Computer Methods and Programs in Biomedicine 282 (2026) 109409

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Calibrated ROI-gated conditional computation for high-throughput and backbone-agnostic brain tumor MRI classification

Ashraful Alam Nirob [a], Tasnim Sakib Apon [b], Anika Tahsin [a], Md. Golam Rabiul Alam [a], Sandra Costanzo [c], Giancarlo Fortino [c], Andrea Aliverti [d], Mohammad Mehedi Hassan [e,*]

[a] Department of Computer Science and Engineering, BRAC University, Kha 224 Pragati Sarani, Merul Badda, Dhaka, 1212, Bangladesh
[b] College of Liberal Arts and Sciences, University of Maine, Orono, 04469, ME, United States
[c] Department of Informatics, Modeling, Electronics, and Systems, University of Calabria, Rende, 87036, Italy
[d] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy
[e] Information Systems Department, College of Computer and Information Sciences, King Saud University, Riyadh, 11543, Saudi Arabia

ARTICLE INFO

Keywords:
Brain tumor MRI
Multi-class classification
ROI-gated spatial transformer
Conditional computation
Token pruning
Calibration (ECE)
Efficient inference

ABSTRACT

Background and Objective: Multi-class brain tumor classification from magnetic resonance imaging must achieve high diagnostic accuracy while maintaining low inference latency and reliable confidence for real-time clinical deployment. High-capacity deep learning models are computationally expensive, and naive acceleration may discard important diagnostic information and reduce confidence reliability. This study proposes a trainable, backbone-agnostic efficiency layer that reallocates computation toward diagnostically relevant regions while preserving global image context and confidence reliability.

Methods: The proposed framework learns differentiable spatial localization to identify candidate regions of interest and performs compute-controlled selection to retain informative regions for classification. This enables conditional computation without requiring external segmentation. The method was evaluated on a revised multi-source corpus constructed from three public brain tumor MRI datasets, including the Fernando Feltrin collection, the Nickparvar dataset, and the BRISC dataset, after harmonization of the classification images used in this study. Experiments were conducted across three magnetic resonance imaging modalities (T1, contrast-enhanced T1, and T2) and multiple convolutional neural network backbones using a controlled inference protocol with consistent hardware and batch settings.

Results: The proposed method achieved consistent efficiency improvements across modalities and backbones. Inference throughput increased by 2.3–5.7 times while maintaining classification accuracy comparable to strong baseline models. Calibration analysis showed that removing the calibration component increased expected calibration error from 0.0100 to 0.0535, indicating reduced confidence reliability. Visual analysis confirmed that the model consistently focused on clinically relevant tumor regions.

Conclusions: The proposed framework improves efficiency while preserving diagnostic accuracy and confidence reliability. By combining adaptive region selection, compute control, and calibration, the method enables fast and trustworthy brain tumor classification suitable for resource-constrained and real-time clinical environments.

1. Introduction

Brain tumor classification from MRI is central to neuro-oncology, yet scaling accurate multi-class decision support remains challenging due to substantial inter- and intra-tumor heterogeneity and the growing volume of imaging studies. Globally, brain and central nervous system cancers accounted for approximately 321,731 new cases and 248,500 deaths in 2022 [1]. Current clinical taxonomy is guided by the 2021 WHO Classification of Tumours of the Central Nervous System, which integrates histopathology with molecular markers and grades tumors using Arabic numerals (1–4) [2–4]. MRI remains the workhorse modality because of its superior soft-tissue contrast, with complementary sequences (T1, contrast-enhanced T1, and T2) supporting

* Corresponding author.
E-mail addresses: ashraful.alam5315@gmail.com (A.A. Nirob), tapon@uvm.edu (T.S. Apon), anika.tahsin5@g.bracu.ac.bd (A. Tahsin), rabiul.alam@bracu.ac.bd (M.G.R. Alam), sandra.costanzo@unical.it (S. Costanzo), giancarlo.fortino@unical.it (G. Fortino), andrea.aliverti@polimi.it (A. Aliverti), mmhassan@ksu.edu.sa (M.M. Hassan).

https://doi.org/10.1016/j.cmpb.2026.109409
Received 19 February 2026; Received in revised form 16 April 2026; Accepted 26 April 2026
Available online 30 April 2026

diagnosis, treatment planning, and prognosis estimation. However, the rapid growth of MRI workloads also makes computational throughput and deployment efficiency critical considerations for scalable clinical decision support.

Despite its clinical value, MRI-based tumor classification is difficult to operationalize at scale when deployment constraints are taken seriously. High-capacity deep models can impose substantial inference cost that limits throughput in routine pipelines, especially in settings requiring near real-time screening or high-volume triage. In practice, inference may run under heterogeneous compute environments (e.g., shared hospital servers, edge workstations, or cloud backends), where latency and resource variability further constrain deployable model choices. Moreover, clinical use often depends on probability scores (e.g., for triage, selective review, or threshold-based decision support), so confidence reliability must be assessed alongside accuracy; modern deep networks can be systematically miscalibrated and overconfident even when top-1 performance is strong, which is unsafe for triage-oriented workflows [5–7]. These issues become more pronounced under conditional computation, where selectively reducing evidence can shift confidence distributions relative to fully processed baselines. These constraints motivate methods that improve efficiency without compromising diagnostic discrimination or probability reliability.

Motivated by these requirements, this work targets multi-class brain tumor classification from MRI using an ROI-guided conditional-computation pipeline that explicitly controls inference cost while retaining global context. The framework operates as a content-adaptive inference engine: for each image it predicts a small set of candidate regions likely to contain diagnostically relevant evidence, while always preserving a global view of the full slice. Because most MRI slices contain large background regions with limited diagnostic value, focusing computation on lesion-relevant content provides a principled pathway to reduce wasted processing while preserving tumor evidence. A differentiable gating mechanism learns to balance local lesion-centric features against global context, dynamically suppressing background-dominated computation so that most processing is spent only where it contributes to discrimination. The proposed design is backbone-agnostic: crops are encoded by a shared pretrained CNN backbone (instantiated as EfficientNet-B0 in the main implementation and validated across multiple SOTA backbones in experiments), and the gated aggregation produces a compact representation for final classification.

Training optimizes classification performance jointly with two deployment-facing constraints. For conditional computation, ROI gates are trained using a Concrete (Gumbel–Sigmoid) relaxation and an expected ROI-count regularizer that penalizes deviation between the expected number of active ROIs and a target compute budget. For confidence reliability, the objective incorporates a differentiable calibration-aware constraint based on soft binning of confidence, explicitly counteracting the tendency of pruned and conditional-computation models to become overconfident under reduced evidence. At inference, the framework supports a full mode that processes the global crop and all K ROI crops, and a pruned mode that processes the global crop together with only the selected ROIs. In the pruned mode, ROIs are chosen by hard top-$K_{\max}$ selection on gate probabilities with optional thresholding; when $K_{\max} = 0$, inference reduces to a global-only path. This design enables direct benchmarking of predictive metrics together with latency (ms/image) and throughput (images/s) under explicitly controlled computation.

This paper evaluates an ROI-guided classification framework with conditional computation for multi-class brain tumor MRI classification, reporting predictive performance, calibration behavior, and inference efficiency. The primary contributions are as follows:

- We introduce a *trainable, backbone-agnostic* ROI-guided conditional-computation framework that always retains a global crop while learning K candidate ROIs per image via differentiable grid-sampling.
- We couple ROI selection to prediction through a differentiable gating-and-aggregation mechanism that prioritizes lesion-relevant evidence while preserving global context, enabling consistent training-to-inference behavior under content-adaptive computation.
- We explicitly control compute using Concrete (Gumbel–Sigmoid) ROI gating with an expected ROI-count regularizer targeting a user-specified compute level and support pruning-aware inference via hard top-$K_{\max}$ selection with optional thresholding (including a global-only path when $K_{\max} = 0$).
- We incorporate a differentiable calibration-aware constraint to maintain reliable probabilities under conditional computation and evaluate confidence reliability using calibration metrics (ECE) alongside discrimination metrics under both full and pruned inference.
- Across T1, T1CE/T1C+, and T2 modalities and heterogeneous backbones, we demonstrate consistent accuracy–efficiency pareto improvements, achieving 2.3×–5.7× throughput gains with no statistically significant accuracy penalty ($p > 0.05$) and statistically significant efficiency gains ($p < 0.001$).

The remainder of this paper is organized as follows. Section 2 reviews prior work on brain tumor classification, attention mechanisms, and efficient inference strategies. In particular, we emphasize systems-aware efficiency and resource-constrained deployment considerations that motivate conditional computation. Section 3 details the proposed methodology, including the dataset characteristics, the ROI parameterization and gating formulation, the shared backbone instantiation, and the calibration-aware training objective. Section 4 presents the experimental analysis, defining the evaluation metrics and reporting quantitative results across T1, T1C+, and T2 modalities with a focus on discrimination, reliability, and computational efficiency. Section 5 discusses the implications of the results, analyzing modality-specific behavior, ROI interpretability, statistical significance, and limitations relevant to real-world deployment. Finally, Section 6 summarizes the contributions and concludes the study.

2. Related work

Brain tumor recognition from medical imaging has been studied across classical machine learning pipelines and modern deep learning systems, using MRI sequences (including T1-weighted contrast-enhanced) and CT depending on the clinical objective. Early machine learning pipelines typically followed a staged design—ROI localization, handcrafted feature extraction, optional feature selection, and a final classifier. Zacharaki et al. [8] exemplified this paradigm for multi-type tumor categorization on MRI. While such approaches can be effective under carefully engineered ROIs and features, they are sensitive to design choices and offer limited leverage over end-to-end optimization and computational efficiency.

Deep learning models reduce reliance on handcrafted features by learning discriminative representations directly from images, which has led to strong performance in MRI tumor classification. Transfer learning with pretrained CNN backbones is widely adopted for multi-class settings; Deepak et al. [9] modified GoogleNet and reported strong accuracy on the figshare dataset, illustrating the practical utility of pretrained visual representations when labeled medical data are limited. In parallel, ROI- and patch-centric strategies remain common: Cheng et al. [10] augmented tumor regions and evaluated feature representations over subregions, reinforcing that focusing computation on diagnostically relevant areas can improve separability even when the downstream classifier is unchanged. However, many ROI-centric pipelines either assume external segmentation or employ handcrafted ROI heuristics, which can decouple localization from classification and limit robustness under tumor heterogeneity (e.g., variable scale, multifocal patterns) and acquisition variability. Consequently, weakly

supervised and end-to-end region selection is increasingly adopted to couple localization and recognition within a single objective, aiming to reduce unnecessary computation on background-dominated tissue while preserving tumor evidence.

Recent studies on brain tumor MRI segmentation studies have further strengthened the role of lesion-focused modeling by combining encoder–decoder architectures with attention and transformer-based context aggregation. Nguyen-Tat et al. proposed a hybrid 3D UNet framework that integrates contextual transformer and double-attention modules for improved tumor delineation in MRI images [11]. More recently, MaxViT-based segmentation designs have also been explored for brain tumor analysis; DoubleBlock-ViT combines a 3D transformer encoder with dual-path skip connections to preserve both global context and fine spatial structure during segmentation [12]. In addition, QMaxViT-Unet+ extends this direction to scribble-supervised medical image segmentation, showing that reduced-annotation training can still support effective boundary-aware dense prediction [13]. These studies are relevant because they highlight the value of spatially focused and attention-enhanced modeling for tumor analysis. However, they primarily target dense segmentation and generally depend on pixel-level or scribble-level supervision, whereas the present work addresses backbone-agnostic and compute-controlled multi-class classification without requiring external segmentation annotations.

This efficiency requirement is also reflected in systems research that targets practical deployment constraints for deep models. Recent works have explored hierarchical coarse-to-fine inference for edge environments [14] and end–edge–cloud collaborative inference with latency-aware partitioning [15], highlighting that adaptive computation is often necessary when inference resources are limited or variable. Complementarily, efficiency-preserving learning under distributed constraints (e.g., incremental/federated training with knowledge distillation) has been studied for medical image classification [16], indicating that both training- and inference-time constraints can shape the design space of deployable medical AI. Hardware-level acceleration further reinforces this motivation; dynamic FPGA reconfiguration has been used to scale embedded CNN inference under tight compute budgets [17]. Together, these lines of work motivate algorithmic approaches that achieve input-adaptive efficiency while maintaining clinically meaningful accuracy.

To operationalize these efficiency goals within a deep learning architecture, differentiable localization mechanisms such as Spatial Transformer Networks (STN) enable end-to-end learning of spatial manipulation (e.g., cropping and warping) within the network [18]. When multiple regions (or instances) are extracted per image, the final prediction typically depends on an aggregation operator over region features. Multiple-instance learning (MIL) provides a standard formulation for this setting, and attention-based MIL introduces learnable instance weighting to form a bag-level representation [19]. While STN/MIL-style formulations provide a principled interface for learning region representations without dense annotation, they do not, by themselves, address *compute control*: without explicit mechanisms for selection or early exiting, multi-region pipelines may still incur substantial cost, and aggregation can be sensitive to noisy or redundant crops.

For models that must select a subset of regions or computations, continuous relaxations of discrete decisions are commonly used. The Concrete distribution [20] and the Gumbel-Softmax estimator [21] provide differentiable approximations for sampling-based gating. Sparsity and conditional computation can also be trained using stochastic gates with differentiable expected objectives; Louizos et al. [22] introduced hard-concrete gating for optimizing an expected L_0 objective, which is closely related to compute-controlled selection. Beyond region selection, dynamic inference methods allocate computation adaptively per input: Graves [23] introduced adaptive computation time to vary the number of computational steps, and BlockDrop [24] demonstrated dynamic layer execution to reduce inference cost while maintaining accuracy. Related ideas have been explored in token- or patch-based selection for efficiency, including DynamicViT [25] and TokenLearner [26], where only a subset of informative tokens is retained for downstream processing. Nevertheless, most dynamic inference and token pruning methods are primarily optimized for 2D natural-image benchmarks; in medical imaging, aggressive evidence reduction can be risky when lesions are small, heterogeneous, or sparsely represented. This motivates reporting not only classification metrics but also efficiency indicators such as latency (ms/image) and throughput (images/s) when proposing cost-aware medical classifiers.

Accordingly, reliable deployment requires not only accurate predictions but also calibrated confidence estimates. Temperature scaling and related post-hoc calibration procedures are widely used baselines for improving probability calibration in neural networks [5]. Recent work also studies calibration objectives that are differentiable and can be optimized during training [27,28], and calibration has been analyzed explicitly in medical imaging classification settings where miscalibrated confidence can affect thresholded decision rules [29]. This consideration becomes especially salient for conditional computation: when the model adaptively drops regions, layers, or tokens, confidence estimates may shift relative to fully processed baselines, motivating explicit calibration-aware design and evaluation. Backbone design further mediates the accuracy–efficiency trade-off; EfficientNet [30] provides a modern convolutional family spanning efficient to large-scale regimes and is commonly used as a strong feature extractor in imaging pipelines, making it a practical reference point for assessing backbone-agnostic efficiency layers.

