Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Bridging the Speed–Accuracy Gap: RT-DETR versus YOLOv8s for Real-Time Brain MRI Tumor Detection

Amir Mahdi Sedghi , ***Shala Nemati*** *

Department of Computer Engineering and Artificial Intelligence Research Institute, Shahrekord University, Shahrekord, Iran.

Article Info

Article History:
Received 03 November 2025
Reviewed 07 Januarys 2026
Revised 01 February 2026
Accepted 17 February 2026

Keywords:
Brain MRI
Brain tumor detection
Real-time object detection
Transformer based detection
YOLOv8s

*Corresponding Author's Email Address: *s.nemati@sku.ac.ir*

Abstract

Background and Objectives: Brain tumor detection in MRI images is critical for early diagnosis and effective treatment, yet manual interpretation is time-consuming and prone to variability. Deep learning models such as YOLO have advanced real-time object detection, but their speed–accuracy tradeoff remains a challenge for medical tasks involving small or low-contrast lesions. The potential of transformer-based detectors like RT-DETR to simultaneously improve accuracy and maintain real-time speed in clinical settings is not well established.
Methods: This study performed a controlled head-to-head comparison between the proposed model (RT-DETR-L-based model) and the YOLOv8s models using a curated, single-class brain tumor MRI dataset of 300 images. Both models were trained and evaluated under identical conditions with comprehensive data augmentation strategies, and their performance was assessed using standard object detection metrics including precision, recall, specificity, and mean Average Precision (mAP) across multiple Intersection over Union (IoU) thresholds.
Results: The proposed model achieved higher localization fidelity and overall accuracy compared to YOLOv8s, with mAP@0.5:0.95 of 0.493 versus 0.421 and mAP@0.5 of 0.963 versus 0.941. Precision and specificity for the proposed model reached 1.000, eliminating false positives, while recall was slightly lower than YOLOv8s (0.925 vs. 0.932), indicating a marginal increase in missed detections. Qualitative analysis confirmed robust detection across various tumor sizes and intensities, though some small or low-contrast lesions were missed.
Conclusion: Proposed model surpasses YOLOv8s in accuracy and specificity for real-time brain tumor detection in MRI images, offering a promising balance between speed and precision. However, its slightly lower recall underscores the need for further refinement to minimize false negatives. The findings suggest transformer-based detectors can narrow the speed–accuracy gap in medical imaging, but broader validation and optimization for resource-constrained environments are required for clinical deployment. Future work should focus on enhancing sensitivity and generalizability through advanced augmentation, larger datasets, and ensemble approaches.

How to cite this paper:
A. M. Sadeghi, S. Nemati, "Bridging the speed–accuracy gap: RT-DETR versus YOLOv8s for real-time brain MRI tumor detection," J. Electr. Comput. Eng. Innovations, 14(2): 311-326, 2026.
DOI: 10.22061/jecei.2026.12463.883
URL: https://jecei.sru.ac.ir/article_2543.html

Introduction

Brain tumors, though relatively rare compared with other systemic cancers, impose a disproportionate burden in terms of mortality, morbidity, and years of life lost, particularly in younger and middle-aged adults [1]. Prognosis strongly depends on timely and accurate characterization of lesions; delays or uncertainty in initial identification can adversely affect therapy planning and longitudinal monitoring [2]. Magnetic Resonance Imaging (MRI) is the primary non-invasive modality for brain tumor assessment because of its superior soft-tissue contrast and multi-parametric capability [3], [4]. However, manual interpretation and lesion localization are time-consuming, susceptible to inter-observer variability, and increasingly strained by rising imaging volumes and workforce limitations [5]. These factors motivate automated systems that can accelerate reliable triage, focus expert attention, and serve as consistent region-of-interest (ROI) proposal mechanisms.

The rapid progress of deep learning in general computer vision—spanning large-scale classification (e.g., ImageNet breakthroughs), object detection (Faster R-CNN, YOLO families), and instance segmentation (Mask R-CNN)—has catalyzed analogous advances in medical image analysis [6]-[13]. Benchmark datasets such as COCO and PASCAL VOC have driven architectural innovation and optimization for accuracy and latency [14]-[15]. These advances have translated to neuroimaging tasks including tumor detection and segmentation, where supervised convolutional neural networks (CNNs) and, more recently, self-configuring frameworks have set strong baselines [5], [16]-[18]. In clinical pipelines, detection (localizing candidate lesion regions) and segmentation (delineating precise boundaries) are complementary. Robust detection can act as a computationally efficient precursor to fine-grained volumetric segmentation or uncertainty estimation, reducing the search space and enabling adaptive resource allocation [16]-[18].

Despite their popularity for real-time applications, one-stage CNN detectors such as YOLO variants prioritize speed and may underperform on challenging medical scenarios characterized by small, low-contrast, or heterogeneous lesions embedded in complex anatomical backgrounds [11], [12], [19], [20]. Empirical studies and engineering analyses highlight an inherent speed–accuracy tension; systems optimized exclusively for maximal frame rate may sacrifice contextual reasoning critical for subtle pathology [21]-[24]. In clinical decision support, the acceptable operating point must balance rapid inference (to enable workflow integration, e.g., prospective triage or intra-session feedback) with high sensitivity and localization fidelity to minimize missed lesions while avoiding excessive false positives that erode trust and efficiency.

