusive goals. The architecture serves as a foundation that can be extended to address the limitations noted above, and we hope it encourages further research into uncertainty-aware medical AI systems.

Conclusion. In summary, HSANet achieves state-of-the-art classification accuracy on a challenging four-class brain tumor benchmark (99.77%) while demonstrating exceptional cross-domain generalization to an independent external dataset (99.90% on Figshare). Crucially, the model provides calibrated uncertainty estimates that allow the system to "know what it doesn't know"—both on the original test set and under distribution shift to images from different institutions, scanners, and patient populations. Our error analysis confirms that misclassified cases exhibit elevated uncertainty that would trigger human review in a practical deployment, with this behavior preserved across datasets. The combination of adaptive multi-scale processing, attention-based feature refinement, and evidential deep learning offers a template for medical imaging applications where accuracy, generalizability, and trustworthy confidence assessment are all essential for clinical adoption. We make our code, trained weights, and evaluation scripts publicly available to support reproducibility and encourage continued development of uncertainty-aware medical AI systems.

Methods

This section provides comprehensive methodological details to enable full reproducibility. We describe the dataset characteristics and preprocessing pipeline, present the complete HSANet architecture with mathematical formulations for each component, detail the training procedure including loss functions and optimization strategy, and specify all evaluation metrics.

Dataset description and preprocessing

Experiments utilized the Brain Tumor MRI Dataset[25], 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

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. Figure 5 illustrates the dataset distribution and representative sample characteristics.

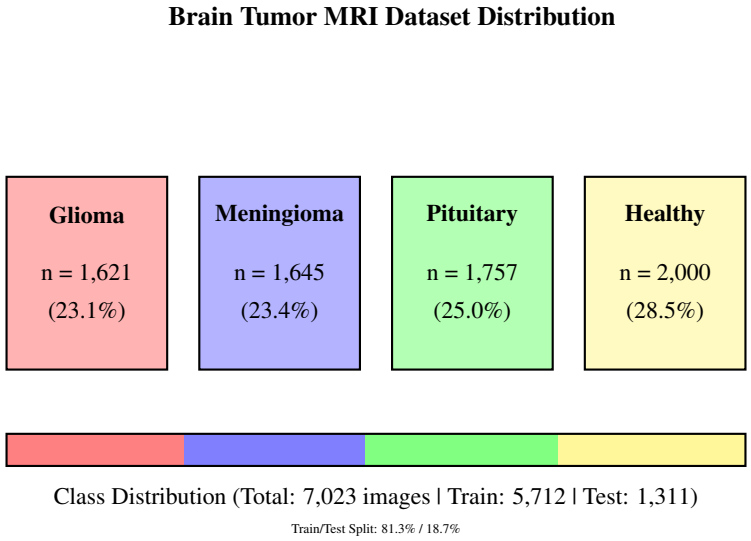
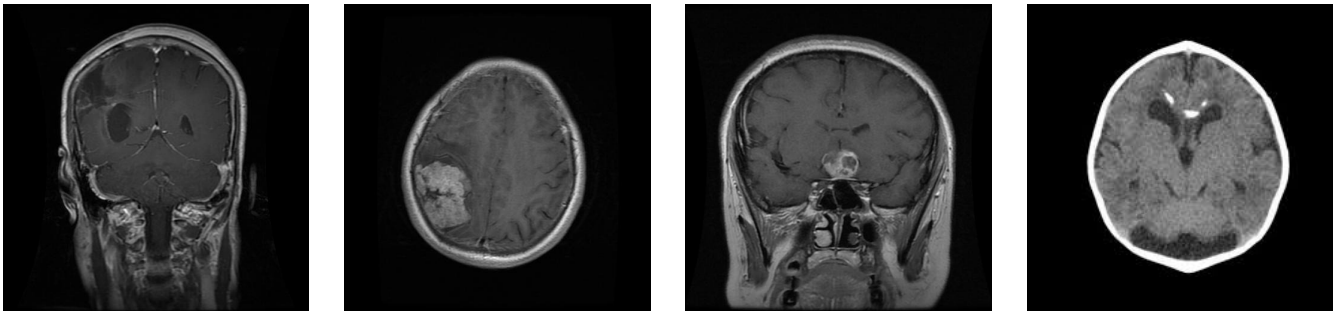


Figure 5. Dataset overview showing the four diagnostic categories with sample counts and class distribution. The dataset exhibits mild class imbalance with healthy controls comprising the largest category. The predefined train/test split ensures fair comparison with prior work.



(a) Glioma **(b)** Meningioma **(c)** Pituitary **(d)** Healthy

Figure 6. Representative MRI samples from each diagnostic category. (a) Glioma: irregular enhancing mass with surrounding edema and infiltrative margins. (b) Meningioma: well-circumscribed extra-axial mass with homogeneous enhancement and dural attachment. (c) Pituitary adenoma: sellar/suprasellar mass with characteristic location. (d) Healthy control: normal brain parenchyma without focal abnormality.