Summary and gap. Taken together, prior literature establishes (i) ROI-focused strategies that emphasize diagnostically relevant computation [10], (ii) differentiable localization and multi-instance aggregation that enable weakly supervised region modeling [18,19], (iii) continuous relaxations and stochastic gating that make discrete selection trainable [20–22], and (iv) dynamic inference mechanisms that reduce cost via input-adaptive execution [23–26]. However, ROI modeling and compute control are often treated separately, leaving localization and conditional computation weakly coupled. Moreover, efficiency-oriented selection yields input-dependent inference paths and selectively discards evidence, which can destabilize predictive confidence relative to fully processed baselines. Yet, calibration is rarely treated as a first-class constraint alongside throughput and latency. These gaps motivate a unified approach that jointly learns region selection and conditional computation, targeting an accuracy–efficiency–reliability Pareto frontier under resource-constrained medical deployment.

3. Methodology

This section presents the proposed ROI-gated conditional-computation framework for multi-class brain tumor MRI classification. We first describe the dataset and notation, then detail the ROI parameterization, probabilistic gating, and STN-based cropping with a shared backbone and gated attention aggregation. Finally, we define the training objective with ROI-budget and calibration constraints and describe the pruned inference procedure for efficient, confidence-preserving prediction.

3.1. Data description

We have employed three different types of MRI images to classify brain tumors accumulated and curated by Fernando Feltrin [31]. This dataset consists of 15 different classes of brain tumors reported by radiologists in actual patients, ranging from the most common to irregular tumor types and the most frequent to rare conditions. In parallel with the normal brain stage images, other classes are astrocytoma, carcinoma, ependymoma, ganglioglioma, germinoma, glioblastoma, granuloma, medulloblastoma, meningioma, neurocytoma, oligodendroglioma, papilloma, schwannoma, and tuberculoma. In common with many distinct classes, this dataset provides three different types of MRI images that no other dataset can provide.

Table 1
Summary of the three public datasets used in the revised study. The original manuscript used a curated 15-class subset of the Fernando Feltrin collection, while the revised version additionally incorporates the classification images from the Nickparvar and BRISC datasets.

Source	MRI type	Label space	Images used	Use in this study
Fernando Feltrin [31]	T1, T1C+, T2	44 classes at source; 15 classes retained in this paper	4479	Base dataset used in the original manuscript
Nickparvar [32]	Brain MRI	4 classes: glioma, meningioma, no tumor, pituitary	7022	Added classification images after label harmonization
BRISC [33]	T1	4 classes: glioma, meningioma, pituitary, no tumor	6000	Added classification images only; segmentation masks were not used

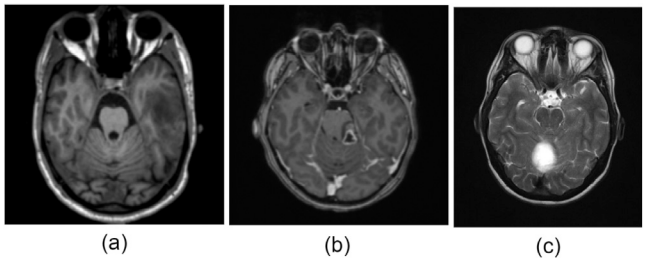

Fig. 1. Sample images from various types of images. Here, (a) is T1 or normal images; (b) is T1C+, contrast-enhanced T1, and finally, (c) is T2 or magnetic resonance images (MRI).

We expanded the dataset by combining the curated 15-class subset [31] with additional public classification images from the Nickparvar brain tumor MRI dataset [32] and the BRISC annotated dataset [33]. The first source forms the base dataset used in the original manuscript and provides multi-modal MRI slices spanning T1, contrast-enhanced T1, and T2 images, whereas the Nickparvar and BRISC sources provide additional four-class classification images. Because the three sources do not share an identical label space, direct merging was performed only for semantically matched classes after label harmonization. In particular, normal and no-tumor images were unified into a single normal category, and meningioma images were merged directly across sources. By contrast, the glioma label in the Nickparvar and BRISC datasets was not treated as interchangeable with the finer-grained astrocytoma, glioblastoma, and oligodendroglioma categories of the original 15-class set, and the pituitary category was not included in the final 15-class merged corpus. The BRISC segmentation masks were not used in the present study; only the classification images were retained. To provide a concise overview of the revised data construction process, Table 1 summarizes the three public sources used in this study and clarifies how each source contributes to the final harmonized corpus (see Fig. 1).

Three types of MRI images provided by Fernando Feltrin datasets are:

- T1 : Weighted images
- T1C+ : Contrast-enhanced T1
- T2 : Weighted images

These are characterized by relaxation time. T1-weighted images are produced by using short TE and TR times. TE refers to Time to Echo, which is the time between the delivery of the RF pulse and the receipt of the echo signal, and TR refers to Repetition Time, which is the amount of time between successive pulse sequences applied to the same slice. The time constant T1 (longitudinal relaxation time) represents how quickly excited protons return to equilibrium. It measures the time the spinning protons take to reorient with the externally applied magnetic field. T2-weighted images are produced by using long TE and TR times. The time constant T2 is a measurement of the dephasing of transverse magnetization. T1 C+ is the T1 with enhanced contrasted images by infusing Gadolinium (Gad), a non-toxic paramagnetic contrast enhancement agent. These images are instrumental in looking at vascular structures and breakdowns in the blood–brain barrier [e.g., tumors, abscesses, and inflammation (herpes simplex encephalitis, multiple sclerosis, and similar conditions)].

Table 2
Magnetic Resonance Imaging (MRI) Basics.

Tissue	T1-Weighted	T2-Weighted
CSF	Dark	Bright
White Matter	Light	Dark Gray
Context	Gray	Light Gray
Fat (within bone marrow)	Bright	Light
Inflammation (infection, demyelination)	Dark	Bright

MRI provides exquisite details of the brain, spinal cord, and vascular anatomy. We have opted for MRI images over alternatives due to their ability to detect flowing blood and enigmatic vascular malformations, which is an advantage over CT. It can also detect demyelinating disease and has no beam-hardening artifacts, such as those that can be seen with CT. Thus, the posterior fossa is more easily visualized on MRI than CT. Furthermore, imaging is executed in the absence of ionizing radiation (see Fig. 2) (see Table 2).

In T1 images, the fat regions come out bright, but the Cerebrospinal Fluid (CSF) is dark because fat protons realign early and CSF protons realign late. But the scenario is different in T2 weighted images, CSF being the bright region and fatrich tissues are on the darker side [34]. As some tumors behave differently under contrast, these variations of images extended the classification range in a broader and more significant way in diagnosing brain tumors. The dataset contains around 4479 brain images with two views, Axial and Coronal.

The images are presented in three file formats: JPEG (3212 images), JPG (1266 images), and WebP (1 image). The majority of the images have dimensions of 630 × 630 and have a maximum file size of 50 KB. The majority of the class possesses over a hundred images of brain tumors. The following section will enumerate the distribution of 4479 images among the fourteen divisions (see Table 3).

Astrocytoma is one of the most common brain tumors that develop in the Central Nervous System (CNS). It can be cancerous (benign) and noncancerous (malignant), and according to WHO (World Health Organization), it comprises four grades ranging from Grades I (benign) to Grades IV (malignant). Children get affected most by Grade I; on the other hand, adults get affected most by Grade IV (around 24%) [35].

Another most common tumor is carcinoma, making up 80% to 90% of cancer diagnoses, and it starts to form in epithelial tissue. It can form

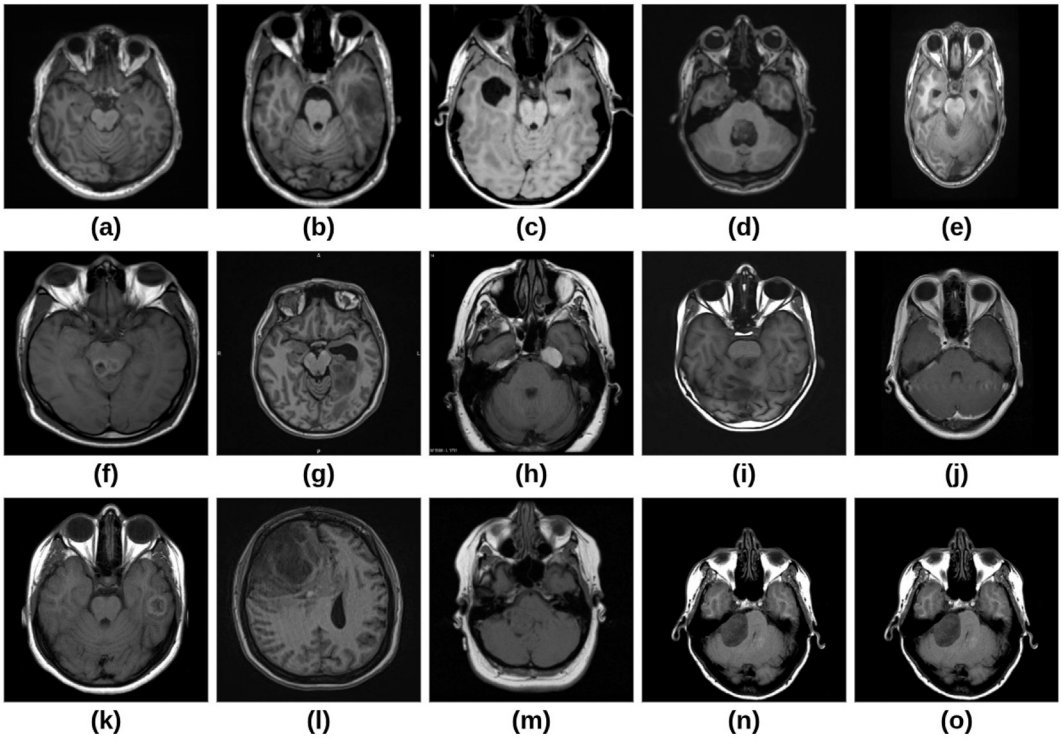

Fig. 2. Sample Images from various classes. Here, (a) is Normal, (b) is Astrocytoma, (c) is Carcinoma, (d) is Ependimoma, (e) is Ganglioglioma, (f) is Germinoma, (g) is Glioblastoma, (h) is Granuloma, (i) is Medulloblastoma, (j) is Meningioma, (k) is Neurocitoma, (l) is Oligodendroglioma, (m) is Papilloma, (n) is Schwannoma and finally (o) is Tuberculoma.

Table 3
Brain tumor dataset distribution among the classes of the tumor.

Class	T1	T1C+	T2
Normal	251	0	271
Astrocitoma	176	233	171
Carcinoma	66	112	73
Ependimoma	45	48	57
Ganglioglioma	20	18	23
Germinoma	27	40	33
Glioblastoma	55	94	55
Granuloma	30	31	17
Meduloblastoma	23	67	41
Meningioma	272	369	233
Neurocitoma	130	223	104
Oligodendroglioma	86	72	66
Papiloma	66	108	63
Schwannoma	148	194	123
Tuberculoma	28	84	33

in different parts of the body. Carcinoma is rare in children but risky for older people(around 65 or older). It spreads in five stages, and the stage is called the metastatic stage [36].

Ependymoma tumor has three grades in its period, Grade I and Grade II, called low grade, which grows slowly in the brain. Both children and adults can be affected. The last Grade is a malignant stage, which means cancerous. As it is a CNS tumor, it rarely spreads outside CNS and has a chance of 83.3% survival rate for 5 years for ependymoma [37].

Ganglioglioma is a low-grade tumor of mixed cell types, and it is a sporadic brain tumor. Children with specific genetic syndromes are at higher risk of developing this tumor [38].

Germinoma brain tumors appear primarily in teenagers and young people. Although it is a cancerous tumor, it forms in the human brain very rarely. It accounts for 3.8% of primary brain tumors in children and adolescents [39].

Glioblastoma tumors are malignant grade IV tumors. It penetrates neighboring brain regions and infiltrates them diffusely and commonly arises as de novo, which means to begin as a grade IV tumor with no evidence. It represents around 14% of brain tumors and is slightly more common in men than women. More than 12,000 glioblastoma cases are diagnosed each year in USA [40].

Granuloma is not cancerous, and it is a benign type of tumor. It forms when cells of the human immune system (macrophages) are not able to destroy something they see as dangerous. It is not that common and rare case in general. However, sometimes granuloma can be found in people with other cancers like skin lymphomas [41].

Another primary central nervous system (CNS) tumor is medulloblastoma, which develops mainly in the brain and spine and also forms in the cerebellum. It is a grade IV tumor, which means it is cancerous. It primarily affects children, more common in men than women and in white and Hispanic people. The Five-year survival rate for medulloblastoma is around 72.1% depending on some factors [42].

Meningioma is one of the most common brain cancers that form in the spine and brain. It has four grade stages, and among them, the last Grade (malignant) is rare in people. Meningiomas are more common in females, but grades II and III occur more often in males. They are most common in black people, followed by white people, and then Asian-Pacific Islanders [43].

Neurocytoma and Papilloma brain tumors are rare, and both are benign brain tumors. Neurocytoma comprises less than 1% of brain tumors, and it occurs in ventricles, which are the fluid-filled spaces within the brain [44]. On the other hand, maximum papillomas are caused by low-risk HPV forms and are common in young children [45].

Oligodendroglioma tumor is a growth of cells that begins in cells called oligodendrocytes. It is most common in adults and responsible for about 3% to 4% of all brain tumors. Up to 80% of people with oligodendroglioma will have a seizure because of this cancer, and its grade III stages are metastatic (cancerous) [46].

Schwannoma is another rare brain tumor, which is also called acoustic neuroma. It is noncancerous and begins in the nerve that

connects the brain with the ear. Around 8 out of 100 people are affected by this tumor, and the risk increases for older people [47].

Tuberculoma, the final sample of the dataset, is another rare, non-cancerous brain lesion caused by Mycobacterium tuberculosis infection. This tumor-like mass, often located in the CNS (usually the brain or spinal cord), can lead to severe neurological symptoms depending on its size and location. It mainly affects those in regions with high tuberculosis prevalence, with children and immunocompromised individuals being particularly vulnerable. Though rare, it can cause complications if left untreated, like increased intracranial pressure or seizures, and accounts for about 5%–10% of intracranial tuberculosis cases in TB-endemic areas.

Data splitting and leakage control. All experiments were conducted using fixed train, validation, and test directories prepared prior to model training. For datasets with source-defined partitions, we preserved the official split structure. In particular, the Nickparvar dataset was used according to its provided training/testing partition, in which approximately 22% of images are reserved for testing, and the BRISC dataset was used according to its published clean stratified split (5000 training slices and 1000 test slices) [32,33]. For the remaining images, the split was performed at the slice level and implemented through mutually exclusive train/validation/test folders before loading. No image from the validation or test folders was used during optimization, augmentation, or checkpoint selection. Because patient identifiers were not consistently available to our split-construction pipeline across all merged public sources, strict patient-level separation could not be enforced uniformly; therefore, the reported results should be interpreted as slice-level generalization rather than strict patient-level generalization.