Transformer based vision models have recently redefined representational capacity by leveraging global self-attention to model long-range dependencies that conventional convolutions capture only implicitly or with deeper hierarchical stacking [16], [25], [26]. DETR introduced a set prediction paradigm for object detection, simplifying post-processing but initially incurring convergence and efficiency challenges [16]. Subsequent refinements—deformable attention, hierarchical backbones, and improved query selection—have narrowed the latency gap with optimized CNN detectors [18], [19]. RT-DETR exemplifies this trajectory by incorporating efficiency-oriented architectural choices to bring transformer detectors into the real-time regime while retaining the advantages of global context modeling and anchor-free design [19]. Concurrently, transformer or hybrid architectures have demonstrated promise across diverse medical imaging tasks, including segmentation and multi-modal fusion, suggesting potential gains for lesion detection in data-constrained, high-variance settings [25], [26].

Applying such architectures to brain MRI tumor detection is compelling for several reasons: (1) global attention can integrate dispersed anatomical cues and subtle intensity transitions; (2) query-based, anchor-free formulations may better accommodate scale variability without extensive hand-tuned priors; (3) improved box quality (higher localization precision across stricter overlap thresholds) can facilitate downstream segmentation refinement; and (4) if inference latency approaches that of optimized CNN baselines, clinical adoption barriers related to throughput diminish [19], [20], [25]. Nevertheless, it remains an open question whether modern real-time transformer detectors can simultaneously preserve near-YOLO speed and deliver measurable accuracy improvements within the constraints of modest, single-class medical datasets—conditions under which overfitting, limited diversity, and noisy annotations can blunt theoretical architectural advantages [5], [20].

While RT-DETR has been explored in other medical domains such as chest X-ray disease detection [27] and general medical object detection benchmarks [28], this work establishes the first focused evaluation of applying a recent real-time transformer detector (RT-DETR) specifically to Brain MRI tumor detection—and assesses its ability to outperform a strong, widely used convolutional baseline (YOLO) while maintaining clinically viable inference speed. Our contributions are: (i) introducing and validating the application of RT-DETR for brain MRI tumor detection as a novel use of a real-time transformer in medical imaging; (ii) a controlled

head-to-head comparison of RT-DETR versus a lightweight YOLO variant on a curated single-class brain tumor MRI detection task; (iii) a detailed analysis of the accuracy–latency trade-off, emphasizing localization quality (multi-threshold mean Average Precision) and sensitivity to subtle lesions; and (iv) a discussion of deployment implications for using transformer-based detectors as region-of-interest generators or decision support components within clinical workflows. By demonstrating how a transformer architecture can narrow the traditional speed–accuracy divide in neuro-oncologic detection, this study supports principled model selection and lays groundwork for expansion to multi-class, multi-institutional, and segmentation-augmented pipelines.

Related Works

Early computer-aided brain tumor analysis relied on hand-crafted feature pipelines coupled with classical classifiers. Bahadure et al. combined Berkeley Wavelet Transform (BWT)–based tissue segmentation with statistical texture descriptors and an SVM, achieving 96.51% accuracy and 97.72% sensitivity, illustrating the discriminative value of multi-scale wavelet representations yet inheriting dependence on sequential preprocessing (skull stripping, contrast enhancement) and limited scalability to heterogeneous MRI protocols [29]. Moving toward reduced manual feature dependence, Farnoosh and Noushkaran proposed an iterative co-clustering plus K-Means (ICCK) strategy on BraTS2019, reaching 99.28% accuracy but only 82.41% sensitivity, underscoring that high global accuracy can mask clinically critical false negatives when unsupervised blocks miss subtle tumor voxels [30].

The transition to deep learning introduced multi-stage and modular CNN pipelines. Gunasekara et al. integrated a classification CNN, region-based localization (R-CNN), and Chan–Vese active contours, obtaining a Dice score of 0.92 while demonstrating that classical variational refinement can still sharpen CNN coarse masks in limited data settings [31]. Veeramuthu et al. explored combined feature- and image-based classifiers (CFIC) to fuse handcrafted statistics with deep representations, improving robustness (98.97% accuracy; 98.86% sensitivity) but increasing architectural complexity and maintenance burden [32]. Mercaldo et al. then established a compact single-class brain MRI object detection baseline using a one-stage YOLO model on 300 images, reporting precision 0.943, recall 0.932, and mAP@0.5 0.941, thereby demonstrating that high real-time detection performance is achievable even under pronounced data scarcity [20]. Concurrently, lightweight U-Net derivatives, such as the pared architecture evaluated by Walsh et al. on BITE data (mean IoU 89%), showed that carefully pruned encoder–decoder designs can approach heavier models without aggressive augmentation when anatomical variability is moderate [33].

Recent progress centers on attention, multi-modal semantics, and refined detection heads. Transformer backbones and hierarchical self-attention, exemplified by Swin Transformer classification of four tumor categories (97% accuracy) [34], highlight improved global context modeling over purely convolutional inductive biases. Extending semantic fusion, a CLIP-guided 3D U-Net framework introduced multi-level language–vision alignment to raise whole-tumor Dice by 4.8% over a baseline 3D U-Net, evidencing the promise of cross-modal priors for ambiguous boundaries [35]. On the detection side, building on the small-data YOLO baseline of Mercaldo et al. [20], which reported strong mAP@0.5 (0.941) but lower multi-threshold mAP@0.5:0.95 (0.421), medical adaptations of YOLO incorporate specialized attention and loss refinements: YOLO-NeuroBoost reported mAP up to 99.48% on Br35H via dynamic kernel selection, CBAM attention, and Inner-GIoU loss emphasizing small, overlapping lesions [36]; an attention-driven YOLOv5m variant with Enhanced Spatial Attention reduced false positives in multi-type detection [37]. Parallel efforts tailor lightweight detection: MobileNetV2-SSD with a modified low-level FPN substantially boosted recall (~98%) for small meningioma loci while trading some precision (~89%) [38].