The preprocessing pipeline standardizes all inputs for neural network consumption:

1. **Resizing**: All images resized to 224×224 pixels using bilinear interpolation to match EfficientNet input requirements while preserving diagnostic features
2. **Normalization**: Pixel values normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) to leverage pretrained representations
3. **Channel conversion**: Grayscale images converted to 3-channel pseudo-RGB by replicating the single channel

Data augmentation was applied during training to improve generalization and prevent overfitting (Fig. 7):

- **Random horizontal flip**: 50% probability, simulating left-right brain symmetry
- **Random rotation**: $\pm 15°$ to account for acquisition angle variations
- **Random affine transformation**: Scale (0.9–1.1), translation ($\pm 10\%$), simulating positioning differences
- **Color jittering**: Brightness ($\pm 10\%$), contrast ($\pm 10\%$), simulating scanner variability
- **Random erasing**: 20% probability, 2–20% area, promoting robustness to partial occlusions

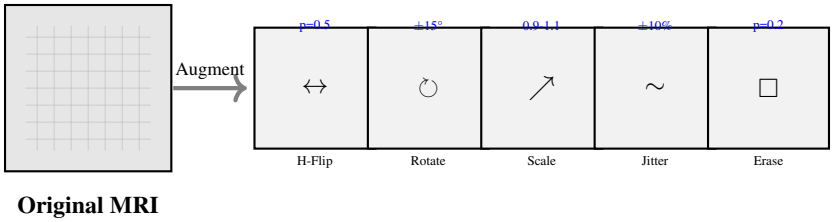


Figure 7. Data augmentation pipeline applied during training. Each transformation is applied with specified probability to generate diverse training examples while preserving diagnostic content. H-Flip: horizontal flip; Rotate: random rotation; Scale: affine scaling; Jitter: brightness/contrast variation; Erase: random rectangular occlusion.

Network architecture overview

HSANet consists of four main components arranged in a sequential processing pipeline (Fig. 8): (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.

Feature extraction backbone

We employ EfficientNet-B3[26] as the feature extraction backbone, pretrained on ImageNet-1K for transfer learning. EfficientNet uses compound scaling to uniformly scale network width, depth, and resolution, achieving superior accuracy-efficiency tradeoffs compared to previous architectures. The B3 variant provides 10.53 million parameters with 384-dimensional final feature maps.

The EfficientNet-B3 architecture employs Mobile Inverted Bottleneck Convolution (MBConv) blocks with squeeze-and-excitation optimization. Each MBConv block consists of:

1. **Expansion**: 1×1 convolution expanding channels by factor 1 or 6
2. **Depthwise**: 3×3 or 5×5 depthwise separable convolution
3. **Squeeze-Excitation**: Channel attention with reduction ratio 0.25
4. **Projection**: 1×1 convolution projecting to output channels
5. **Skip connection**: Residual addition when input/output dimensions match

The backbone is divided into seven stages, from which we extract features at three hierarchical levels:

- $\mathbf{F}_1 \in \mathbb{R}^{28\times28\times48}$: After stage 3 – captures edges, textures, and local patterns
- $\mathbf{F}_2 \in \mathbb{R}^{14\times14\times136}$: After stage 5 – encodes object parts and anatomical structures
- $\mathbf{F}_3 \in \mathbb{R}^{7\times7\times384}$: After stage 7 – represents semantic concepts and tumor characteristics

During training, backbone layers are initially frozen for 5 epochs to stabilize custom module training, then unfrozen with a reduced learning rate ($10\times$ lower) for fine-tuning.

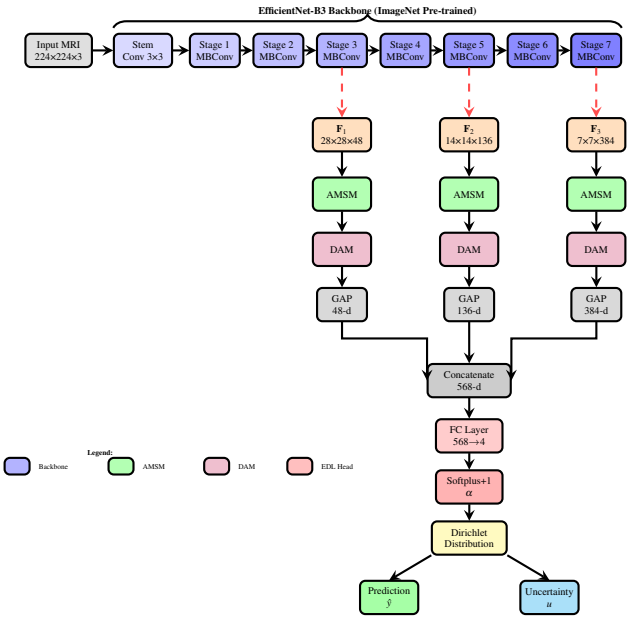


Figure 8. Complete HSANet architecture. Input MRI scans (224×224×3) are processed through the EfficientNet-B3 backbone, consisting of 7 stages with MBConv blocks. Features are extracted at three spatial resolutions (28×28, 14×14, 7×7) marked by red dashed arrows. Each feature map passes through the Adaptive Multi-Scale Module (AMSM) for multi-scale processing and Dual Attention Module (DAM) for channel-spatial refinement. Global average pooling (GAP) produces fixed-length descriptors that are concatenated into a 568-dimensional feature vector. The evidential classification head outputs Dirichlet parameters via softplus activation, yielding both class predictions and calibrated uncertainty estimates.