3.2. Problem formulation and notation

Let an input MRI slice be denoted by $\mathbf{x} \in [0,1]^{H\times W\times 3}$ with class label $y \in \{1,\dots,C\}$. Our model predicts (i) a class posterior over C tumor categories and (ii) a calibrated confidence score derived from the posterior. The method constructs a *set* of $K{+}1$ image crops per input: one *global* crop and K *learned* region-of-interest (ROI) crops. We index crops by $k \in \{0,1,\dots,K\}$ where $k=0$ denotes the global crop and $k \geq 1$ denotes ROIs.

3.3. ROI parameterization and probabilistic gating

Given $\mathbf{x}$, a lightweight ROI predictor outputs K triplets of crop parameters and K gating logits. For each ROI $k \in \{1,\dots,K\}$, the predictor emits an unconstrained vector $\mathbf{r}_k = (r_k^{(x)}, r_k^{(y)}, r_k^{(s)})$ and a gate logit a_k. We map these outputs to valid crop centers and scales using bounded parameterizations:

$$\begin{aligned} c_{x,k} &= \sigma\left(r_k^{(x)}\right), \quad c_{y,k} = \sigma\left(r_k^{(y)}\right), \\ s_k &= s_{\min} + \left(s_{\max} - s_{\min}\right)\sigma\left(r_k^{(s)}\right), \end{aligned} \tag{1}$$

where $\sigma(\cdot)$ is the logistic sigmoid, $(c_{x,k}, c_{y,k}) \in (0,1)^2$ is the ROI center in normalized image coordinates, and $s_k \in [s_{\min}, s_{\max}]$ controls the crop extent. Eq. (1) ensures each ROI is always well-defined and remains within a controlled scale range.

We convert each gate logit a_k into an ROI retention probability p_k and define the expected number of selected ROIs:

$$p_k = \sigma(a_k), \qquad \mathbb{E}[K] = \sum_{k=1}^{K} p_k. \tag{2}$$

Eq. (2) provides a differentiable estimate of how many ROIs the model intends to use for an input batch, which we later constrain via a budget loss.

During training, we require discrete-like gating while preserving gradients; therefore we use a Concrete (Gumbel-Sigmoid) relaxation:

$$g_k = \sigma\left(\frac{a_k + \log u_k - \log(1-u_k)}{T_g}\right), \qquad u_k \sim \mathcal{U}(\varepsilon, 1-\varepsilon), \tag{3}$$

where $T_g > 0$ is the gate temperature and ε prevents numerical overflow. Eq. (3) yields stochastic gates $g_k \in (0,1)$ that approach binary decisions as $T_g \to 0$. At evaluation time, we set $g_k = p_k$ (deterministic gating) to eliminate sampling noise (see Fig. 3).

Finally, we define the global crop as a fixed "always-on" element:

$$(c_{x,0}, c_{y,0}, s_0, g_0) = (0.5, 0.5, 1.0, 1.0), \tag{4}$$

so crop $k=0$ always represents the full-image context and is never pruned (Eq. (4)).

3.4. STN cropping and shared backbone feature extraction

We extract crops using a Spatial Transformer Network (STN) sampler with a differentiable grid. Let $(u,v) \in [-1,1]^2$ be canonical coordinates on a $P\times P$ crop grid. For crop k, we map grid points to normalized image coordinates:

$$\begin{bmatrix} x_k(u,v) \\ y_k(u,v) \end{bmatrix} = \begin{bmatrix} (2c_{x,k}-1) \\ (2c_{y,k}-1) \end{bmatrix} + s_k \begin{bmatrix} u \\ v \end{bmatrix}. \tag{5}$$

Eq. (5) defines a translation by the ROI center and an isotropic scaling by s_k, producing a continuous sampling grid. The crop $\mathbf{z}_k \in \mathbb{R}^{P\times P\times 3}$ is then obtained via bilinear sampling of $\mathbf{x}$ on this grid (standard STN sampling).

Each crop is passed through a single shared feature extractor (backbone) to produce an embedding vector:

$$\mathbf{f}_k = \phi(\mathbf{z}_k) \in \mathbb{R}^d, \qquad k \in \{0,1,\dots,K\}, \tag{6}$$

where $\phi(\cdot)$ is the backbone network and d is the embedding dimension. Eq. (6) enforces parameter sharing across global and ROI crops, ensuring that ROI specialization arises from *where* we sample, not from separate backbones.

3.5. Gated patch attention and classification

After computing $\{\mathbf{f}_k\}_{k=0}^{K}$, we aggregate them with an attention mechanism that is explicitly modulated by the gates. We first compute a scalar compatibility score for each crop using a learned projection vector $\mathbf{w} \in \mathbb{R}^d$ and bias b:

$$s_k = \mathbf{w}^\top \mathbf{f}_k + b. \tag{7}$$

Eq. (7) assigns a raw importance logit to each crop embedding.

We then incorporate gating by adding $\log(\tilde{g}_k)$ to the attention score, where $\tilde{g}_k = \max(g_k, \varepsilon)$:

$$\alpha_k = \frac{\exp\left(s_k + \log(\tilde{g}_k)\right)}{\sum_{j=0}^{K} \exp\left(s_j + \log(\tilde{g}_j)\right)}. \tag{8}$$

Eq. (8) ensures that a suppressed ROI ($g_k \approx 0$) receives negligible attention mass even if s_k is large.

The final pooled representation is the attention-weighted sum:

$$\mathbf{h} = \sum_{k=0}^{K} \alpha_k \mathbf{f}_k. \tag{9}$$

Eq. (9) fuses global context and discriminative ROIs into a single vector.

Finally, we produce classification logits:

$$\boldsymbol{\ell} = \mathbf{W}_c \mathbf{h} + \mathbf{b}_c \in \mathbb{R}^C, \tag{10}$$

where $(\mathbf{W}_c, \mathbf{b}_c)$ are learned classifier parameters. Eq. (10) defines the only class-predictive head used in both training and inference.

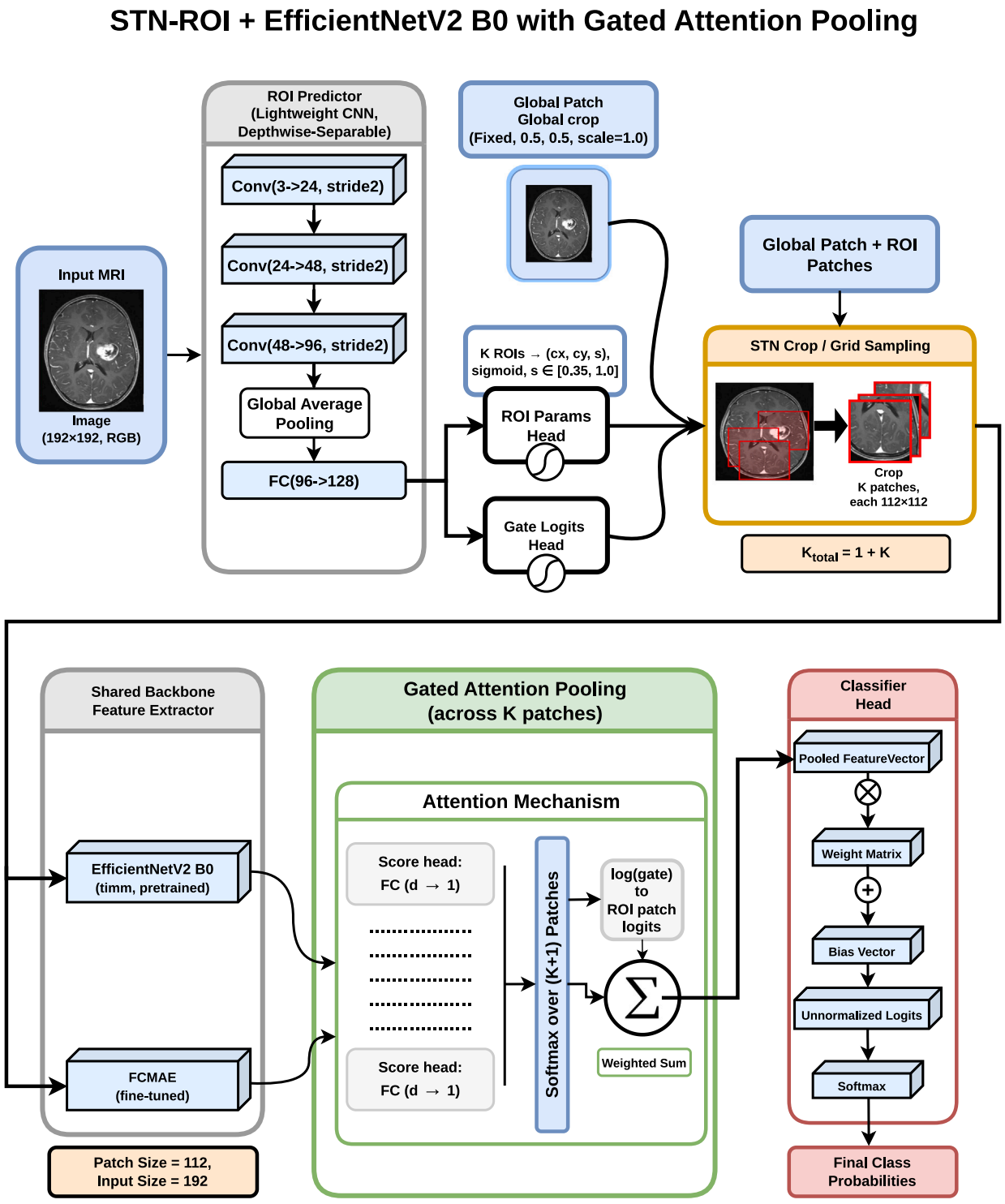

Fig. 3. Overview of the proposed STN-ROI + EfficientNet-B0 framework with gated attention pooling for brain MRI classification. A lightweight depthwise-separable ROI predictor estimates K ROI parameters (c_x, c_y, s) and corresponding gate logits, while a fixed global crop is always included, yielding $K_{\text{total}} = K + 1$ patches (global + ROIs). All patches share a single EfficientNet-B0 backbone (FCMAE fine-tuned) to extract features, which are then aggregated via gated attention: patch-level attention scores are modulated by $\log(\text{gate})$, normalized with a softmax across patches, and used to compute a weighted sum of patch features. The pooled representation is passed to a linear classifier to produce final class logits; a budget loss encourages the expected number of active ROIs to match the target K, and an optional Soft-ECE regularizer improves calibration during training.

3.6. Training objective: Accuracy, ROI budget, and calibration constraint

Let $\mathbf{q} = \mathrm{softmax}(\boldsymbol{\ell})$ be the predicted posterior. We train with a cross-entropy objective (optionally with label smoothing):

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{c=1}^{C} \tilde{y}_c \log q_c, \tag{11}$$

where $\tilde{\mathbf{y}}$ is the (possibly smoothed) one-hot target. Eq. (11) directly optimizes classification performance.

To enforce a controllable compute budget, we penalize deviation between the expected ROI count and a target value K_{tgt}:

$$\mathcal{L}_{\mathrm{bud}} = \lambda_{\mathrm{bud}} \left(\mathbb{E}[K] - K_{\mathrm{tgt}}\right)^2, \tag{12}$$

where $\mathbb{E}[K]$ is defined in Eq. (2). Eq. (12) is the mechanism that pushes the gating network toward a desired average number of ROIs.

We additionally incorporate a differentiable calibration-aware constraint using a Soft-ECE-style regularizer. For each sample i in a mini-batch, define the confidence (maximum predicted probability) and the true-class probability:

$$c_i = \max_c \ \mathrm{softmax}\left(\frac{\boldsymbol{\ell}_i}{T_{\mathrm{ece}}}\right)_c, \qquad t_i = \mathrm{softmax}\left(\frac{\boldsymbol{\ell}_i}{T_{\mathrm{ece}}}\right)_{y_i}, \tag{13}$$

where $T_{\mathrm{ece}} > 0$ is used only inside this regularizer. Eq. (13) makes both c_i and t_i differentiable functions of logits.

We compute soft bin memberships over B confidence bins with edges $\{e_b\}_{b=0}^{B}$, using a smoothness parameter δ:

$$w_{i,b} = \sigma\left(\frac{c_i - e_{b-1}}{\delta}\right) - \sigma\left(\frac{c_i - e_b}{\delta}\right). \tag{14}$$

Eq. (14) assigns each sample a *soft* membership across bins, replacing hard binning by a differentiable relaxation.

For each bin b, define the (soft) average confidence and (soft) average true-class probability:

$$\bar{c}_b = \frac{\sum_i w_{i,b} c_i}{\sum_i w_{i,b}}, \qquad \bar{t}_b = \frac{\sum_i w_{i,b} t_i}{\sum_i w_{i,b}}. \tag{15}$$

Eq. (15) yields per-bin quantities that remain differentiable with respect to logits through $w_{i,b}$.

The Soft-ECE penalty is then:

$$\mathcal{L}_{\mathrm{ece}} = \sum_{b=1}^{B} \left(\frac{\sum_i w_{i,b}}{\sum_{b'} \sum_i w_{i,b'}}\right) \left(\bar{c}_b - \bar{t}_b\right)^2. \tag{16}$$

Eq. (16) penalizes systematic gaps between confidence and "soft accuracy proxy" (t_i) within each confidence region.

The final training objective is the weighted sum:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{bud}} + \lambda_{\mathrm{ece}} \mathcal{L}_{\mathrm{ece}}. \tag{17}$$

Eq. (17) explicitly separates (i) classification fit, (ii) ROI usage control, and (iii) calibration pressure.

3.7. Pruned inference for efficient confidence-preserving prediction

At inference time, we optionally restrict computation to at most $K_{\max}$ ROIs. We first compute p_k from Eq. (2) and keep the top-$K_{\max}$ ROIs by probability. We then apply a retention threshold τ to suppress low-utility ROIs:

$$m_k = \mathbb{I}[p_k > \tau], \qquad \hat{g}_k = p_k \cdot m_k. \tag{18}$$

Eq. (18) defines a deterministic pruning mask and the final inference-time gate value $\hat{g}_k$. The attention fusion remains unchanged except that g_k in Eq. (8) is replaced by $\hat{g}_k$, which preserves the same gated-attention semantics while skipping feature extraction for pruned ROIs.