Efficiency constraints in resource-limited or point-of-care contexts have motivated edge-friendly detectors: a MobileNet-backboned RetinaNet variant attained balanced large vs. small lesion AP with markedly reduced computational load, demonstrating that architectural depth can be sacrificed if multi-scale feature fusion and focal optimization are preserved (preprint) [39].

Across these detection frameworks, a persistent tension emerges between maximizing recall for subtle, low-contrast foci and suppressing false positives that erode clinical trust—particularly acute in single-class triage workflows.

Given the chronic scarcity of labeled medical images, specialized augmentation and synthetic data have become pivotal. Lesion-aware compositing (CarveMix) improved segmentation generalization by explicitly modeling lesion context transitions [40], while GAN-based augmentation (StyleGANv2-ADA) markedly elevated radiogenomic classification accuracy across diverse datasets by enriching minority morphological patterns [41]. These approaches complement architectural advances, yet may not fully resolve the speed–accuracy trade-offs inherent in real-time detection pipelines.

Prior brain MRI work has either (i) emphasized segmentation quality over latency (transformer or 3D fusion methods), (ii) optimized lightweight CNN detectors sometimes at a sensitivity cost for diminutive or low-contrast tumors, or (iii) boosted sensitivity via augmentation/ensembles without systematically quantifying multi-threshold localization fidelity under real-time constraints; moreover, while Mercaldo et al. reported both mAP@0.5 and a substantially reduced mAP@0.5:0.95 (0.421), they did not benchmark against transformer-based detectors nor analyze architectural trade-offs under identical training regimes [20]. Direct, controlled head-to-head evaluation of a modern real-time transformer detector (RT-DETR) against a strong lightweight YOLO baseline on a constrained, single-class medical detection task remains underexplored. Our study addresses this gap by jointly analyzing speed-conscious architecture choice, strict IoU-stratified mAP, and the precision–recall balance, thereby clarifying whether global attention can narrow the historical speed–accuracy divide in clinically oriented brain tumor detection.

Method

This section presents detailed information about the dataset used, the data preparation procedure, and the data augmentation strategies applied to improve deep learning model performance in detecting and localizing brain tumors. The primary objective is to establish a sound basis for fair and scientific comparison between the proposed approach and existing methods—particularly the YOLO model—and to clarify the preprocessing and augmentation processes that enhance model generalizability.

A. Dataset

In this study, the same dataset employed in the reference YOLO-based brain cancer detection work was used [20]. The rationale for selecting this dataset is to ensure fair comparative conditions between the proposed model and YOLO so that the effect of architecture and processing methods can be evaluated without interference from extraneous variables. Using a shared dataset enables more precise analysis of performance differences and prevents bias arising from disparate data sources.

The dataset contains 300 brain MRI images, each representing a slice of the human brain. The images cover different anatomical orientations and exhibit substantial variation in spatial position, size, and intensity (contrast). Although cases originate from different underlying tumor pathologies (e.g., meningioma, pituitary adenoma, glioma), the provided annotations define a single detection class ‘tumor’; subtype labels are not part of the dataset [20]. The images are in JPG format, and the dataset is publicly accessible via the Roboflow platform [42]. Fig. 1 shows two sample images illustrating the morphological and dimensional diversity of the tumors.

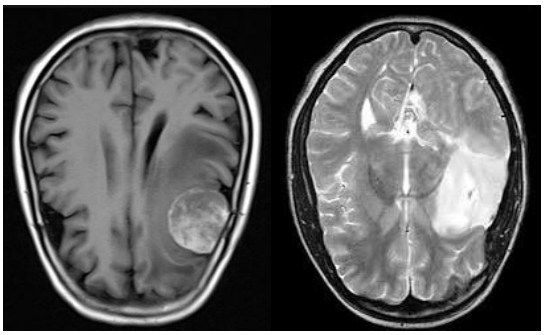

Fig. 1: MRI dataset samples showing variations in tumor morphology, size, and intensity.

B. Data Augmentation

One of the fundamental challenges in training deep learning models—especially in medical imaging—is the scarcity of labeled data [36]. Although a corpus of 300 images is acceptable for preliminary experimentation, it is typically insufficient for training a high-capacity deep network and may lead to model overfitting. To increase generalization capability and reduce the likelihood of overfitting, the use of data augmentation techniques during training is essential [43].

Data augmentation refers to a set of procedures that generate new, diverse samples by applying controlled, stochastic transformations with specified probabilities to the original images [44]. These transformations may include rotation, translation, scaling, random cropping, flipping, brightness and contrast adjustment, random noise injection, or combinations thereof. The central aim is to simulate diverse real-world conditions and enable the model to learn stable, invariant features insensitive to minor variations.

In this study, a comprehensive collection of augmentation techniques was applied to the training images to make the model more robust against spatial and photometric variations and to enhance its ability to generalize to unseen instances. As shown in Table 1, the augmentation strategy encompassed color space modifications (HSV hue ±1.5%, saturation ±70%, value ±40%), geometric transformations (50% horizontal flip probability, ±10% translation, ±50% scaling), and advanced techniques including 100% mosaic augmentation probability and 40% random erasing. This multi-faceted approach enabled the model to correctly detect tumors even when they differed from those seen during training. Overall, data augmentation is recognized as a key factor in the success of deep learning models in

medical image analysis and plays a critical role in performance improvement.

It is important to note that we utilized an on-the-fly augmentation pipeline provided by the Ultralytics framework. Unlike offline augmentation where a fixed set of images is generated prior to training, our approach applies stochastic transformations dynamically in memory during each training batch. Consequently, the model is exposed to a slightly different variation of the training samples in every epoch, effectively expanding the diversity of the dataset beyond the initial 300 images and mitigating overfitting risks.