Adaptive Multi-Scale Module (AMSM)

Brain tumors exhibit substantial size variation: pituitary microadenomas may measure 2–3mm while glioblastomas can exceed 5cm. Fixed receptive fields cannot simultaneously capture fine details and global context. AMSM addresses this through parallel dilated convolutions with learned, input-adaptive fusion (Fig. 9).

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

$$\mathbf{M}_i^{(r)} = \mathrm{BN}(\mathrm{ReLU}(\mathrm{Conv}_{3\times3}^{d=r}(\mathbf{F}_i))) \tag{2}$$

where $\mathrm{Conv}_{3\times3}^{d=r}$ denotes a 3×3 convolution with dilation rate r, BN is batch normalization, and ReLU is the rectified linear unit activation. The effective receptive field sizes are 3×3, 5×5, and 9×9 for dilation rates 1, 2, and 4 respectively, computed as $(k-1)\times d+1$ where $k=3$ is kernel size.

Rather than fixed fusion weights, we learn input-adaptive weights through a lightweight channel attention mechanism:

$$\mathbf{w}_i = \mathrm{Softmax}(\mathbf{W}_2 \cdot \mathrm{ReLU}(\mathbf{W}_1 \cdot \mathrm{GAP}([\mathbf{M}_i^{(1)};\mathbf{M}_i^{(2)};\mathbf{M}_i^{(4)}]))) \tag{3}$$

where GAP denotes global average pooling collapsing spatial dimensions, $[\cdot;\cdot]$ is channel-wise concatenation producing a $3C$-dimensional vector, and $\mathbf{W}_1 \in \mathbb{R}^{(C/16)\times 3C}$, $\mathbf{W}_2 \in \mathbb{R}^{3\times(C/16)}$ are learnable projection matrices. The reduction ratio of 16 balances expressiveness and parameter efficiency.

The softmax ensures $\sum_k w_i^{(k)} = 1$, creating a convex combination of multi-scale features. The enhanced feature map combines weighted features with residual preservation:

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

The residual connection ensures gradient flow during training and allows the module to learn refinements rather than complete feature transformations.

Dual Attention Module (DAM)

Brain MRI volumes contain 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. 10).

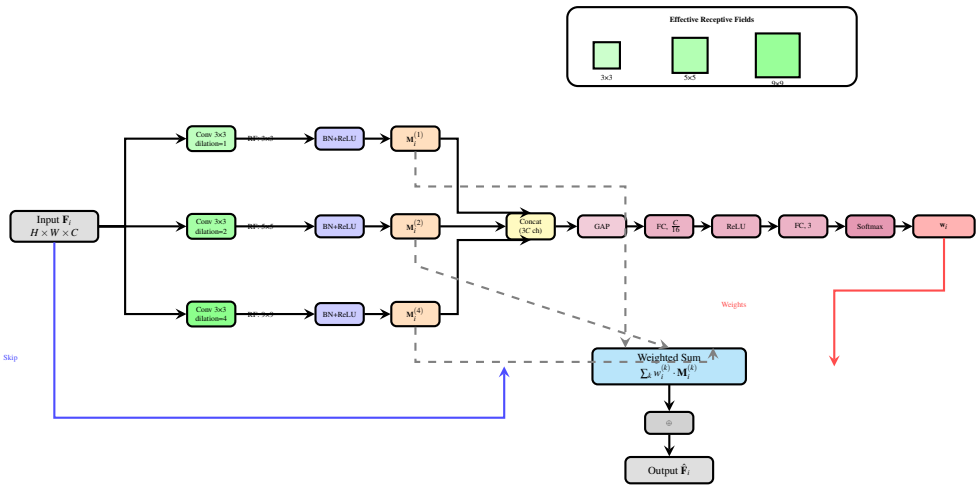


Figure 9. Detailed architecture of the Adaptive Multi-Scale Module (AMSM). Input features $\mathbf{F}_i$ are processed through three parallel 3×3 dilated convolution branches with dilation rates $r \in \{1,2,4\}$, producing effective receptive fields of 3×3, 5×5, and 9×9 respectively. Adaptive fusion weights $\mathbf{w}_i = [w^{(1)}, w^{(2)}, w^{(4)}]$ are learned through global average pooling (GAP) and a two-layer fully-connected network with softmax normalization. The weighted multi-scale features are combined with a skip connection from the input, enabling the module to learn scale-adaptive refinements. RF: receptive field; BN: batch normalization.