Methodological distinction from token-pruning methods. Although the proposed framework also performs input-adaptive computation reduction, it differs fundamentally from transformer token-pruning methods such as DynamicViT and TokenLearner. DynamicViT prunes tokens *inside* a vision transformer by using intermediate token features to predict token importance and by applying differentiable attention masking across transformer layers [25]. TokenLearner likewise reduces the token set that is processed by subsequent transformer modules by learning adaptive tokens for ViT-style architectures [26]. In contrast, the proposed framework performs compute control in image space before backbone encoding: a lightweight ROI predictor operates directly on the input image, STN-based sampling extracts a global crop together with a small set of candidate ROIs, and only those retained crops are encoded by the shared backbone. Consequently, the efficiency layer does not depend on transformer blocks, self-attention matrices, or internal token representations, and can therefore be coupled with heterogeneous backbones, including convolutional and transformer encoders, through the same crop-encode-aggregate interface. This design also preserves backbone-specific input conventions: CNN backbones can be evaluated under the fixed resized setting used in this research, whereas ViT-based backbones can be used at their native image size because the ROI mechanism operates before backbone-specific tokenization or feature extraction. In this sense, the proposed method is a backbone-agnostic spatial compute reallocation mechanism rather than an architecture-specific token sparsification module. Moreover, the expected-ROI budget constraint and the global-only inference mode provide direct control over the amount of spatial evidence processed at test time.

4. Experimental analysis

This section describes the experimental setup and evaluation methodology used to assess the proposed framework across MRI modalities and backbones. We report discrimination, calibration, and efficiency metrics and present qualitative and quantitative results to characterize the accuracy–efficiency–reliability trade-off.

4.1. Evaluation metrics

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be the evaluation set with $y_i \in \{1, \ldots, C\}$. The classifier outputs a probability vector $\hat{\mathbf{p}}_i \in [0,1]^C$, and the predicted class is $\hat{y}_i = \arg\max_c \hat{p}_{i,c}$. We report three metric groups. First, classification metrics quantify discriminative performance under multi-class imbalance by including macro-averaged scores (each class receives equal weight). Second, probabilistic metrics assess whether predicted confidences are reliable, which is critical when probabilities are used for triage or downstream decision support. Third, efficiency metrics summarize inference cost using latency and throughput.

4.1.1. Classification performance metrics

Accuracy measures overall correctness:

$$\mathrm{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\hat{y}_i = y_i]. \tag{19}$$

To avoid performance being dominated by frequent classes, we report macro-averaged precision and recall. Let TP_c, FP_c, and FN_c

Algorithm 1: Training and Pruned Inference of ROI-Gated Classifier

Data: Image set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$
Hyperparams: ROI count K, crop size P, target ROI count K_{tgt}, gate temperature T_g
Hyperparams: Weights $\lambda_{\text{bud}}, \lambda_{\text{ece}}$, pruning parameters $K_{\max}, \tau$ (inference only)

Initialize: ROI predictor ρ, STN sampler S, shared backbone ϕ, token scorer ψ, classifier head θ_c

/* Training Phase */
for *each minibatch* $(\mathbf{x}, y) \in D$ **do**
  // 1) ROI parameters and gates
  Predict $\{(c_{x,k}, c_{y,k}, s_k)\}_{k=1}^{K}$ and gate logits $\{a_k\}_{k=1}^{K}$ using $\rho(\mathbf{x})$ // Eq. (1)
  Compute ROI probabilities $p_k = \sigma(a_k)$ and $\mathbb{E}[K] = \sum_{k=1}^{K} p_k$ // Eq. (2)
  Sample stochastic training gates g_k via Concrete relaxation with temperature T_g // Eq. (3)
  Set global crop gate $g_0 = 1$ and global crop parameters fixed // Eq. (4)
  // 2) Differentiable cropping and feature extraction
  Extract crops $\mathbf{z}_k = S(\mathbf{x}; c_{x,k}, c_{y,k}, s_k, P)$ for $k = 0, \ldots, K$ // Eq. (5)
  Encode features $\mathbf{f}_k = \phi(\mathbf{z}_k)$ for $k = 0, \ldots, K$ // Eq. (6)
  // 3) Gate-modulated token pooling and prediction
  Compute token scores $s_k = \psi(\mathbf{f}_k)$ // Eq. (7)
  Compute gate-modulated weights $\alpha_k = \text{softmax}(s_k + \log(\tilde{g}_k))$ with $\tilde{g}_k = \max(g_k, \epsilon)$ // Eq. (8)
  Pool representation $\mathbf{h} = \sum_{k=0}^{K} \alpha_k \mathbf{f}_k$ and logits $\boldsymbol{\ell} = \theta_c(\mathbf{h})$ // Eqs. (9)--(10)
  // 4) Losses and update
  Compute $\mathcal{L}_{\text{CE}}$ // Eq. (11)
  Compute compute regularizer $\mathcal{L}_{\text{bud}} = \lambda_{\text{bud}}(\mathbb{E}[K] - K_{\text{tgt}})^2$ // Eq. (12)
  Compute Soft-ECE penalty $\mathcal{L}_{\text{ece}}$ // Eq. (16)
  Update parameters to minimize $\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{bud}} + \lambda_{\text{ece}}\mathcal{L}_{\text{ece}}$ // Eq. (17)
end
/* Inference Phase */
Procedure PrunedInference($\mathbf{x}, K_{\max}, \tau$):
  Predict $\{(c_{x,k}, c_{y,k}, s_k)\}_{k=1}^{K}$ and logits $\{a_k\}_{k=1}^{K}$ using $\rho(\mathbf{x})$
  Compute probabilities $p_k = \sigma(a_k)$ // Eq. (2)
  if $K_{\max} = 0$ **then**
    Extract only the global crop; compute logits via Eqs. (6)–(10) with $g_0 = 1$
    return $\text{softmax}(\boldsymbol{\ell})$
  end
  Select indices $\mathcal{I} = \text{top-}K_{\max}(p_k)$
  Apply thresholding within $\mathcal{I}$: $m_k = \mathbb{I}[p_k > \tau]$ and $\hat{g}_k = p_k \cdot m_k$ // Eq. (18)
  // Guarantee non-empty ROI set
  if $\sum_{k\in\mathcal{I}} m_k = 0$ **then**
    Let $k^\star = \arg\max_{k\in\mathcal{I}} p_k$; set $m_{k^\star} = 1$ and $\hat{g}_{k^\star} = p_{k^\star}$
  end
  Extract crops only for the global crop and ROIs with $m_k = 1$; encode with shared backbone ϕ
  Compute pooled logits using Eqs. (7)–(10) with gates $g_0 = 1$ and ROI gates $g_k \leftarrow \hat{g}_k$
  return $\text{softmax}(\boldsymbol{\ell})$

denote the standard per-class counts for class c.

$$\text{Prec}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}. \tag{20}$$

$$\text{Rec}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}. \tag{21}$$

Macro F1 is computed by averaging *per-class* F1 scores (not by taking the F1 of macro precision/recall). For each class, $\text{F1}_c = \frac{2\,\text{Prec}_c\,\text{Rec}_c}{\text{Prec}_c + \text{Rec}_c}$, and

$$\text{F1}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\text{F1}_c. \tag{22}$$

We additionally report macro AUC under a one-vs-rest (OvR) decomposition, which evaluates ranking quality of class probabilities beyond a single decision threshold. Let AUC_c be the ROC AUC for class c treated as positive against all remaining classes [48,49]. The macro average is

$$\text{AUC}^{\text{OvR}}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\text{AUC}_c. \tag{23}$$

4.1.2. Calibration and probabilistic metrics

Let $q_i = \max_c \hat{p}_{i,c}$ denote the predicted confidence for sample i. Mean confidence summarizes how peaked the predictive distributions are:

$$\overline{\text{Conf}} = \frac{1}{N}\sum_{i=1}^{N} q_i. \tag{24}$$

Negative log-likelihood (NLL) penalizes low probability assigned to the true class and is a strictly proper scoring rule:

$$\text{NLL} = -\frac{1}{N}\sum_{i=1}^{N}\log \hat{p}_{i,y_i}. \tag{25}$$

The multiclass Brier score measures the squared error between the predicted probability vector and the one-hot target $\mathbf{e}_{y_i}$ [50]:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{p}}_i - \mathbf{e}_{y_i}\right\|_2^2. \tag{26}$$

Expected calibration error (ECE) measures the discrepancy between predicted confidence and empirical accuracy via confidence binning [51]. Let $q_i \in [0,1]$ denote the predicted confidence of sample i (e.g., the maximum softmax probability) and let $\hat{y}_i$ and y_i be the predicted and ground-truth labels, respectively. We partition the interval $[0,1]$ into M=15 equal-width bins $\{B_m\}_{m=1}^{M}$, where

$$B_m = \left(\frac{m-1}{M}, \frac{m}{M}\right], \quad m = 1, \ldots, M. \tag{27}$$

For each bin B_m, we compute the empirical accuracy

$$\text{Acc}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m}\mathbb{I}\left[\hat{y}_i = y_i\right], \tag{28}$$

and the mean confidence

$$\text{Conf}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m} q_i, \tag{29}$$

where $|B_m|$ is the number of samples whose confidence falls into bin B_m and N is the total number of samples. The ECE with M=15 bins is then

$$\text{ECE}_{15} = \sum_{m=1}^{15}\frac{|B_m|}{N}\left|\text{Acc}(B_m) - \text{Conf}(B_m)\right|. \tag{30}$$

4.1.3. Efficiency and computational metrics

Inference efficiency is reported using wall-clock time. Let T denote the total elapsed time (in seconds) to process N images under a fixed inference configuration. Average latency is

$$\text{Latency (ms/img)} = 10^3\frac{T}{N}, \tag{31}$$

and throughput is

$$\text{Throughput (img/s)} = \frac{N}{T}. \tag{32}$$

For efficiency benchmarking, inference latency and throughput were measured on a single workstation equipped with an NVIDIA RTX 3080 GPU, an AMD Ryzen 5 5600G CPU, and 16 GB DDR4 RAM. Benchmarking was performed with a fixed batch size of 16 and four data-loading workers, with pinned host memory enabled for the evaluation loaders. To ensure fair comparison across heterogeneous backbones while respecting architecture-specific input conventions, convolutional backbones were evaluated at a fixed input resolution of 192 × 192, whereas ViT-based models were evaluated at their native input resolution (e.g., 224 × 224 or the backbone-specific default size). For each backbone, the raw model and the proposed framework were benchmarked under the same input-size setting. Before timing, each model was executed for 10 warm-up batches to stabilize GPU execution. Wall-clock time was then measured with a performance counter under inference mode, with CUDA synchronization applied immediately before and after each forward pass.

4.2. Experimental results

Our empirical goal is not to advocate a particular backbone, but to validate that the proposed framework behaves as a *backbone-agnostic efficiency layer*: it should preserve (or improve) recognition performance while substantially reducing inference latency. Unless noted otherwise, all reported metrics are averaged over three random seeds (0, 24, 48), and efficiency is measured according to the fixed benchmarking protocol described in Section 4.1.3.

Unless noted otherwise, all reported metrics are averaged over three random seeds (0, 24, 48), and efficiency is measured under a fixed inference batch size (batch size = 16) on the same hardware to ensure fair and consistent timing.

4.2.1. ROI visualization and interpretability

Fig. 4 provides qualitative support for the framework's central claim: efficiency gains are achieved by learning to discard spatial redundancy rather than by relying on a specific architecture. Early in training, the candidate set often includes partially informative regions (e.g., tissue boundaries or broad anatomical context), and some crops can be weakly relevant. As optimization progresses, the selected crops become more stable and repeatedly center on lesion-bearing regions across diverse appearances, while progressively excluding large background areas that do not contribute to class discrimination. This training-time behavior aligns with the observed test-time speedups: by confining the most expensive feature extraction to a small set of candidate regions, the framework reduces the effective spatial workload while retaining the structures that dominate class evidence.

Importantly, the ROI behavior is not merely "highlighting something": the selected crops exhibit *consistency* across epochs and across cases, which is a stronger interpretability signal than a single attention map. When occasional off-target crops appear (typically among the lower-ranked candidates), their presence is mitigated by the top-K selection and by the fact that the final decision aggregates evidence across candidates. This offers a plausible mechanism for why accuracy remains stable even when inference becomes substantially faster.

4.2.2. Modality-wise quantitative results and backbone-agnostic trends

Across all three modalities (Tables 4–6), the proposed framework produces a consistent pareto improvement: it moves the operating point toward substantially higher throughput at comparable (and in some cases higher) accuracy and macro-F1. This pattern is visible even when comparing against the strongest accuracy-oriented baselines, which are typically heavier models or models that retain full-resolution processing.

T1-weighted MRI (Table 4). On T1, the framework attains 0.961 ± 0.004 accuracy and 0.939 ± 0.004 macro-F1, while achieving 5189.0 ± 156.5 img/s. The best-accuracy baseline (DenseNet-121) reaches 0.964 ± 0.002 accuracy at only 913.1 ± 67.7 img/s. The absolute accuracy difference is -0.003 (within seed variability), whereas throughput improves by $\approx 5.68\times$ relative to the strongest accuracy baseline. In latency terms, this corresponds to reducing per-image time from 1.099 ms/img (DenseNet-121) to 0.193 ms/img. Notably, even against the fastest baseline in this table (ConvNeXtV2-Atto), the framework still provides a $\approx 2.57\times$ throughput increase, indicating that the speedup is not a trivial artifact of comparing against slow networks.

T1C+ MRI (Table 5). On contrast-enhanced scans, the framework achieves the best overall accuracy (0.980 ± 0.009) while simultaneously delivering 5065.1 ± 265.4 img/s. Relative to the best baseline accuracy (EfficientNet-B0, 0.975 ± 0.004), the framework increases throughput by $\approx 4.11\times$ (from 1231.7 ± 60.7 to 5065.1 ± 265.4 img/s). This modality is particularly informative because enhancement patterns can be spatially localized; the ROI mechanism in Fig. 4 provides a direct explanation for why reducing the processed field-of-view does not compromise recognition.

T2-weighted MRI (Table 6). T2 shows the clearest separation in terms of combined accuracy and efficiency. The framework attains 0.951 ± 0.021 accuracy and 0.937 ± 0.020 macro-F1 at 5206.9 ± 185.9 img/s, outperforming the best baseline accuracy (EfficientNet-B0, 0.927 ± 0.007) while being $\approx 3.99\times$ faster (baseline 1303.4 ± 24.7 img/s). While this accuracy gain is larger in magnitude than in T1/T1C+, the standard deviation is also higher, suggesting that T2 performance may be more sensitive to seed-level variation and/or class composition in the splits. A conservative interpretation is therefore appropriate: the framework is *at least* accuracy-preserving on T2 while delivering large speedups, and it frequently improves performance in practice.