Table 1: Data augmentation configuration for RT-DETR model training

Aug. Type	Enabled Tech.	Parameters
Color Space	HSV variations	H: ±1.5%, S: ±70%, V: ±40%
Geometric	Horizontal flip, Translation, Scaling	50% flip prob., ±10% translation, ±50% scaling
Advanced	Mosaic, Random erasing, RandAugment	100% mosaic prob., 40% erasing prob.
Disabled	Vertical flip, Rotation, Shear, Perspective, MixUp, CutMix, Copy-paste	N/A
Training Strategy	Close mosaic in final epochs	Last 10 epochs

C. Proposed Model

The RT-DETR-L (Real-Time Detection Transformer – Large) model is a recent object detection architecture first introduced by Baidu's research team in 2023 [19]. Its primary goal is to provide a transformer-based detection framework that combines the high accuracy of attention-driven models with practical speed suitable for real-time applications. RT-DETR is one of the earliest transformer-based, end-to-end, anchor-free detection models that addresses speed and complexity limitations of earlier architectures such as DETR [45].

The input to RT-DETR is an RGB image of fixed size. The output is a set of bounding boxes accompanied by class labels and confidence scores for each detected object.

This output structure is fully analogous to standard object detection models such as YOLO and Faster R-CNN, enabling direct performance comparison.

Fig. 2 provides a schematic overview of the proposed model. It comprises seven principal components:

1. Backbone: A high-capacity CNN producing a multi-scale pyramid of progressively lower-resolution, semantically enriched feature maps.
2. Multi-Scale Feature Maps: Hierarchically aligned tensors that preserve spatial correspondence while aggregating cross-level contextual semantics.
3. Efficient Hybrid Encoder: A lightweight module combining multi-scale attention with local mixing (feed-forward/convolutional) to capture long-range dependencies without sacrificing fine detail, all under real-time latency constraints.
4. IoU-Aware Query Selection: Quality-driven filtering that retains only top object queries (by predicted IoU/quality), suppressing low-value or redundant tokens before decoding to cut computation.
5. Decoder: Stacked layers apply cross-attention to encoded features and self-attention among the remaining queries, iteratively refining spatial localization and class discrimination.
6. Detection Head: Parallel heads mapping refined query embeddings to class logits and normalized box parameters (center x, y, width, height), often with iterative refinement across decoder layers.
7. Outputs: A compact, non-redundant set of class–box predictions with confidence/quality scores, typically usable directly (often NMS-free due to query selection) or in downstream post-processing pipelines.

During training, each component fulfills a distinct role.

The backbone extracts low-level and mid-level features; the Transformer module leverages attention to encode long-range dependencies and contextual information; and the detection head predicts bounding box coordinates and class membership probabilities based on the Transformer output.

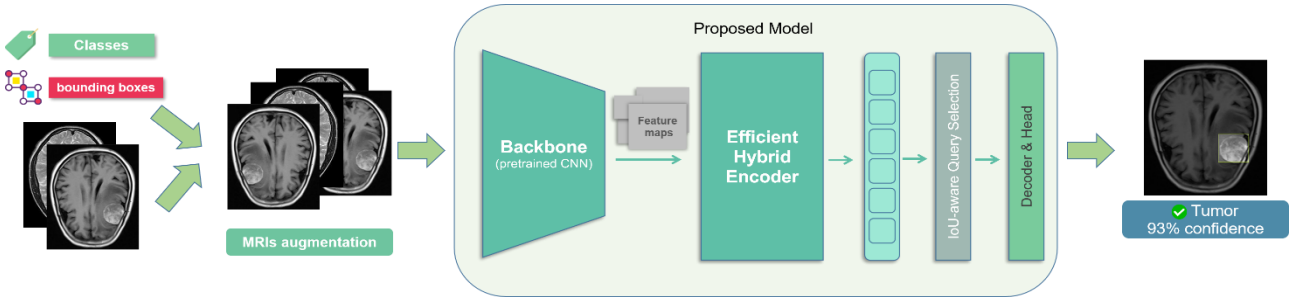

Fig. 2: Proposed model architecture diagram showcasing its main components.

Table 2: Layer-wise structure of proposed model and their functions.

layer	from	n	params	module	arguments
0	-1	1	25248	modules.block.HGStem	[3, 32, 48]
1	-1	6	155072	modules.block.HGBlock	[48, 48, 128, 3, 6]
2	-1	1	1408	modules.conv.DWConv	[128, 128, 3, 2, 1, False]
3	-1	6	839296	modules.block.HGBlock	[128, 96, 512, 3, 6]
4	-1	1	5632	modules.conv.DWConv	[512, 512, 3, 2, 1, False]
5	-1	6	1695360	modules.block.HGBlock	[512, 192, 1024, 5, 6, True, False]
6	-1	6	2055808	modules.block.HGBlock	[1024, 192, 1024, 5, 6, True, True]
7	-1	6	2055808	modules.block.HGBlock	[1024, 192, 1024, 5, 6, True, True]
8	-1	1	11264	modules.conv.DWConv	[1024, 1024, 3, 2, 1, False]
9	-1	6	6708480	modules.block.HGBlock	[1024, 384, 2048, 5, 6, True, False]
10	-1	1	524800	modules.conv.Conv	[2048, 256, 1, 1, None, 1, 1, False]
11	-1	1	789760	modules.transformer.AIFI	[256, 1024, 8]
12	-1	1	66048	modules.conv.Conv	[256, 256, 1, 1]
13	-1	1	0	modules.upsampling.Upsample	[None, 2, 'nearest']
14	7	1	262656	modules.conv.Conv	[1024, 256, 1, 1, None, 1, 1, False]
15	[-2, -1]	1	0	modules.conv.Concat	[1]
16	-1	3	2232320	modules.block.RepC3	[512, 256, 3]
17	-1	1	66048	modules.conv.Conv	[256, 256, 1, 1]
18	-1	1	0	modules.upsampling.Upsample	[None, 2, 'nearest']
19	3	1	131584	modules.conv.Conv	[512, 256, 1, 1, None, 1, 1, False]
20	[-2, -1]	1	0	modules.conv.Concat	[1]
21	-1	3	2232320	modules.block.RepC3	[512, 256, 3]
22	-1	1	590336	modules.conv.Conv	[256, 256, 3, 2]
23	[-1, 17]	1	0	modules.conv.Concat	[1]
24	-1	3	2232320	modules.block.RepC3	[512, 256, 3]
25	-1	1	590336	modules.conv.Conv	[256, 256, 3, 2]
26	[-1, 12]	1	0	modules.conv.Concat	[1]
27	-1	3	2232320	modules.block.RepC3	[512, 256, 3]
28	[21, 24, 27]	1	7303907	modules.head.RTDETRDecoder	[1, [256, 256, 256]]