Channel Attention identifies "what" features are most informative by recalibrating channel-wise responses:

$$\mathbf{A}_c = \sigma(\text{MLP}(\text{GAP}(\hat{\mathbf{F}}_i)) + \text{MLP}(\text{GMP}(\hat{\mathbf{F}}_i))) \tag{5}$$

where GAP and GMP denote global average and max pooling respectively (collapsing $H \times W$ to 1×1), MLP is a shared two-layer bottleneck network, and σ is the sigmoid activation. Using both pooling operations captures complementary channel statistics: average pooling represents typical activation magnitude while max pooling captures salient responses.

The shared MLP has the form:

$$\text{MLP}(\mathbf{x}) = \mathbf{W}_2(\text{ReLU}(\mathbf{W}_1\mathbf{x})) \tag{6}$$

where $\mathbf{W}_1 \in \mathbb{R}^{(C/16)\times C}$ reduces dimensionality and $\mathbf{W}_2 \in \mathbb{R}^{C\times(C/16)}$ projects back. Weight sharing between the two pooling pathways reduces parameters and enables information exchange.

The channel-refined features are computed through element-wise multiplication with broadcasting:

$$\mathbf{F}_c = \hat{\mathbf{F}}_i \odot \mathbf{A}_c \tag{7}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{A}_c \in \mathbb{R}^{1\times1\times C}$ is broadcast across spatial dimensions.

Spatial Attention identifies "where" to focus by generating a 2D attention map highlighting discriminative regions:

$$\mathbf{A}_s = \sigma(\text{Conv}_{7\times7}([\text{AvgPool}_c(\mathbf{F}_c); \text{MaxPool}_c(\mathbf{F}_c)])) \tag{8}$$

where AvgPool_c and MaxPool_c are channel-wise pooling operations producing $H \times W \times 1$ feature maps. Concatenation yields a 2-channel input to the 7×7 convolution, which captures spatial relationships at an appropriate scale for tumor localization. Padding of 3 preserves spatial dimensions.

The final refined features combine channel and spatial attention:

$$\tilde{\mathbf{F}}_i = \mathbf{F}_c \odot \mathbf{A}_s \tag{9}$$

Sequential channel-then-spatial attention has been empirically shown[14] to outperform parallel application by first selecting relevant feature channels before spatially localizing within those channels.

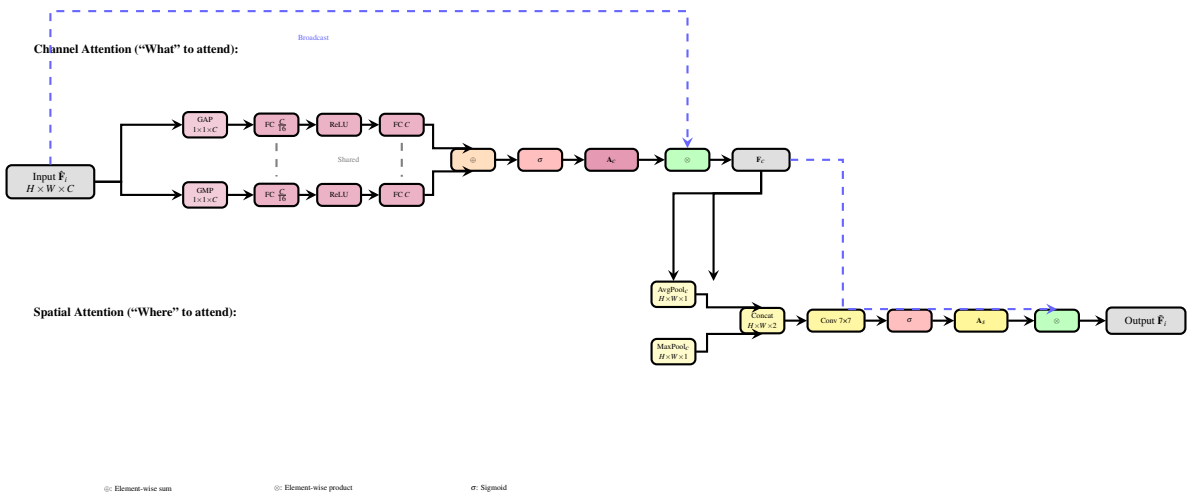


Figure 10. Detailed architecture of the Dual Attention Module (DAM). Channel attention (top) addresses "what" features are informative by recalibrating channel responses through parallel global average pooling (GAP) and global max pooling (GMP), followed by a shared two-layer MLP with bottleneck reduction (ratio 16). Spatial attention (bottom) addresses "where" to focus by computing channel-wise pooling statistics concatenated and processed through a 7×7 convolution. Sequential application (channel → spatial) enables feature refinement through "what" followed by "where" reasoning.