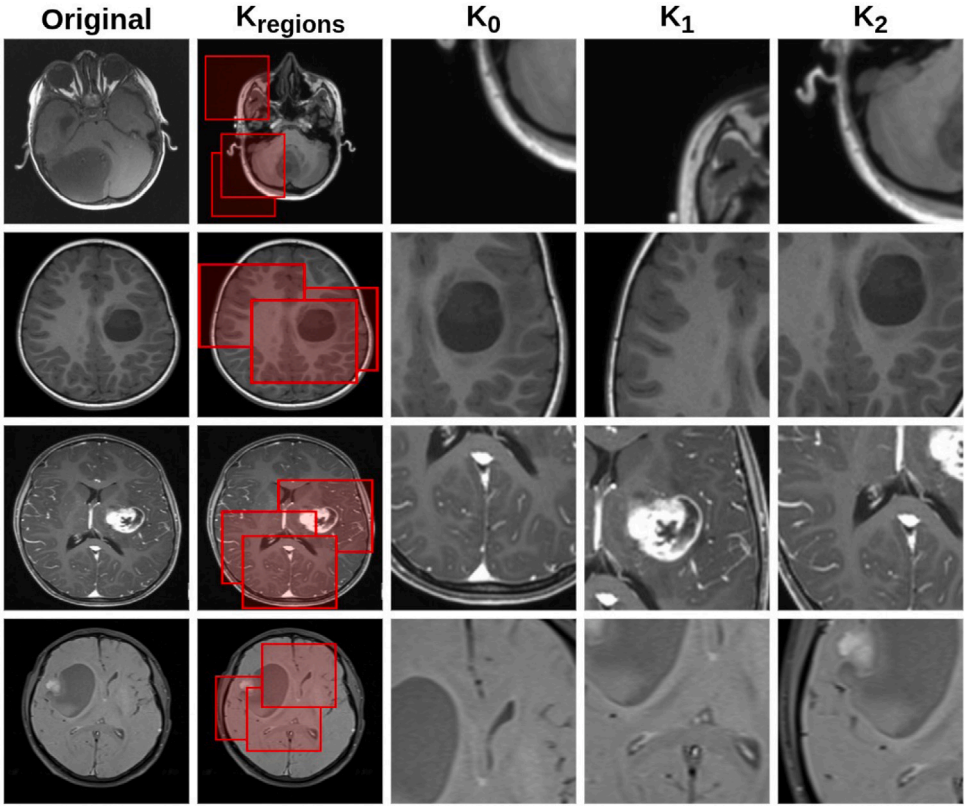

Fig. 4. Progressive refinement of ROI selection during training. For each example, the first column shows the original MRI slice and the second column shows the top-K candidate regions proposed at the corresponding training epoch (red boxes). The remaining columns visualize the cropped regions (K_0–K_2). As training proceeds, the selected regions become more consistent and increasingly concentrate on clinically relevant tumor areas, indicating improved localization and reduced background emphasis over epochs.

Table 4
Validation results on T1-weighted MRI reported as mean ± std over three seeds (0, 24, 48). For each seed, the checkpoint is selected by validation accuracy. Arrows indicate whether higher (↑) or lower (↓) values are preferred. Best and second-best results are **bold** and underlined, respectively. Rows are sorted by throughput (img/s) in ascending order.

Model	Params (M)	Classification					Calibration				Efficiency	
		Acc ↑	P_{macro} ↑	R_{macro} ↑	$F1_{macro}$ ↑	AUC_{macro} ↑	Conf.	ECE_{15} ↓	NLL ↓	Brier ↓	ms/img ↓	img/s ↑
DenseNet-121	6.97	**0.964±0.002**	**0.961±0.007**	**0.946±0.007**	**0.950±0.008**	**0.9992±0.0003**	0.935±0.021	0.036±0.021	**0.156±0.027**	0.065±0.010	1.099±0.081	913.1±67.7
VGG-19-BN	139.64	0.958±0.005	0.953±0.026	0.933±0.013	0.936±0.006	0.9982±0.0008	0.976±0.003	0.037±0.005	0.234±0.017	0.080±0.007	1.017±0.004	983.0±3.7
VGG-16-BN	134.33	0.959±0.003	0.952±0.010	0.927±0.007	0.932±0.010	0.9984±0.0014	**0.984±0.002**	0.032±0.007	0.264±0.085	0.074±0.011	0.875±0.001	1142.8±0.9
Inception-V3	21.82	0.951±0.002	0.935±0.005	0.930±0.017	0.927±0.008	0.9971±0.0007	0.975±0.005	0.037±0.004	0.269±0.035	0.089±0.005	0.801±0.103	1261.1±153.2
ConvNeXtV2-Tiny	27.88	0.873±0.137	0.870±0.113	0.789±0.203	0.795±0.195	0.9820±0.0273	0.855±0.194	0.049±0.034	0.449±0.425	0.192±0.195	0.782±0.005	1279.2±8.7
Xception41	24.95	0.959±0.009	0.951±0.009	0.937±0.020	0.937±0.015	0.9990±0.0003	0.959±0.012	**0.021±0.008**	0.162±0.020	0.065±0.012	0.667±0.058	1505.4±124.4
EfficientNet-B0	4.03	0.958±0.002	0.939±0.011	0.940±0.015	0.934±0.005	0.9991±0.0004	0.969±0.012	0.033±0.002	0.180±0.034	0.073±0.004	0.659±0.059	1526.3±132.4
MobileNetV3-L	4.22	0.921±0.007	0.902±0.017	0.871±0.021	0.873±0.016	0.9964±0.0008	0.930±0.006	0.029±0.010	0.286±0.031	0.123±0.010	0.506±0.027	1981.9±104.2
ConvNeXtV2-Atto	3.39	0.921±0.012	0.909±0.024	0.868±0.029	0.876±0.017	0.9972±0.0012	0.916±0.005	0.033±0.004	0.268±0.067	0.118±0.022	0.495±0.016	2022.5±66.0
Proposed Model	4.05	0.961±0.004	0.945±0.010	**0.946±0.004**	0.939±0.004	0.997±0.002	0.936±0.017	0.037±0.017	0.187±0.023	**0.062±0.002**	**0.193±0.006**	**5189.0±156.5**

Conf.: mean confidence (average maximum predicted probability). ECE_{15} uses 15 equal-width bins over $[0, 1]$. Values are mean ± std.

Table 5
Validation results on T1 contrast-enhanced (T1C+) MRI reported as mean ± std over three seeds (0, 24, 48). For each seed, the checkpoint is selected by validation accuracy. Arrows indicate whether higher (↑) or lower (↓) values are preferred. Best and second-best results are **bold** and underlined, respectively. Rows are sorted by throughput (img/s) in ascending order.

Model	Params (M)	Classification					Calibration				Efficiency	
		Acc ↑	P_{macro} ↑	R_{macro} ↑	$F1_{macro}$ ↑	AUC_{macro} ↑	Conf.	ECE_{15} ↓	NLL ↓	Brier ↓	ms/img ↓	img/s ↑
DenseNet-121	6.97	0.961±0.003	0.958±0.012	0.928±0.008	0.937±0.006	0.9971±0.0012	0.962±0.008	0.017±0.006	0.150±0.015	0.063±0.003	1.448±0.135	694.9±66.8
Inception-V3	21.82	0.972±0.005	0.959±0.020	0.954±0.018	0.955±0.018	0.9983±0.0010	0.987±0.002	0.026±0.004	0.173±0.037	0.052±0.009	1.071±0.125	941.8±103.9
VGG-19-BN	139.64	0.965±0.004	0.951±0.007	0.935±0.015	0.939±0.008	0.9978±0.0007	0.989±0.001	0.030±0.004	0.255±0.019	0.064±0.005	1.016±0.001	983.8±1.2
VGG-16-BN	134.33	0.971±0.001	0.962±0.015	0.953±0.012	0.956±0.009	**0.9994±0.0001**	0.989±0.004	0.026±0.003	0.159±0.028	0.053±0.004	0.889±0.026	1126.1±32.7
EfficientNet-B0	4.03	0.975±0.004	0.963±0.010	0.955±0.013	0.957±0.003	0.9983±0.0009	0.986±0.004	**0.016±0.004**	0.135±0.024	0.043±0.006	0.813±0.039	1231.7±60.7
Xception41	24.95	0.971±0.004	**0.968±0.013**	0.944±0.016	0.952±0.017	0.9986±0.0007	0.979±0.001	0.020±0.001	0.130±0.014	0.049±0.005	0.797±0.033	1256.0±53.3
ConvNeXtV2-Tiny	27.88	0.942±0.041	0.922±0.047	0.890±0.059	0.896±0.063	0.9911±0.0026	0.964±0.039	0.035±0.016	0.229±0.124	0.086±0.058	0.766±0.002	1305.4±3.0
MobileNetV3-L	4.22	0.946±0.003	0.930±0.024	0.897±0.027	0.910±0.026	0.9977±0.0017	0.957±0.010	0.027±0.011	0.203±0.041	0.090±0.006	0.644±0.064	1564.0±155.4
ConvNeXtV2-Atto	3.39	0.948±0.008	0.946±0.011	0.888±0.018	0.898±0.015	0.9936±0.0030	0.947±0.016	0.021±0.004	0.186±0.033	0.079±0.011	0.572±0.039	1753.0±115.0
Proposed Model	4.05	**0.980±0.009**	0.963±0.019	**0.963±0.012**	**0.962±0.016**	0.996±0.002	0.954±0.008	0.035±0.008	**0.111±0.020**	**0.036±0.008**	**0.198±0.011**	**5065.1±265.4**

Conf.: mean confidence (average maximum predicted probability). ECE_{15} uses 15 equal-width bins over $[0, 1]$. Values are mean ± std.

Table 6
Validation results on T2-weighted MRI reported as mean ± std over three seeds (0, 24, 48). For each seed, the checkpoint is selected by validation accuracy. Arrows indicate whether higher (↑) or lower (↓) values are preferred. Best and second-best results are **bold** and underlined, respectively. Rows are sorted by throughput (img/s) in ascending order.

Model	Params (M)	Classification					Calibration				Efficiency	
		Acc ↑	P_{macro} ↑	R_{macro} ↑	$F1_{macro}$ ↑	AUC_{macro} ↑	Conf.	ECE_{15} ↓	NLL ↓	Brier ↓	ms/img ↓	img/s ↑
DenseNet-121	6.97	0.908±0.004	0.890±0.004	0.885±0.009	0.882±0.009	0.9907±0.0009	0.897±0.033	0.042±0.020	0.344±0.022	0.146±0.002	1.261±0.056	794.3±35.5
VGG-19-BN	139.64	0.905±0.016	0.903±0.026	0.887±0.017	0.888±0.018	0.9950±0.0013	0.964±0.012	0.069±0.010	0.482±0.026	0.161±0.025	1.017±0.002	983.1±2.3
Inception-V3	21.82	0.892±0.011	0.899±0.031	0.861±0.016	0.873±0.020	0.9927±0.0037	0.946±0.001	0.071±0.016	0.462±0.112	0.184±0.028	0.955±0.074	1051.2±83.5
VGG-16-BN	134.33	0.910±0.010	0.894±0.011	0.903±0.005	0.893±0.007	0.9933±0.0005	**0.967±0.007**	0.065±0.011	0.530±0.072	0.159±0.022	0.875±0.002	1143.1±3.2
ConvNeXtV2-Tiny	27.88	0.914±0.011	0.896±0.014	0.862±0.026	0.865±0.024	0.9908±0.0019	0.949±0.021	0.050±0.016	0.341±0.068	0.136±0.018	0.777±0.004	1286.2±6.0
EfficientNet-B0	4.03	0.927±0.007	0.916±0.019	0.897±0.019	0.899±0.011	0.9936±0.0017	0.951±0.009	0.043±0.009	0.327±0.044	0.119±0.002	0.767±0.015	1303.4±24.7
Xception41	24.95	0.923±0.009	0.922±0.010	0.879±0.013	0.887±0.012	0.9892±0.0050	0.944±0.011	0.045±0.009	0.315±0.059	0.130±0.021	0.770±0.072	1306.5±116.2
MobileNetV3-L	4.22	0.868±0.012	0.855±0.015	0.828±0.007	0.833±0.005	0.9854±0.0021	0.874±0.011	0.044±0.005	0.517±0.039	0.209±0.018	0.576±0.032	1739.3±100.0
ConvNeXtV2-Atto	3.39	0.862±0.010	0.835±0.046	0.795±0.007	0.801±0.027	0.9821±0.0022	0.867±0.009	**0.040±0.003**	0.503±0.037	0.214±0.019	0.523±0.012	1913.2±43.7
Proposed Model	4.05	**0.951±0.021**	**0.957±0.022**	**0.926±0.017**	**0.937±0.020**	**0.998±0.001**	0.939±0.022	**0.040±0.004**	**0.187±0.066**	**0.075±0.030**	**0.192±0.007**	**5206.9±185.9**

Conf.: mean confidence (average maximum predicted probability). ECE_{15} uses 15 equal-width bins over $[0, 1]$. Values are mean ± std.

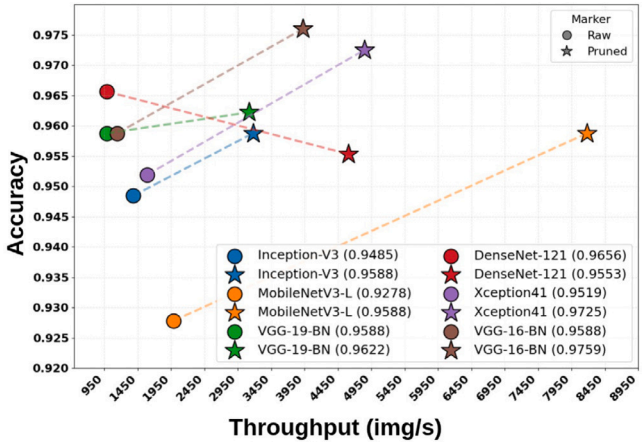

Fig. 5. Accuracy–throughput trade-off on the T1-weighted MRI setting across multiple SOTA backbones. Circles indicate the unmodified (raw) models, and stars indicate the same backbones evaluated with the proposed framework. Dashed lines connect each backbone's paired raw and framework results, highlighting how the framework shifts the operating point.

Statistical perspective (seed-level variability). To avoid over-claiming from overlapping error bars, we evaluate improvements through the lens of seed variability. Using a conservative Welch's two-sample t-test based on the reported mean ± std across three seeds, throughput gains over the best-accuracy baseline are statistically decisive in all modalities (T1: $p = 5.98 \times 10^{-5}$; T1C+: $p = 9.93 \times 10^{-4}$; T2: $p = 6.25 \times 10^{-4}$). In contrast, accuracy differences versus the best-accuracy baseline are not statistically significant under this conservative test (T1: $p = 0.331$; T1C+: $p = 0.449$; T2: $p = 0.178$). These results support the main claim the framework is designed to substantiate: the efficiency gains are robust and far exceed run-to-run variance, whereas accuracy changes are within the range of seed variability.