Table 2 details the layer-wise architecture of proposed model (RT-DETR-based model); each layer performs a specific processing function. For example, convolutional layers in the backbone extract primary features, encoder and decoder layers in the Transformer model non-local relations, and linear layers in the head predict coordinates and classes.

Among RT-DETR-L's advantages over YOLOv8s is its attention-based architecture [19]. The attention mechanism enables RT-DETR-L to learn richer spatial and contextual relationships than purely convolutional networks like YOLOv8s, yielding improved performance in detecting small or overlapping objects. Both RT-DETR-L and YOLOv8s employ anchor-free designs—predicting

bounding boxes directly without predefined anchor templates. However, incorporation of attention in RT-DETR-L enhances accuracy, especially in more complex scenarios [19].

Regarding speed, RT-DETR employs Transformer optimizations such as dynamic heads and efficient attention to bring inference latency closer to that of YOLO models [46]. Although its speed is slightly lower than YOLOv8, this difference is generally negligible in medical applications where diagnostic accuracy supersedes raw throughput. Specifically, for brain tumor detection, priority lies with accuracy and reliability.

From an architectural technique perspective, both RT-DETR and YOLO utilize convolutional backbones for feature extraction; the principal divergence lies in the detection head and bounding box prediction mechanism. RT-DETR, by removing anchors and using dynamic queries, yields a simpler and more flexible structure and, through set-based loss (e.g., Hungarian matching), achieves more precise alignment between predictions and ground truth objects. Additionally, the use of Transformer attention facilitates modeling of global, non-local relationships that are not explicitly captured in YOLO, which emphasizes local features [19].

On the reference COCO dataset, RT-DETR-Large (RT-DETR-L) has achieved higher mAP than YOLOv8s. For example, RT-DETR-L attains mAP = 53.0% on COCO val2017, whereas YOLOv8s achieves mAP = 44.9% under the same evaluation settings [19], [47]. This performance gap primarily stems from RT-DETR's use of transformer-based attention mechanisms that more effectively capture global context and its IoU-aware query selection and advanced loss formulations, which together improve both localization and classification precision compared to the convolution-only backbone and loss strategies employed by YOLOv8s.

As illustrated in Fig. 3, the proposed model (the large variant of RT-DETR (RT-DETR-L)) comprises approximately 33 million parameters, which is substantially higher than the 11 million parameters of YOLOv8s [47]. While RT-DETR-L incorporates structural optimizations and efficient attention mechanisms that partially alleviate its computational demands, the model nonetheless necessitates significantly greater computational resources for both training and maintenance compared to YOLOv8s. This increased resource requirement may present challenges for deployment in environments with limited computational capacity.

A key strength of proposed model (RT-DETR-based model) is its ease of fine-tuning [48]. Like YOLO, it is fully supported by the Ultralytics package and can be readily retrained on new datasets. Fine-tuning allows the model to adapt to dataset-specific characteristics, improving performance in specialized tasks such as brain tumor detection.

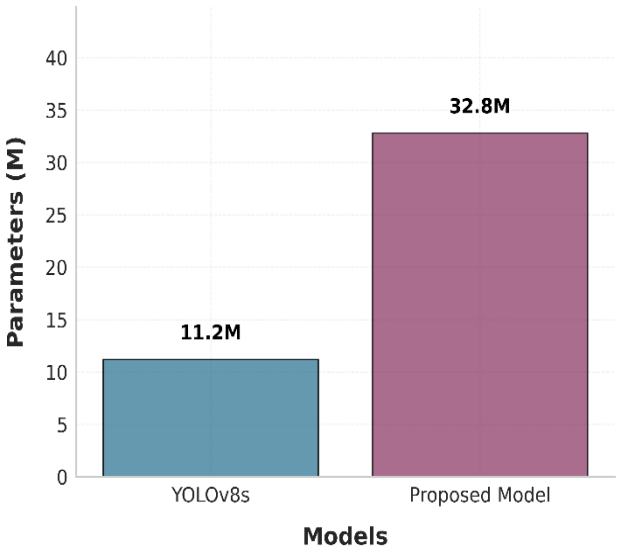

Fig. 3: Parameter comparison.

In medical imaging, model selection must consider accuracy, processing speed, and memory usage [6]. While RT-DETR-L delivers high accuracy in identifying and delineating complex objects, its larger parameter count and computational demands may challenge real-time deployment or use in hardware-limited systems (e.g., edge devices or portable units). However, in clinical and research centers equipped with sufficient infrastructure, this model can significantly enhance diagnostic precision and reliability. Overall, operational adoption of RT-DETR-L in medical applications depends directly on hardware availability, required response time, and the relative importance of accuracy; in scenarios where accuracy outweighs speed and memory constraints, proposed model (RT-DETR-L-based model) is a suitable option. However, optimizing the model's memory footprint through techniques such as quantization or pruning is essential to enable deployment in resource-constrained environments, such as portable medical units or edge devices.