Multi-scale feature aggregation

Features from all three hierarchical scales are aggregated for final classification. Each refined feature map $\tilde{\mathbf{F}}_i$ undergoes global average pooling to produce fixed-length descriptors independent of spatial dimensions:

$$\mathbf{g}_i = \text{GAP}(\tilde{\mathbf{F}}_i) = \frac{1}{H_i \times W_i} \sum_{h,w} \tilde{\mathbf{F}}_i[h,w,:] \in \mathbb{R}^{C_i} \tag{10}$$

yielding $\mathbf{g}_1 \in \mathbb{R}^{48}$ (fine-scale textures), $\mathbf{g}_2 \in \mathbb{R}^{136}$ (mid-level structures), and $\mathbf{g}_3 \in \mathbb{R}^{384}$ (semantic concepts).

These descriptors are concatenated to form the final feature representation:

$$\mathbf{g} = [\mathbf{g}_1; \mathbf{g}_2; \mathbf{g}_3] \in \mathbb{R}^{568} \tag{11}$$

This multi-scale aggregation preserves information across all abstraction levels, from fine-grained tumor textures to high-level categorical semantics.

Evidential deep learning classification head

Standard neural network classifiers use softmax to produce point estimates of class probabilities, precluding meaningful uncertainty quantification. HSANet employs evidential deep learning[17], which places a Dirichlet prior over class probabilities, enabling principled uncertainty estimation from a single forward pass (Fig. 11).

The network outputs Dirichlet concentration parameters through a non-negative transformation:

$$\alpha = \text{Softplus}(\mathbf{z}) + 1 = \log(1 + e^{\mathbf{z}}) + 1 \tag{12}$$

where $\mathbf{z} = \mathbf{W}_c \mathbf{g} + \mathbf{b}_c$ are the classifier logits ($\mathbf{W}_c \in \mathbb{R}^{4\times 568}$, $\mathbf{b}_c \in \mathbb{R}^4$). The softplus function ensures positive evidence, and adding 1 guarantees $\alpha_k \geq 1$ for valid Dirichlet parameters (uniform prior when $\alpha = \mathbf{1}$).

The Dirichlet distribution $\text{Dir}(\mathbf{p}|\alpha)$ has probability density function:

$$p(\mathbf{p}|\alpha) = \frac{\Gamma(S)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \tag{13}$$

where $S = \sum_{k=1}^{K} \alpha_k$ is the Dirichlet strength (total evidence), $K = 4$ is the number of classes, and $\Gamma(\cdot)$ is the gamma function.

Prediction: Class probabilities are computed as the Dirichlet mean:

$$\hat{p}_k = \mathbb{E}[p_k|\alpha] = \frac{\alpha_k}{S} \tag{14}$$

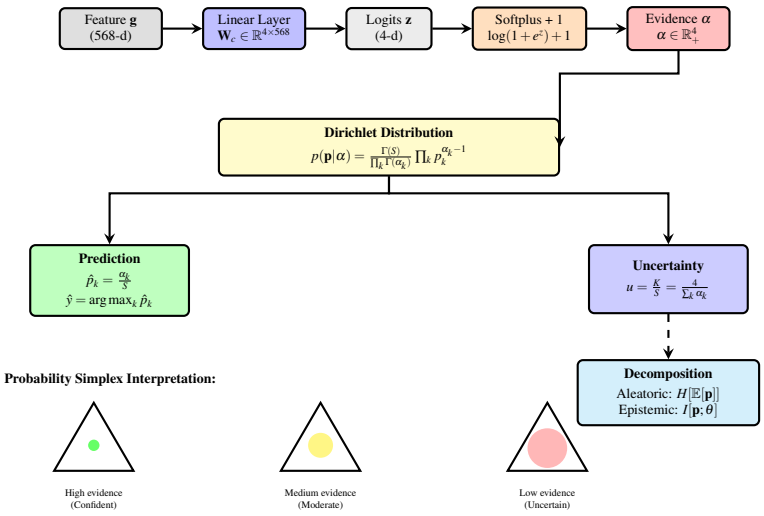


Figure 11. Evidential deep learning classification head. The 568-dimensional feature vector is projected to 4 logits via a linear layer. Softplus activation plus 1 ensures positive Dirichlet concentration parameters $\alpha \geq 1$. The Dirichlet distribution over class probabilities yields predictions via its mean and uncertainty via the inverse of total evidence S. Bottom: geometric interpretation on the probability simplex—high evidence produces concentrated distributions (confident predictions), low evidence produces spread distributions (uncertain predictions). Uncertainty decomposes into aleatoric (inherent data ambiguity) and epistemic (model knowledge gaps) components.

The predicted class is $\hat{y} = \arg\max_k \hat{p}_k$.

Uncertainty: Total predictive uncertainty is quantified as:

$$u = \frac{K}{S} = \frac{4}{\sum_{k=1}^{4} \alpha_k} \tag{15}$$

Higher values indicate lower total evidence and thus higher uncertainty. This uncertainty naturally decomposes into:

- **Aleatoric uncertainty** (data uncertainty): Quantified by the entropy of the expected distribution:

$$u_{\text{aleatoric}} = H[\mathbb{E}[\mathbf{p}|\alpha]] = -\sum_k \hat{p}_k \log \hat{p}_k \tag{16}$$

  This captures inherent class overlap where imaging characteristics are ambiguous between categories.