Fig. 5 further isolates the backbone-agnostic nature of the framework by comparing each backbone *with and without* the wrapper under identical timing conditions. Across six diverse architectures (MobileNetV3-L, Inception-V3, VGG-16/19, DenseNet-121, Xception41), the framework yields a consistent rightward shift in throughput, ranging from 2.30× to 4.69× (mean ≈ 3.46×), while accuracy changes remain small in absolute terms (typically within ±1–3%). Importantly, the gains are not restricted to either lightweight or heavyweight networks: the framework improves throughput on compact models (e.g., MobileNetV3-L) and also substantially accelerates slower, higher-capacity backbones. This is precisely the intended operating regime: the framework should be deployable as an inference-time efficiency layer without requiring architectural redesign.

Table 7
Ablation summary on T1 with Xception41. Mean±std over 3 seeds (0, 24, 48).

Variant	Acc↑	ECE↓	Thr. (img/s)↑
Raw	0.9519 ± 0.0091	0.0174 ± 0.0292	1590 ± 5
Proposed	**0.9725 ± 0.0040**	**0.0100 ± 0.0130**	**3079 ± 25**
−ROI	0.9622 ± 0.0072	0.0205 ± 0.0157	2217 ± 88
−Calibration	0.9691 ± 0.0034	0.0535 ± 0.0121	3074 ± 2
−ROI −Calibration	0.9587 ± 0.0121	0.0559 ± 0.0080	2003 ± 7

4.2.3. ROC curves and class-wise behavior

To complement the aggregate classification metrics, Fig. 6 summarizes the class distribution of the curated 15-class corpus across T1, T1C+, and T2 modalities and presents the corresponding modality-wise confusion matrices, thereby linking class support with the observed class-wise prediction behavior.

While macro-AUC values are high in most settings, Fig. 7 reveals meaningful heterogeneity at the class level. Common pathologies exhibit near-ceiling ROC behavior across modalities and backbones, indicating strong separability when sufficient positive examples exist. However, several rare subtypes show AUC values reported near chance (e.g., Carcinoma, Ependymoma, Ganglioglioma, Germinoma, Glioblastoma, Granuloma, Medulloblastoma, Neurocytoma, Papilloma, Tuberculoma; exact instances vary by modality/backbone panel). This pattern should not be treated as a minor footnote: it indicates that for a subset of classes, one-vs-rest discrimination can be unstable under limited support.

A critical nuance is that one-vs-rest AUC becomes ill-defined when a class has too few positives (or is absent) in a particular split; standard evaluation pipelines often default to 0.5 in such degenerate cases. The presence of multiple 0.5 entries across different panels is therefore consistent with severe class imbalance and split-level scarcity, rather than implying that the model "always guesses" those categories. Practically, the framework does not eliminate this limitation because it operates as an efficiency wrapper: it improves compute allocation but inherits the underlying data imbalance sensitivity of the base classifier.

This observation motivates two concrete reporting and modeling recommendations for a Q1-ready manuscript: (i) accompany ROC with per-class support counts (or at least flag classes below a minimum support threshold) and optionally add macro-averaged PR curves, which are often more informative under imbalance; and (ii) consider imbalance-aware training within the framework (e.g., class-balanced sampling, focal loss, or cost-sensitive calibration), which targets the rare-class failure mode directly without compromising the backbone-agnostic design principle.

4.2.4. Ablation study: ROI vs calibration contributions

Table 7 disentangles the roles of ROI selection and calibration in the proposed framework compared to the baseline model. Three conclusions emerge.

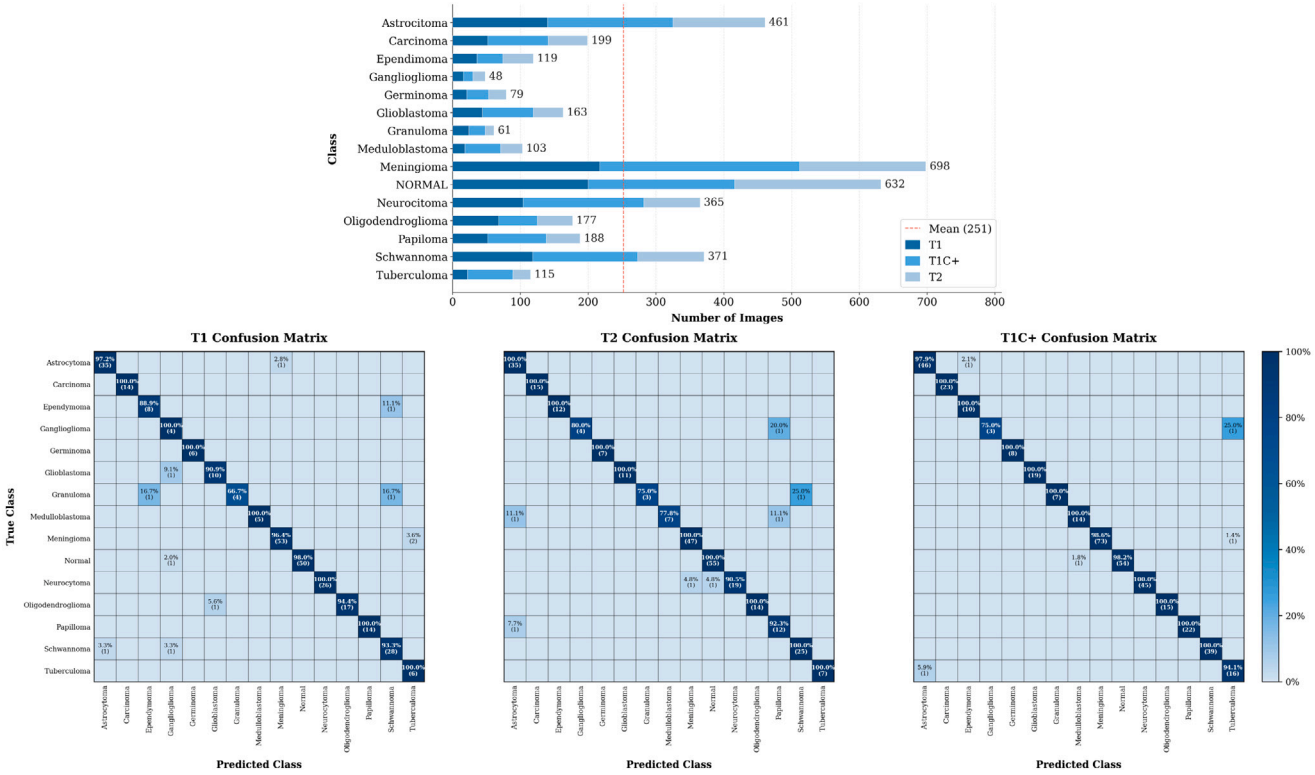

Fig. 6. Overview of the curated 15-class Fernando Feltrin corpus and modality-wise classification behavior. The top panel shows the class distribution across T1, contrast-enhanced T1 (T1C+), and T2 MRI modalities, together with the mean class count indicated by the dashed vertical line. The bottom panels present row-normalized confusion matrices for the proposed framework on T1, T2, and T1C+ MRI, respectively. Rows denote true classes and columns denote predicted classes; each cell reports the percentage of samples assigned to the corresponding predicted class, with the raw count shown in parentheses. Overall, the matrices indicate strong class-wise separability across modalities, while also revealing the relative difficulty of several low-support categories.

First, the full framework nearly doubles throughput over the raw configuration (3079 vs 1590 img/s; ≈ 1.94×) while improving accuracy (from 0.9519 to 0.9725). Second, removing ROI selection reduces throughput from 3079 to 2217 img/s, demonstrating that spatial selection is a primary driver of efficiency (the full framework is ≈ 1.39× faster than the no-ROI variant under the same fixed-cost setting). Third, removing calibration leaves throughput essentially unchanged (3074 img/s) but substantially degrades calibration quality (ECE increases from 0.0100 to 0.0535). This separation of effects is desirable: ROI is responsible for compute savings, while calibration stabilizes confidence estimates under aggressive pruning—a key property for safety-critical deployment where well-calibrated probabilities matter as much as accuracy.

Robustness to ROI specification is evaluated by systematically perturbing ROI geometry at inference, varying both ROI scale (0.70×–1.30×) and ROI center location (0–16 px jitter). As shown in Fig. 8, classification performance remains highly stable across the entire perturbation grid: Accuracy and Macro-F1 exhibit negligible variation, indicating that the decision function is not dependent on a narrowly tuned ROI size or a precise ROI placement. This stability is complemented by consistent predictive reliability, where ECE (15 bins) and NLL change only marginally under increasingly aggressive scale and position shifts. These perturbations intentionally emulate realistic ROI-selection variability, including scenarios where enlarged ROIs may incorporate more surrounding context (a potential smoothing/masking effect) and where shifted ROIs may partially overlap neighboring tissue (a potential "hybrid" ROI). The observed invariance in both discriminative performance and calibration suggests that ROI-patch inference leverages robust, pathology-relevant cues rather than brittle, ROI-specific artifacts, supporting reliable deployment under practical ROI imprecision.

Takeaway. Across modalities and across heterogeneous backbones, the experimental evidence supports a consistent narrative: the framework acts as a backbone-agnostic wrapper that reallocates computation to informative spatial regions, yielding large and robust throughput gains, while keeping accuracy within the variability induced by random initialization and data ordering. The ROC analysis clarifies where the approach does *not* solve the problem (rare-class instability under extreme imbalance), and the ablation study explains why the full design is needed (ROI for speed, calibration for reliability).

4.3. Additional multi-source evaluation on an external MRI corpus

To examine whether the observed behavior extends beyond the primary 15-class corpus, we conducted supplementary experiments on a harmonized four-class MRI corpus constructed from the Nickparvar dataset [32] and the BRISC dataset [33]. These additional experiments were designed to broaden the empirical basis of the study by evaluating the candidate backbones under the same three-seed protocol and efficiency benchmarking setup used in the main experiments. Because the current results are obtained on a harmonized external corpus rather than through a strict train-on-one-dataset/test-on-another transfer setting, they should be interpreted as a multi-source replication analysis rather than as definitive evidence of fully independent cross-dataset generalization. Table 8 summarizes the backbone-level results on the additional harmonized external MRI corpus and highlights the accuracy throughput trade-off achieved by the proposed pruned framework.

Across all evaluated backbones, performance on the additional external corpus remained consistently strong. Validation accuracy ranged from 0.9946 ± 0.0000 to 0.9976 ± 0.0010, while macro-F1 ranged from 0.9946 ± 0.0000 to 0.9976 ± 0.0010, and macro-AUC remained near ceiling across all baseline models. In absolute terms, DenseNet-

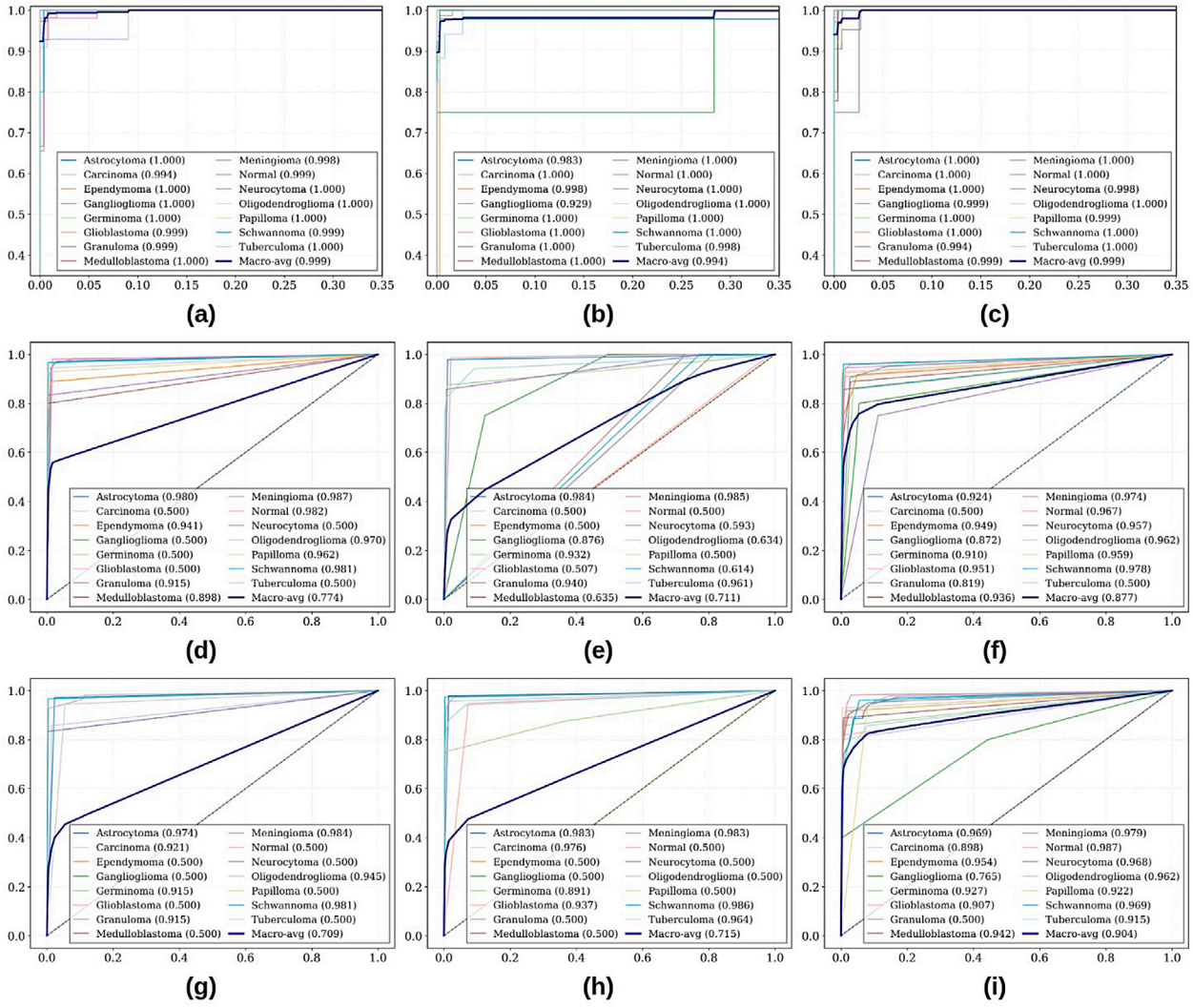

Fig. 7. Multi-class ROC curves across MRI modalities and backbones. Columns from left to right correspond to T1, T1C+, and T2. Rows from top to bottom correspond to EfficientNet-B0, Xception41, and DenseNet-121. Each subplot (a–i) reports one-vs-rest ROC curves for all classes, along with the macro-average (thick curve); the diagonal indicates chance-level performance.

Table 8
Results on the additional harmonized external MRI corpus constructed from the Nickparvar and BRISC datasets. All rows are reported as mean ± std over three seeds (0, 24, 48). Arrows indicate whether higher (↑) or lower (↓) values are preferred. Best and second-best results are **bold** and underlined, respectively. Rows are sorted by throughput (img/s) in ascending order.