In the test phase, the trained RT-DETR model processes an MRI image as input and outputs the tumor bounding box coordinates, class label, and confidence score for each prediction. These outputs can then be used for model performance evaluation or to support clinical decision-making.

The proposed model was trained in the Kaggle environment on an NVIDIA Tesla P100 GPU (CUDA-enabled) with 16 GB RAM using the default Ultralytics configuration (batch size 16, input resolution 640×640, 50 epochs) and the SGD optimizer with an initial learning rate of 0.01. After each epoch, performance was

evaluated on a held-out validation set to track optimization progress. The single checkpoint achieving the highest validation means Average Precision (mAP) over the 50 epochs was retained and then used once to generate test-set predictions. The test set remained completely unseen during training and model selection, ensuring an unbiased estimate of generalization and limiting overfitting.

Results and Discussion

In this section, we examine and analyze the performance of the proposed model in comparison with the YOLOv8s model for brain tumor detection and localization in MRI images. First, the evaluation metrics employed in this study are introduced; then the quantitative and analytical results of the models are presented.

Finally, the strengths and weaknesses of the models and future recommendations are discussed.

In Fig. 4, the normalized confusion matrix of proposed model on the test data is shown.

The confusion matrix provides a visualization of the distribution of the model's predictions across different classes and was specifically computed using the same fixed IoU threshold of 0.5 and a confidence threshold of 0.25 as YOLOv8s, ensuring a consistent basis for direct comparison between the two models. As observed, proposed model correctly identified 97% of tumor samples, which is 2% higher than the value obtained using YOLOv8s [20]. These thresholds were chosen to align the evaluation metrics and facilitate a fair comparison of classification performance, while filtering out low-confidence detections. Additionally, proposed model demonstrated a very low false positive rate, indicating its strong ability to distinguish between tumorous and healthy regions.

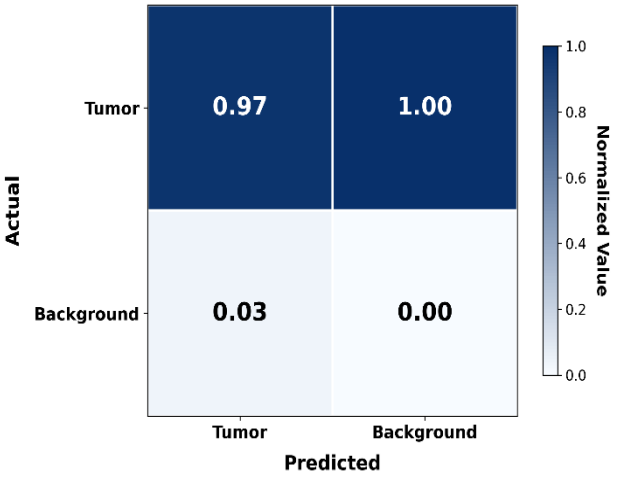

Fig. 4: Normalized confusion matrix of proposed model predictions on test data.

A. Evaluation Metrics

In this study, to evaluate and compare model performance, standard object detection metrics were used, including Precision, Recall (Sensitivity), Specificity, and mean Average Precision (mAP). The confusion matrix was computed using a fixed IoU threshold of 0.5, while the values reported in the results chart in Fig. 5 were calculated using aggregated thresholds across IoU values ranging from 0.5 to 0.95. This distinction is necessary to account for the difference in recall values observed between the confusion matrix and the results table.

The mathematical definitions of these metrics are given below.

Precision is a metric that indicates, among all regions the model has labeled as tumor, how many are truly tumor.

This is especially important in medical applications, because higher precision means fewer false positives, which reduces unnecessary patient anxiety, additional costs, and unwarranted treatments. It is computed as shown in:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

where TP (True Positive) is the number of correctly detected positive samples and FP (False Positive) is the number of incorrectly detected positive samples.

Recall (Sensitivity) measures the model's ability to correctly identify all positive samples (true tumors). It reflects how many real tumors the model has not missed.

The high importance of recall in medicine is due to the need to minimize false negatives, thereby reducing the likelihood of overlooking real tumors and preventing irreversible consequences for the patient. It is computed as:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where FN (False Negative) is the number of missed positive samples.

Specificity is the complementary metric to recall and shows how well the model correctly identifies healthy (non-tumorous) samples. High specificity means fewer false positives and helps avoid unnecessary treatments. It is computed as:

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

where TN (True Negative) is the number of correctly identified negative samples.

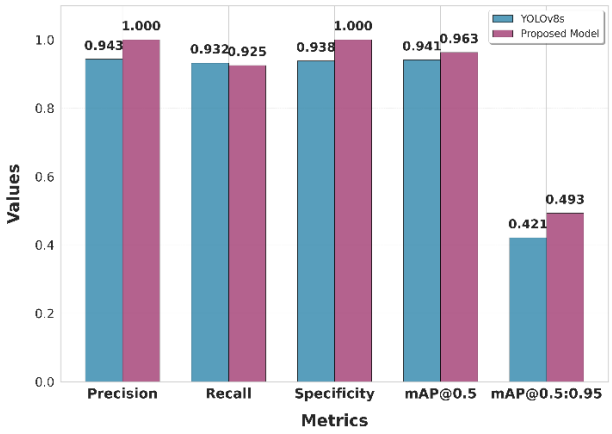

Fig. 5: Comparison of proposed model and YOLOv8s performance metrics.