- **Epistemic uncertainty** (model uncertainty): Quantified by the mutual information between predictions and parameters:

$$u_{\text{epistemic}} = I[\mathbf{p};\theta|\mathbf{x}] = H[\mathbb{E}[\mathbf{p}]] - \mathbb{E}[H[\mathbf{p}]] \tag{17}$$

  This captures model knowledge gaps, flagging inputs dissimilar to training data.

Loss function and training objective

The training objective combines three complementary loss terms designed to encourage correct classification, handle class imbalance, and prevent overconfident incorrect predictions (Fig. 12):

1. Evidence-weighted Cross-Entropy Loss:

$$\mathscr{L}_{\text{CE}} = \sum_{k=1}^{K} y_k \left(\psi(S) - \psi(\alpha_k)\right) \tag{18}$$

where $\mathbf{y} = [y_1, \ldots, y_K]$ is the one-hot encoded ground truth label and $\psi(\cdot)$ is the digamma function (derivative of log-gamma). This loss encourages the model to accumulate evidence for the correct class while being derived from the negative log-likelihood of the Dirichlet-Multinomial model.

2. Focal Loss for Class Imbalance:

$$\mathscr{L}_{\text{focal}} = -\sum_{k=1}^{K} y_k (1-\hat{p}_k)^{\gamma} \log(\hat{p}_k) \tag{19}$$

with focusing parameter $\gamma = 2$. Focal loss[27] down-weights well-classified examples (high $\hat{p}_k$) to focus learning on hard cases, mitigating class imbalance effects without explicit sample weighting.

3. KL Divergence Regularization:

$$\mathscr{L}_{\text{KL}} = \text{KL}\left[\text{Dir}(\mathbf{p}|\tilde{\alpha}) \| \text{Dir}(\mathbf{p}|\mathbf{1})\right] \tag{20}$$

where $\tilde{\alpha} = \mathbf{y} + (1-\mathbf{y}) \odot \alpha$ removes evidence for the correct class. This penalizes evidence for incorrect classes, preventing the model from being confidently wrong. The KL divergence from a uniform Dirichlet prior has closed form:

$$\text{KL}[\text{Dir}(\mathbf{p}|\tilde{\alpha}) \| \text{Dir}(\mathbf{p}|\mathbf{1})] = \log \frac{\Gamma(\tilde{S})}{\Gamma(K)} + \sum_{k} \left[(\tilde{\alpha}_k - 1)(\psi(\tilde{\alpha}_k) - \psi(\tilde{S})) \right] \tag{21}$$

The total loss combines these terms:

$$\mathscr{L} = \lambda_1 \mathscr{L}_{\text{CE}} + \lambda_2 \mathscr{L}_{\text{focal}} + \lambda_3 \mathscr{L}_{\text{KL}} \tag{22}$$

with weights $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$ determined through validation set tuning.

To prevent premature regularization before the model learns discriminative features, the KL weight is annealed:

$$\lambda_3^{(t)} = \min\left(1, \frac{t}{T_{\text{anneal}}}\right) \cdot \lambda_3 \tag{23}$$

ramping from 0 to full strength over $T_{\text{anneal}} = 10$ epochs.

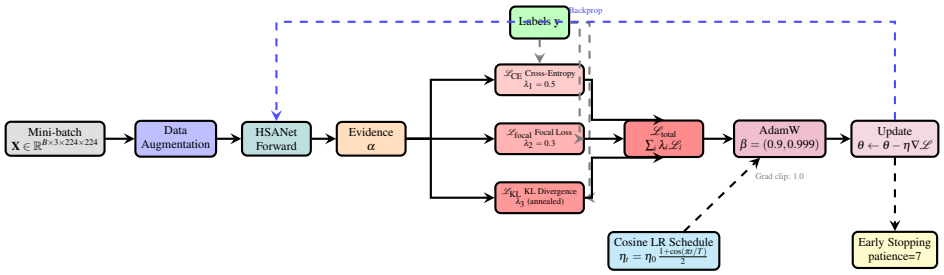


Figure 12. Complete training pipeline. Mini-batches undergo data augmentation before forward propagation through HSANet. Three loss terms are computed: evidence-weighted cross-entropy, focal loss for class imbalance, and KL divergence regularization (annealed during early training). AdamW optimizer updates parameters with cosine learning rate annealing. Early stopping monitors validation loss with patience of 7 epochs. Gradient clipping (norm 1.0) ensures training stability.