Model	Params (M)	Acc ↑	$F1_{macro}$ ↑	ECE_{15} ↓	ms/img ↓	img/s ↑
DenseNet-121	6.96	**0.998±0.001**	**0.998±0.001**	**0.002±0.001**	0.705±0.004	1419.1±7.7
Swin-T	27.52	0.997±0.000	0.997±0.000	**0.002±0.000**	0.576±0.063	1748.7±189.3
VGG-19-BN	139.60	0.997±0.001	0.997±0.001	0.004±0.000	0.547±0.015	1827.8±48.5
Inception-V3	21.79	0.997±0.001	0.997±0.001	0.003±0.001	0.547±0.040	1834.7±130.7
ConvNeXtV2-Tiny	27.87	0.997±0.001	0.997±0.001	0.003±0.000	0.510±0.009	1959.4±36.1
VGG-16-BN	134.29	0.997±0.001	0.997±0.001	0.003±0.000	0.477±0.018	2098.2±77.9
EfficientNet-B0	4.01	0.997±0.001	0.997±0.001	0.004±0.001	0.452±0.022	2215.5±102.6
Xception41	24.93	0.997±0.001	0.997±0.001	0.004±0.000	0.430±0.010	2327.8±56.3
MobileNetV3-L	4.21	0.997±0.001	0.997±0.001	0.003±0.001	0.381±0.014	2630.3±93.4
ConvNeXtV2-Atto	3.39	0.995±0.000	0.995±0.000	0.005±0.000	0.360±0.005	2781.7±35.1
Proposed (Pruned, calib.)	4.05	0.9935±0.0026	0.9935±0.0026	0.0023±0.0012	**0.0629±0.0010**	**15899.5±259.7**

ECE_{15} uses 15 equal-width bins over $[0, 1]$. All rows are reported as mean ± std over three seeds on the external MRI corpus.

121 achieved the highest validation accuracy (0.9976 ± 0.0010), whereas ConvNeXtV2-Atto delivered the highest baseline throughput (2781.7 ± 35.1 img/s). For the proposed framework, aggregating the calibrated results over three seeds (0, 24, 48), the full model achieved 0.9953 ± 0.0020 accuracy and 0.9953 ± 0.0020 macro-F1 with ECE_{15} of 0.0020 ± 0.0006, while the pruned model retained 0.9935 ± 0.0026 accuracy and 0.9935 ± 0.0026 macro-F1 with ECE_{15} of 0.0023 ± 0.0012. Although pruning introduced a small absolute accuracy reduction relative to full inference (approximately 0.18 percentage points), the pruned configuration achieved 15899.5 ± 259.7 img/s, corresponding to approximately 5.72× higher throughput than the fastest baseline and 11.20× relative to the best-accuracy baseline.

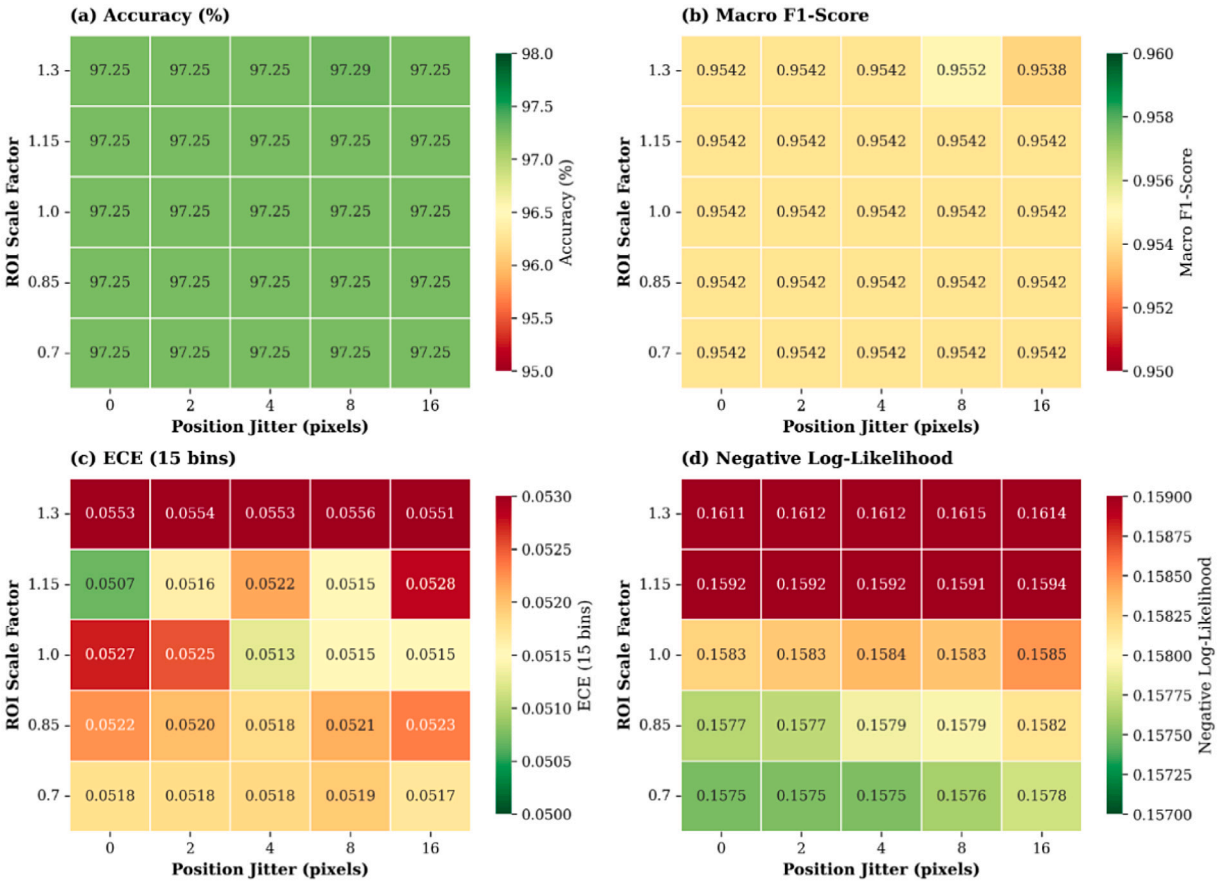

Fig. 8. ROI robustness under controlled scale and center perturbations. Heatmaps report mean performance across repeated runs for combinations of ROI scale factor (y-axis) and position jitter (x-axis, pixels): (a) Accuracy (%), (b) Macro F1-score, (c) Expected Calibration Error (ECE; 15 bins), and (d) Negative Log-Likelihood (NLL). Overall, accuracy and macro-F1 remain highly stable across the tested perturbation range, while calibration (ECE) and NLL show only minor variations, indicating the classifier's predictions are largely insensitive to moderate ROI size and center shifts.

Table 9
Full versus pruned evaluation of the proposed framework on the additional harmonized external MRI corpus, reported as mean ± std over three seeds (0, 24, 48).

Variant	Acc ↑	P_{macro} ↑	R_{macro} ↑	$F1_{macro}$ ↑	ECE_{15} ↓	NLL ↓	Brier ↓	img/s ↑
Full (calib.)	**0.9953 ± 0.0020**	**0.9953 ± 0.0020**	**0.9953 ± 0.0020**	**0.9953 ± 0.0020**	**0.0020 ± 0.0006**	**0.0213 ± 0.0061**	**0.0081 ± 0.0028**	–
Pruned (calib.)	0.9935 ± 0.0026	0.9935 ± 0.0026	0.9935 ± 0.0026	0.9935 ± 0.0026	0.0023 ± 0.0012	0.0295 ± 0.0122	0.0108 ± 0.0040	**15899.5 ± 259.7**

Metrics are reported after calibration. Throughput is reported for the optimized pruned inference configuration on the external MRI corpus.

These results suggest that the framework retains its core accuracy–efficiency behavior on the additional external MRI corpus, although a stricter train-on-one-source/test-on-another protocol would still be required to establish fully independent cross-dataset generalization.

To isolate the effect of pruning within the proposed framework, Table 9 compares the calibrated full and pruned inference modes on the same external corpus.

The results indicate that the proposed framework retains very high classification performance on the additional external MRI corpus while delivering a substantial throughput gain relative to the strongest baseline backbones. This suggests that the overall behavior of the framework is not confined to the original primary corpus alone. However, while these additional experiments strengthen the multi-source empirical evaluation of the framework, they should not be interpreted as a strict external-validation protocol, since the present analysis does not isolate train-on-one-source/test-on-another generalization across fully disjoint datasets.

5. Discussion

This work evaluates a framework-level contribution: a backbone-agnostic, content-adaptive inference layer that reallocates computation toward informative spatial regions while preserving discriminative performance and probabilistic reliability. Across three MRI modalities and heterogeneous CNN backbones, the evidence supports a consistent conclusion: the framework yields large and statistically decisive throughput gains, whereas accuracy differences relative to the best-performing baselines are small and statistically indistinguishable under the reported seed variability. The class-wise ROC analysis further clarifies a boundary condition of the approach: spatial compute optimization does not resolve distributional sparsity in rare subtypes. The additional independent-source experiments on the Nickparvar [32] and BRISC [33] datasets further indicate that the proposed framework is not limited to a single curated source, although a measurable cross-dataset gap remains under label and acquisition heterogeneity.

5.1. Backbone-agnostic efficiency as the primary contribution

The framework's generality follows from what it changes—*the spatial allocation of compute*—rather than how a backbone represents images. Because the proposed method operates as a backbone-agnostic efficiency layer rather than a standalone classifier, the absolute predictive performance remains dependent on the representational strength of the underlying backbone. The role of the framework is to reduce redundant spatial computation before backbone encoding, thereby improving throughput while preserving accuracy across heterogeneous encoders. Although the evaluated backbones differ in architectural priors, they all process substantial spatial redundancy in MRI slices, including background, skull/edge artifacts, and largely homogeneous tissue regions that contribute weakly to subtype discrimination. The ROI pathway reduces this redundancy by prioritizing a small set of candidate regions for downstream feature extraction, while maintaining sufficient contextual evidence through top-K selection and aggregation.

This behavior is qualitatively consistent with Fig. 4, where ROI proposals become increasingly stable over training and concentrate around lesion-bearing regions, while deprioritizing non-informative background. Importantly, the paired raw-to-framework shifts in Fig. 5 confirm structural independence from the backbone: for each architecture, the framework moves the operating point toward higher throughput without requiring a backbone-specific redesign. The magnitude of this shift is substantial across both lightweight and heavyweight networks, supporting the claim that the framework acts as a reusable efficiency layer rather than an architecture-tuned enhancement.

5.2. Modality-dependent effects grounded in MRI physics

Although the efficiency improvements are consistent across modalities, the magnitude of performance change differs. The most distinct combined gain appears in T2-weighted imaging (Table 6), whereas T1 and T1C+ more often show accuracy preservation with large speedups. This pattern is aligned with the physical contrast mechanisms of MRI sequences. T2-weighted scans emphasize fluid and edema, which often appear hyperintense relative to surrounding tissue; these signals create strong local intensity gradients around lesions and edema margins. An ROI selector effectively functions as a spatial "hotspot" mechanism: it benefits when discriminative evidence is localized and high-contrast, because candidate crops can reliably capture lesion cores and peritumoral edema while excluding large areas of weakly informative anatomy. This provides a mechanistic explanation for why ROI-driven spatial restriction can be particularly favorable in T2.

In contrast, T1-weighted imaging is more structural and anatomy-driven: tissue interfaces and global context may contribute more subtly to discrimination, and lesions can be less conspicuous without contrast enhancement. Under these conditions, aggressively restricting the processed field-of-view is less likely to yield systematic accuracy gains. The framework's objective, therefore, is not to force improvements on every modality, but to preserve discrimination while achieving decisive runtime acceleration. The empirical results reflect this: across modalities, throughput increases are pronounced, while accuracy changes remain within seed-level variability.

5.3. Statistical interpretation: decisive efficiency gains, indistinguishable accuracy differences

The reported mean ± std values indicate that throughput gains greatly exceed run-to-run variability, whereas accuracy differences often overlap. This distinction is confirmed by conservative statistical testing based on the available summary statistics: throughput improvements over the strongest accuracy baseline are statistically decisive in all modalities (T1: $p = 5.98 \times 10^{-5}$; T1C+: $p = 9.93 \times 10^{-4}$; T2: $p = 6.25 \times 10^{-4}$), while accuracy differences are statistically indistinguishable (T1: $p = 0.331$; T1C+: $p = 0.449$; T2: $p = 0.178$). These results support the core claim of the framework: it provides robust acceleration without evidence of a statistically meaningful loss in classification performance relative to the best baseline under the same experimental protocol.

Beyond point estimates, this framing is important for clinical translation. Inference efficiency is a systems-level property that must be reliable under repeated runs and deployment constraints. In contrast, small absolute accuracy differences on limited-seed experiments should not be over-interpreted without additional paired trials or cross-validation. The appropriate scientific conclusion from the present evidence is therefore that the framework's runtime gains are unequivocal, while the classification performance remains comparable to (and sometimes better than) strong baselines.

5.4. Efficiency–imbalance boundary condition: what ROC reveals

The per-class ROC curves (Fig. 7) reveal a clear dichotomy: common categories exhibit near-ceiling separability, while several rare subtypes show near-chance one-vs-rest AUC in specific modality/backbone panels. This should be treated as a defined boundary condition rather than a defect of the framework. The framework is an accelerator that redistributes spatial computation; it does not introduce new supervision signals, synthetic minority evidence, or a class-balancing objective. Consequently, it cannot be expected to solve distributional sparsity or split-level absence of rare classes. In other words, spatial optimization (ROI selection) improves compute efficiency and can preserve discriminative signal, but it does not address the statistical scarcity that governs rare-class ROC stability.

A further practical nuance is that one-vs-rest AUC can become unstable or degenerate under extreme imbalance, particularly when a class has very few positives (or is absent) in a split; some evaluation pipelines default to 0.5 in such cases. Regardless of the implementation detail, the scientific implication remains the same: rare-class performance is constrained primarily by support and label quality, and efficiency layers do not substitute for imbalance-aware learning. Addressing this boundary condition requires complementary methods (e.g., class-balanced sampling, cost-sensitive losses, or focal objectives), which operate on the data distribution rather than the compute distribution.

5.5. Reliability under acceleration: calibration is not optional

Clinical decision support depends not only on correct predictions but also on well-calibrated confidence. The ablation study (Table 7) isolates two roles: ROI selection primarily governs throughput, whereas calibration is essential for maintaining low ECE under compute reduction. Specifically, removing calibration leaves throughput essentially unchanged but substantially worsens ECE, indicating overconfident miscalibration under aggressive pruning. This separation of responsibilities is desirable: the framework accelerates inference through spatial selection while explicitly preserving probabilistic reliability through calibration, which is critical when probabilities are used for triage, referral thresholds, or human-in-the-loop review.