While precision and recall each measure a single aspect of performance, in practical applications a balance between them is usually required. Therefore, the Precision–Recall curve is commonly used to analyze performance at different decision thresholds.

Average Precision (AP) is the area under the Precision–Recall curve for a given class and provides a combined measure of precision and recall. In other words, AP tells us how well the model performs across all threshold values rather than at a single operating point.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{4}$$

where N is the number of classes and AP_i is the AP for class i.

In this study, mAP was computed in two forms: mAP@0.5 and mAP@0.5:0.95. In the first case, predictions are considered correct when the Intersection over Union (IoU) between the predicted region and the ground-truth region is at least 0.5. IoU measures the overlap between the predicted bounding box and the ground-truth tumor region and is defined as:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{5}$$

The higher the IoU, the more accurate the spatial localization of the model. In the mAP@0.5:0.95 setting, the mean mAP is computed over IoU thresholds from 0.5 to 0.95 in steps of 0.05 to evaluate the model not only at lower thresholds but also at stricter localization accuracies. This metric provides a more comprehensive and realistic picture of the model's performance in precise tumor detection and localization.

In the Precision–Recall curve shown in Fig. 6 corresponding to the proposed model, the AP (mAP for the tumor class) at IoU ≥ 50% is approximately 0.93. The curve was generated using a confidence threshold of 0.001, ensuring that all potential detections were considered during evaluation. This curve aggregates performance across varying IoU thresholds (0.5–0.95), providing a broader view of the model's ability to balance precision and recall at different localization accuracies.

More complete numerical evaluation results of the models are provided in Fig. 5. According to this chart, proposed model outperformed YOLOv8s in both mAP@0.5:0.95 and mAP@0.5.

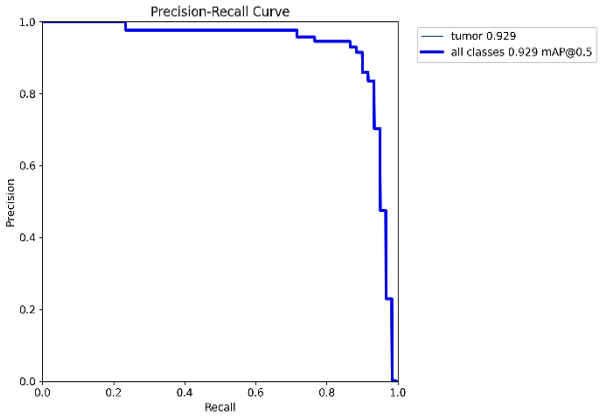

Fig. 6: Precision–Recall curve of proposed model across varying IoU thresholds.

Values reported in Fig. 5 correspond to aggregated thresholds (IoU@0.5:0.95) and a confidence threshold of 0.001 to provide a comprehensive evaluation, whereas values in the confusion matrix reflect a fixed IoU threshold of 0.5 and a confidence threshold of 0.25. Specifically, mAP@0.5:0.95 for proposed model was 0.493 compared to 0.421 for YOLOv8s. Also, mAP@0.5 for RT-DETR-L was 0.963 versus 0.941 for YOLOv8s [20]. These results indicate that proposed model achieved higher localization fidelity and accuracy across stricter IoU thresholds.

To validate the real-time capabilities of the models, we measured the inference latency on an NVIDIA GPU. RT-DETR-L achieved an average inference time of 40.59 ms per image (approx. 24.6 FPS), whereas YOLOv8s achieved 8.58 ms per image (approx. 116.6 FPS). Although RT-DETR-L is computationally heavier due to its transformer blocks, its processing speed remains well within the acceptable range for real-time clinical support systems (typically requiring >20 FPS for smooth visualization), confirming that the substantial accuracy gains do not come at a prohibitive cost to throughput.

However, YOLOv8s achieved slightly higher recall (0.932 vs. 0.925), demonstrating a marginal advantage in sensitivity. This tradeoff highlights that while proposed model effectively reduces false positives, further

refinement is needed to minimize false negatives, which are critical in medical imaging applications.

The proposed model demonstrated zero false positives in the evaluated test set, indicating high precision and specificity. However, its slightly lower recall compared to YOLOv8s (0.925 vs. 0.932) suggests a residual risk of missed tumor detections, particularly for subtle or low-contrast lesions.

In medical imaging, recall is critical to ensure comprehensive detection of tumors, as false negatives can lead to delayed diagnosis and treatment. While proposed model effectively minimizes false positives, further refinement is needed to address false negatives and improve recall without compromising precision. Techniques such as ensemble modeling, hard example mining, and post-processing methods could be explored to enhance sensitivity to subtle lesions.

In comparison of other metrics, both Precision and Specificity for proposed model equaled 1.000, indicating the absence of false positives in the test data. This is particularly important in medical applications, as reducing false positives avoids unnecessary patient anxiety and medical costs. Conversely, the Recall of YOLOv8s was slightly higher than that of proposed model (0.932 vs. 0.925), indicating that YOLOv8s was marginally more sensitive in some instances; however, this gain in sensitivity came with an increase in false positives.

It should be noted that rerunning this experiment may produce slightly different numerical outcomes (<1% variation) due to stochastic factors in weight initialization and the dynamic data augmentation process. However, the performance gap and the ranking between RT-DETR-L and YOLOv8s remained consistent across repeated runs, confirming the robustness of the reported improvements.

B. Impact of False Negatives

False negatives in medical imaging refer to missed detections of actual tumors, which can have severe consequences for patient outcomes. In clinical practice, missing even a small percentage of tumors may result in delayed diagnosis, progression of the disease, and reduced effectiveness of treatments. For example, subtle lesions with low contrast or small size are more likely to be overlooked, especially in noisy or complex imaging scenarios. While proposed model reduces false positives, the presence of false negatives highlights the need for further refinement to ensure comprehensive detection. Improving recall is particularly critical in screening workflows, where early detection is essential to prevent disease progression.