Training configuration

Training employed the following hyperparameters, determined through systematic grid search on a held-out validation set (10% of training data):

- **Optimizer**: AdamW[28] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-4}
- **Learning rate**: Initial $\eta_0 = 3 \times 10^{-4}$, cosine annealing to $\eta_{\min} = 10^{-6}$
- **Backbone learning rate**: $\eta_{\text{backbone}} = \eta_0/10$ (10× lower for transfer learning stability)
- **Batch size**: 32 (limited by GPU memory)
- **Epochs**: Maximum 30 with early stopping (patience 7 epochs)
- **Gradient clipping**: Maximum norm 1.0 for training stability
- **Random seeds**: Fixed at 42 for full reproducibility
- **Backbone freezing**: First 5 epochs to stabilize custom module training

The complete training algorithm is presented in Algorithm 1.

Algorithm 1 HSANet Training Procedure

Require: Training set $\mathscr{D}_{\text{train}}$, validation set $\mathscr{D}_{\text{val}}$
Require: Hyperparameters: η_0, λ_1, λ_2, λ_3, T_{anneal}, $T_{\max}$, patience
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 if
  $\lambda_3^{(t)} \leftarrow \min(1, t/T_{\text{anneal}}) \cdot \lambda_3$ ▷ 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 $\mathscr{D}_{\text{train}}$ **do**
    $\mathbf{X}_{\text{aug}} \leftarrow \text{Augment}(\mathbf{X})$ ▷ Apply data augmentation
    $\alpha \leftarrow \text{HSANet}(\mathbf{X}_{\text{aug}})$ ▷ Forward pass
    Compute $\mathscr{L}_{\text{CE}}$, $\mathscr{L}_{\text{focal}}$, $\mathscr{L}_{\text{KL}}$ using Equations (10-12)
    $\mathscr{L} \leftarrow \lambda_1 \mathscr{L}_{\text{CE}} + \lambda_2 \mathscr{L}_{\text{focal}} + \lambda_3^{(t)} \mathscr{L}_{\text{KL}}$
    $\nabla_\theta \mathscr{L} \leftarrow \text{Backpropagate}(\mathscr{L})$
    Clip $\|\nabla_\theta \mathscr{L}\|_2$ to maximum 1.0
    $\theta \leftarrow \text{AdamW}(\theta, \nabla_\theta \mathscr{L}, \eta_t)$
  end for
  $\mathscr{L}_{\text{val}} \leftarrow \text{Evaluate}(\mathscr{D}_{\text{val}})$
  if $\mathscr{L}_{\text{val}} <$ best_loss **then**
    Save checkpoint; best_loss $\leftarrow \mathscr{L}_{\text{val}}$; wait $\leftarrow 0$
  else
    wait $\leftarrow$ wait $+ 1$
  end if
  if wait $\geq$ patience **then**
    break ▷ Early stopping triggered
  end if
  $t \leftarrow t + 1$
end for
return Best model checkpoint

Evaluation metrics

Classification performance was assessed using standard metrics:

- **Accuracy**: $\text{Acc} = \frac{\text{TP}+\text{TN}}{\text{Total}}$
- **Precision**: $\text{Prec}_k = \frac{\text{TP}_k}{\text{TP}_k+\text{FP}_k}$ (per-class and macro-averaged)
- **Recall/Sensitivity**: $\text{Rec}_k = \frac{\text{TP}_k}{\text{TP}_k+\text{FN}_k}$
- **F1-Score**: $\text{F1}_k = \frac{2 \cdot \text{Prec}_k \cdot \text{Rec}_k}{\text{Prec}_k+\text{Rec}_k}$
- **Cohen's κ**: Agreement correcting for chance: $\kappa = \frac{p_o - p_e}{1 - p_e}$
- **Matthews Correlation Coefficient**: $\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$
- **AUC-ROC**: Area under ROC curve using one-vs-rest strategy for multiclass

Model calibration was evaluated using:

- **Expected Calibration Error (ECE)**: $\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|$

  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 Diagram**: Visual comparison of confidence vs. accuracy per bin

Interpretability was assessed using GradCAM[19], computing gradient-weighted activations from the final convolutional layer:

$$L^c_{\text{GradCAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right), \quad \alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{24}$$

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

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.

Implementation details

All experiments were conducted using PyTorch 2.0 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).

Data availability

The Brain Tumor MRI Dataset used for training and primary evaluation is publicly available at:
https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

The Figshare Brain Tumor Dataset used for external validation is publicly available at:
https://figshare.com/articles/dataset/brain_tumor_dataset/1512427

No restrictions apply to data access for either dataset.