5.6. Limitations and implications

The conclusions should be interpreted within scope. The ROI interpretability evidence is qualitative due to the absence of pixel-wise annotations; future work can quantify localization faithfulness using perturbation-based tests or weak localization proxies. Related segmentation and weak-supervision frameworks, including recent transformer- and MaxViT-based designs for brain tumor delineation, may provide useful directions for incorporating stronger localization supervision in future extensions of the present framework [11–13]. The current statistical evidence is based on three seeds, sufficient to establish decisive throughput gains but not designed to resolve very small accuracy deltas. Finally, the ROC analysis highlights an imbalance boundary

condition that is intrinsic to the dataset and evaluation splits, not unique to the framework.

Overall, the results position the proposed framework as a deployable, architecture-independent accelerator that consistently improves inference throughput while preserving classification performance and calibrated confidence. Its primary strength is robustness across backbones and modalities; its primary boundary condition is rare-class scarcity, which requires distribution-aware learning rather than spatial compute reallocation.

6. Conclusion

This paper presented a backbone-agnostic efficiency framework for multi-class brain tumor MRI classification that redirects inference computation toward diagnostically relevant spatial regions while preserving predictive confidence. This research, experiments were conducted on a harmonized multi-source corpus constructed from three public datasets, namely the Fernando Feltrin dataset [31], the Nickparvar dataset [32], and the BRISC dataset [33]. Across three MRI modalities (T1, T1C+, and T2) and a range of SOTA CNN backbones, the framework consistently improved the accuracy-throughput trade-off: inference throughput increased by multiple-fold without statistically significant degradation in classification performance compared to the strongest baselines under identical evaluation protocols. Qualitative ROI visualizations corroborate the underlying mechanism, revealing progressively refined region proposals that focus on lesion-relevant content while filtering non-informative background, aligned with the observed efficiency gains. The class-wise ROC analysis identifies a limitation of the approach: while the framework maintains separability for well-represented categories, rare subtypes remain constrained by distributional scarcity and limited training support. This highlights a fundamental distinction between spatial compute optimization and class-imbalance mitigation—the proposed framework accelerates inference but does not substitute for distribution-aware learning strategies. The ablation study confirms the necessity of the complete design, showing that ROI selection enables efficiency improvements, while calibration is critical for maintaining confidence reliability under reduced computation.

In summary, the proposed method provides a practical, deployable solution for fast and reliable brain tumor classification that does not depend on a specific backbone and remains compatible with diverse encoders. The multi-source evaluation further broadens the experimental basis of the study by incorporating three public datasets within a unified classification setting. Future work will address rare-class robustness through imbalance-aware objectives incorporated within the framework and will enhance interpretability quantification using weak localization and perturbation-based evaluation protocols.

CRediT authorship contribution statement

Ashraful Alam Nirob: Methodology, Formal analysis, Data curation, Conceptualization. **Tasnim Sakib Apon:** Writing – review & editing, Methodology, Investigation. **Anika Tahsin:** Writing – review & editing, Software, Resources, Formal analysis. **Md. Golam Rabiul Alam:** Writing – review & editing, Validation, Supervision, Methodology. **Sandra Costanzo:** Writing – review & editing, Visualization. **Giancarlo Fortino:** Writing – review & editing, Validation, Resources. **Andrea Aliverti:** Validation, Writing – original draft, Writing – review & editing. **Mohammad Mehedi Hassan:** Writing – review & editing, Writing – original draft, Supervision, Formal analysis.

Code availability

The source code for this work is publicly available on GitHub at https://github.com/Ashraful-Alam-Nirob/Backbone-Agnostic-ROI-Inference.git.

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 are grateful to King Saud University, Riyadh, Saudi Arabia for funding this work through Ongoing Research Funding Program (ORF-2026-18). The authors Fortino and Costanzo also are partially supported by the POS RADIOAMICA project funded by the Italian Minister of Health (CUP: H53C2200065000).

Data availability

The data used in this study are derived from three publicly available brain tumor MRI datasets: the Kaggle dataset “Brain Tumor MRI Images (44 Classes)” by Fernando Feltrin [31] the Kaggle “Brain Tumor MRI Dataset” by Nickparvar [32], and the BRISC annotated dataset [33].

References

[1] International Agency for Research on Cancer, GLOBOCAN 2022 fact sheet: Brain, central nervous system, 2024, Reported global estimates for 2022: 321, 731 new cases and 248, 500 deaths. Global Cancer Observatory.

[2] W.H.O. Classification of Tumours Editorial Board, Central nervous system tumours, fifth ed., in: WHO Classification of Tumours, vol. 6, International Agency for Research on Cancer, Lyon, France, 2021.

[3] D.N. Louis, A. Perry, P. Wesseling, et al., The 2021 WHO classification of tumors of the central nervous system: a summary, Acta Neuropathol. (2021).

[4] A. Guzzo, G. Fortino, G. Greco, M. Maggiolini, Data and model aggregation for radiomics applications: Emerging trend and open challenges, Inf. Fusion 100 (2023) 101923, http://dx.doi.org/10.1016/j.inffus.2023.101923.

[5] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 1321–1330, URL https://proceedings.mlr.press/v70/guo17a.html.

[6] S.R. Khan, M. Sikandar, A. Almogren, I. Ud Din, A. Guerrieri, G. Fortino, Iomt-based computational approach for detecting brain tumor, Future Gener. Comput. Syst. 109 (2020) 360–367, http://dx.doi.org/10.1016/j.future.2020.03.054.

[7] A. Raza, A. Guzzo, M. Ianni, R. Lappano, A. Zanolini, M. Maggiolini, G. Fortino, Federated learning in radiomics: A comprehensive meta-survey on medical image analysis, Comput. Methods Programs Biomed. 267 (2025) 108768, http://dx.doi.org/10.1016/j.cmpb.2025.108768.

[8] E.I. Zacharaki, S. Wang, S. Chawla, D. Soo Yoo, R. Wolf, E.R. Melhem, C. Davatzikos, Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme, Magn. Reson. Med.: An Off. J. Int. Soc. Magn. Reson. Med. 62 (6) (2009) 1609–1618.

[9] S. Deepak, P. Ameer, Brain tumor classification using deep CNN features via transfer learning, Comput. Biol. Med. 111 (2019) 103345.

[10] 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.

[11] T.B. Nguyen-Tat, T.-Q.T. Nguyen, H.-N. Nguyen, V.M. Ngo, Enhancing brain tumor segmentation in MRI images: A hybrid approach using unet, attention mechanisms, and transformers, Egypt. Inform. J. 27 (2024) 100528, http://dx.doi.org/10.1016/j.eij.2024.100528.

[12] T.B. Nguyen-Tat, L.Q. Truong, K.Q. Truong, T.H. Ngo, DoubleBlock-ViT: A MaxViT-based enhancement and dual-path skip connections for brain tumor segmentation in MRI scans, Comput. Methods Programs Biomed. 274 (2026) 109165, http://dx.doi.org/10.1016/j.cmpb.2025.109165.

[13] T.B. Nguyen-Tat, H.-A. Vo, P.-S. Dang, QMaxViT-Unet+: A query-based MaxViT-Unet with edge enhancement for scribble-supervised segmentation of medical images, Comput. Biol. Med. 187 (2025) 109762, http://dx.doi.org/10.1016/j.compbiomed.2025.109762.

[14] Z. Zhang, Y. Zhang, W. Bao, C. Li, D. Yuan, Coarse-to-fine: A hierarchical DNN inference framework for edge computing, Future Gener. Comput. Syst. 157 (2024) 180–192, http://dx.doi.org/10.1016/j.future.2024.03.009.

[15] G. Liu, F. Dai, X. Xu, X. Fu, W. Dou, N. Kumar, M. Bilal, An adaptive DNN inference acceleration framework with end–edge–cloud collaborative computing, Future Gener. Comput. Syst. 140 (2023) 422–435, http://dx.doi.org/10.1016/j.future.2022.10.033.

[16] G. Tricomi, G. Cicceri, I. Ficili, S. Vitabile, G. Merlino, A. Puliafito, HERALD: A hybrid distributed learning incremental & federated solution for knowledge distillation in COVID-19 classification, Future Gener. Comput. Syst. 174 (2026) 107991, http://dx.doi.org/10.1016/j.future.2025.107991.

[17] J. Boudjadar, S.U. Islam, R. Buyya, Dynamic FPGA reconfiguration for scalable embedded artificial intelligence (AI): A co-design methodology for convolutional neural networks (CNN) acceleration, Future Gener. Comput. Syst. 169 (2025) 107777, http://dx.doi.org/10.1016/j.future.2025.107777.

[18] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, 2015, arXiv preprint arXiv:1506.02025.

[19] M. Ilse, J.M. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: Proceedings of the 35th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 2132–2141, URL https://proceedings.mlr.press/v80/ilse18a.html.

[20] C.J. Maddison, A. Mnih, Y.W. Teh, The concrete distribution: A continuous relaxation of discrete random variables, 2016, arXiv preprint arXiv:1611.00712.

[21] E. Jang, S. Gu, B. Poole, Categorical reparameterization with Gumbel-Softmax, in: International Conference on Learning Representations, ICLR, 2017.

[22] C. Louizos, M. Welling, D.P. Kingma, Learning sparse neural networks through L_0 regularization, in: International Conference on Learning Representations, ICLR, 2018, arXiv:1712.01312.

[23] A. Graves, Adaptive computation time for recurrent neural networks, 2016, arXiv preprint arXiv:1603.08983.

[24] Z. Wu, T. Nagarajan, A. Kumar, S. Rennie, L.S. Davis, K. Grauman, R. Feris, BlockDrop: Dynamic inference paths in residual networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018, arXiv:1711.08393.

[25] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, C.-J. Hsieh, DynamicViT: Efficient vision transformers with dynamic token sparsification, in: Advances in Neural Information Processing Systems (NeurIPS), 2021, arXiv:2106.02034.

[26] M.S. Ryoo, A.J. Piergiovanni, A. Arnab, M. Dehghani, A. Angelova, TokenLearner: Adaptive space-time tokenization for videos, in: Advances in Neural Information Processing Systems (NeurIPS), 2021.

[27] A. Karandikar, N. Cain, D. Tran, B. Lakshminarayanan, J. Shlens, M.C. Mozer, R. Roelofs, Soft calibration objectives for neural networks, in: Advances in Neural Information Processing Systems, vol. 34, 2021, URL https://proceedings.neurips.cc/paper/2021/hash/f8905bd3df64ace64a68e154ba72f24c-Abstract.html.

[28] O. Bohdal, Y. Yang, T. Hospedales, Meta-calibration: Learning of model calibration using differentiable expected calibration error, 2021, arXiv preprint arXiv:2106.09613 URL https://arxiv.org/abs/2106.09613.

[29] S. Rajaraman, P. Ganesan, S. Antani, Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks, PLOS ONE 17 (1) (2022) e0262838, http://dx.doi.org/10.1371/journal.pone.0262838, URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0262838.

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

[31] F. Feltrin, Brain tumor MRI images 44 classes, 2023, URL https://www.kaggle.com/datasets/fernando2rad/brain-tumor-mri-images-44c?datasetId=2892866&sortBy=voteCount.

[32] M. Nickparvar, Brain Tumor MRI Dataset, URL https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset, Kaggle dataset.

[33] BRISC Dataset, BRISC: Annotated dataset for brain tumor segmentation and classification, 2025, http://dx.doi.org/10.6084/m9.figshare.30533120, URL https://figshare.com/articles/dataset/_b_BRISC_Annotated_Dataset_for_Brain_Tumor_Segmentation_and_Classification_b_/30533120.

[34] A.J. Farrall, Magnetic resonance imaging, Pr. Neurol. 6 (5) (2006) 318–325, http://dx.doi.org/10.1136/jnnp.2006.091843, arXiv:https://pn.bmj.com/content/6/5/318.full.pdf, URL https://pn.bmj.com/content/6/5/318.

[35] D.A. Reardon, J.N. Rich, H.S. Friedman, D.D. Bigner, Recent advances in the treatment of malignant astrocytoma, J. Clin. Oncol. 24 (8) (2006) 1253–1265.

[36] C.C. medical professional, Carcinoma, 2023, URL https://my.clevelandclinic.org/health/diseases/23180-carcinoma.

[37] S. Zacharoulis, L. Moreno, Ependymoma: an update, J. Child Neurol. 24 (11) (2009) 1431–1438.

[38] J.M. Silver, C.E. Rawlings III, E. Rossitch Jr., S.M. Zeidman, A.H. Friedman, Ganglioglioma: a clinical study with long-term follow-up, Surg. Neurol. 35 (4) (1991) 261–266.

[39] D.S. Osorio, J.C. Allen, Management of CNS germinoma, CNS Oncol. 4 (4) (2015) 273–279.

[40] H.-G. Wirsching, M. Weller, Glioblastoma, Malig. Brain Tumors: State-of-the-Art Treat. (2017) 265–288.

[41] H. Australia, Granulomas, 2021, URL https://www.healthdirect.gov.au/granulomas.

[42] N.E. Millard, K.C. De Braganca, Medulloblastoma, J. Child Neurol. 31 (12) (2016) 1341–1353.

[43] J. Wiemels, M. Wrensch, E.B. Claus, Epidemiology and etiology of meningioma, J. Neurooncol. 99 (2010) 307–314.

[44] S.J. Lee, T.T. Bui, C.H.J. Chen, C. Lagman, L.K. Chung, S. Sidhu, D.J. Seo, W.H. Yong, T.L. Siegal, M. Kim, et al., Central neurocytoma: a review of clinical management and histopathologic features, Brain Tumor Res. Treat. 4 (2) (2016) 49–57.

[45] D. Sethi, R. Arora, K. Garg, P. Tanwar, Choroid plexus papilloma, Asian J. Neurosurg. 12 (01) (2017) 139–141.

[46] A.I. Persson, C. Petritsch, F.J. Swartling, M. Itsara, F.J. Sim, R. Auvergne, D.D. Goldenberg, S.R. Vandenberg, K.N. Nguyen, S. Yakovenko, et al., Non-stem cell origin for oligodendroglioma, Cancer Cell 18 (6) (2010) 669–682.

[47] C.R. U.K., Schwannoma, 2022, URL https://www.cancerresearchuk.org/about-cancer/brain-tumours/types/vestibular-schwannoma.

[48] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36, http://dx.doi.org/10.1148/radiology.143.1.7063747.

[49] D.J. Hand, R.J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, Mach. Learn. 45 (2001) 171–186, http://dx.doi.org/10.1023/A:1010920819831.

[50] G.W. Brier, Verification of forecasts expressed in terms of probability, Mon. Weather Rev. 78 (1) (1950) 1–3, http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[51] M.P. Naeini, G.F. Cooper, M. Hauskrecht, Obtaining well calibrated probabilities using Bayesian binning into quantiles, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, 2015.