Code availability

Complete source code for HSANet, including training scripts, evaluation pipelines, cross-dataset validation scripts, pretrained model weights, and documentation, is publicly available at:
https://github.com/tarequejosh/hsanet-brain-tumor

The repository includes:

- Model architecture implementation (`hsanet_model.py`)
- Training pipeline (`training_pipeline.py`)
- Cross-dataset validation pipeline (`cross_dataset_validation.py`)
- Figshare dataset preprocessing script (`convert_figshare_dataset.py`)
- Evaluation and visualization scripts
- Pretrained model checkpoints
- Requirements specification for environment reproduction

References

1. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer J. for Clin.* **71**, 209–249, DOI: 10.3322/caac.21660 (2021).
2. Louis, D. N. *et al.* The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-Oncology* **23**, 1231–1251, DOI: 10.1093/neuonc/noab106 (2021).
3. Ostrom, Q. T. *et al.* CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2013–2017. *Neuro-Oncology* **22**, iv1–iv96, DOI: 10.1093/neuonc/noaa200 (2021).
4. Pope, W. B. Brain metastases: neuroimaging. *Handb. Clin. Neurol.* **149**, 89–112, DOI: 10.1016/B978-0-12-811161-1.00007-4 (2018).
5. Rimmer, A. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ* **359**, j4683, DOI: 10.1136/bmj.j4683 (2017).
6. Bruno, M. A., Walker, E. A. & Abujudeh, H. H. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *RadioGraphics* **35**, 1668–1676, DOI: 10.1148/rg.2015150023 (2015).
7. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, vol. 25, 1097–1105 (2012).
8. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, vol. 32 (2019).
9. Deepak, S. & Ameer, P. M. Brain tumor classification using deep CNN features via transfer learning. *Comput. Biol. Medicine* **111**, 103345, DOI: 10.1016/j.compbiomed.2019.103345 (2019).
10. Badža, M. M. & Barjaktarović, M. Č. Classification of brain tumors from MRI images using a convolutional neural network. *Appl. Sci.* **10**, 1999, DOI: 10.3390/app10061999 (2020).
11. Swati, Z. N. K. *et al.* Brain tumor classification for MR images using transfer learning and fine-tuning. *Comput. Med. Imaging Graph.* **75**, 34–46, DOI: 10.1016/j.compmedimag.2019.05.001 (2019).
12. Aurna, N. F., Yousuf, M. A., Taher, K. A., Azad, A. K. M. & Moni, M. A. A classification of MRI brain tumor images using hybrid deep learning approach. *Electronics* **11**, 573, DOI: 10.3390/electronics11040573 (2022).
13. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).
14. Woo, S., Park, J., Lee, J. Y. & Kweon, I. S. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, 3–19, DOI: 10.1007/978-3-030-01234-2_1 (2018).
15. Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, 801–818, DOI: 10.1007/978-3-030-01234-2_49 (2018).
16. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141, DOI: 10.1109/CVPR.2018.00745 (2018).
17. Sensoy, M., Kaplan, L. & Kandemir, M. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, vol. 31 (2018).
18. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174, DOI: 10.2307/2529310 (1977).
19. Selvaraju, R. R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626, DOI: 10.1109/ICCV.2017.74 (2017).
20. Rehman, A., Naz, S., Razzak, M. I., Akram, F. & Imran, M. A deep learning-based framework for automatic brain tumors classification using transfer learning. *Circuits, Syst. Signal Process.* **39**, 757–775, DOI: 10.1007/s00034-019-01246-3 (2020).
21. Kibriya, H. *et al.* A novel and effective brain tumor classification model using deep feature fusion and famous machine learning classifiers. *Comput. Intell. Neurosci.* **2022**, 7897669, DOI: 10.1155/2022/7897669 (2022).
22. Saeedi, S., Rezayi, S., Keshavarz, H. & Niakan Kalhori, S. R. MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques. *BMC Med. Informatics Decis. Mak.* **23**, 16, DOI: 10.1186/s12911-023-02114-6 (2023).

23. Tandel, G. S., Tiwari, A. & Kakde, O. G. Performance optimisation of deep learning models using majority voting algorithm for brain tumour classification. *Comput. Biol. Medicine* **169**, 107899, DOI: 10.1016/j.compbiomed.2024.107899 (2024).

24. van Leeuwen, K. G. *et al.* Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur. Radiol.* **31**, 3797–3809, DOI: 10.1007/s00330-021-07892-z (2021).

25. Nickparvar, M. Brain tumor MRI dataset. *Kaggle* https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset (2021).

26. Tan, M. & Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, 6105–6114 (2019).

27. Lin, T. Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988, DOI: 10.1109/ICCV.2017.324 (2017).

28. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

Acknowledgements

The authors acknowledge the Department of Computer Science and Engineering, Daffodil International University, for providing computational resources. We thank the creators of the Brain Tumor MRI Dataset for making their data publicly available for research purposes. We also acknowledge the developers of PyTorch, EfficientNet, and related open-source tools that made this research possible.

Author contributions statement

M.A. conceived the research direction, designed the methodology, and supervised the project. M.T.J.J. developed the software implementation, conducted all experiments, performed data analysis, and wrote the original manuscript draft. M.A.R.J. performed data curation, created visualizations, and contributed to manuscript preparation. M.N.I.I. contributed to experimental validation, literature review, and manuscript revision. All authors reviewed, edited, and approved the final manuscript.

Additional information

Competing interests: The authors declare no competing financial or non-financial interests.