C. Proposed Solutions to Improve Recall

To address the recall limitation, several techniques can be implemented:

1) Ensemble Modeling: Combining predictions from multiple models can improve sensitivity by capturing subtle patterns missed by individual models.
2) Hard Example Mining: Prioritizing challenging samples during training, such as low-contrast or small tumors, can help the model focus on difficult cases and improve recall.
3) Post-Processing Methods: Applying post-processing techniques, such as confidence threshold adjustments or secondary verification steps, can reduce the risk of missed detections.
4) Data Augmentation: Enhancing the diversity of training data through advanced augmentation techniques (e.g., simulating low-contrast or noisy conditions) can improve the model's ability to detect subtle lesions.

By implementing these strategies, proposed model can achieve a better balance between precision and recall, ensuring comprehensive tumor detection while maintaining high accuracy.

D. Training and Validation Analysis

In Fig. 5, the training and evaluation trends of proposed model for the single-class MRI tumor detection task are analyzed based on the plots produced during training. To rigorously assess model performance, a set of quantitative indicators including loss values and accuracy-related metrics on both the training and validation sets were examined.

Based on the presented plots, analysis of proposed model during training shows effective convergence on brain MRI data for tumor detection. The overall objective function of proposed model consists of three main components: Box Loss, Classification Loss, and Distribution Focal Loss (DFL). The Box Loss, responsible for regression of bounding box coordinates, gradually decreased from an initial value of 2.0 to about 1.1, indicating appropriate learning of object localization. The Classification Loss followed a consistent downward trend from 3.0 to 0.5, reflecting improved discriminative capability for tumor class identification. The DFL Loss, used to refine the probability distribution of bounding box boundary predictions, decreased from 2.0 to 1.2. On the validation set, all loss types stabilized after initial fluctuations: val/box_loss around 1.9, val/cls_loss about 0.8, and val/dfl_loss near 1.8—confirming the absence of overfitting and adequate generalization.

Subsequently, precision and recall, as the primary indicators of tumor detection performance, were analyzed. The increase in precision throughout training indicates a reduction in false positive rate and improved reliability in correctly identifying tumors. Similarly, the increase in recall reflects the model’s strong ability to capture all tumor instances (reduction in false negatives).

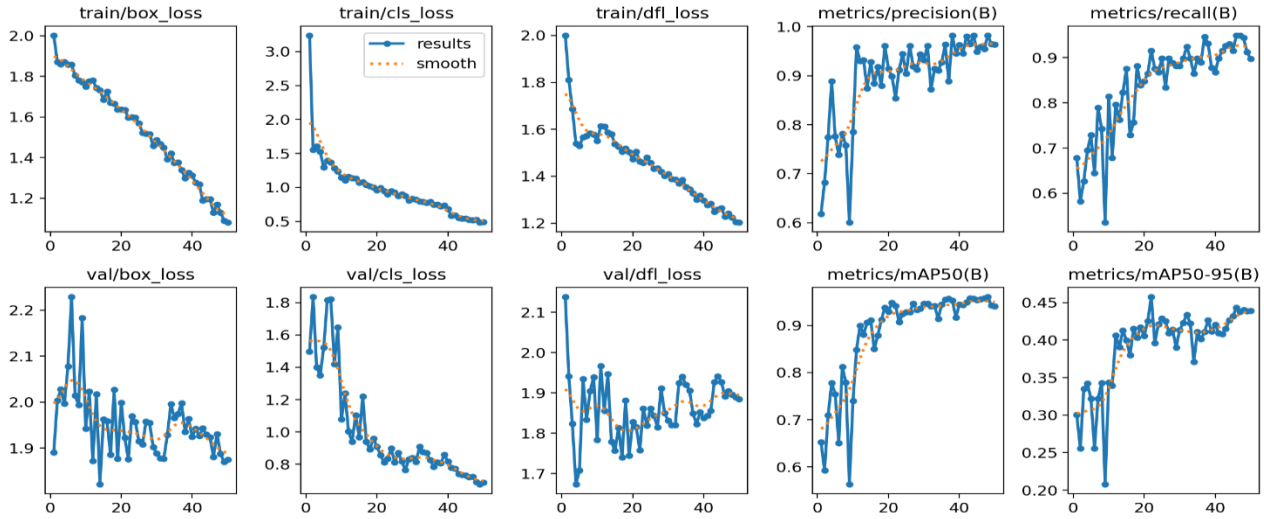

Fig. 5: The experimental analysis results.

Moreover, mean Average Precision at both mAP@0.5 and mAP@0.5:0.95 was examined as comprehensive performance measures across varying IoU overlap thresholds. The rapid rise and subsequent stabilization of these metrics during training demonstrate acceptable performance in accurately detecting and localizing tumor regions in MRI images.

Overall, the declining trend of loss functions and the increasing trend of accuracy metrics and mAP values in both training and validation sets confirm successful learning and proper convergence of proposed model for the single-class MRI tumor detection task. These results highlight the high potential of this model for practical and research applications in medical imaging.

In the following images, sample outputs of proposed model on validation data are shown.

Figure 8 presents the model's predictions while Fig. 9 shows the ground-truth labels. Moreover, mean Average Precision at both mAP@0.5 and mAP@0.5:0.95 was examined as comprehensive performance measures across varying IoU overlap thresholds. The rapid rise and subsequent stabilization of these metrics during training demonstrate acceptable performance in accurately detecting and localizing tumor regions in MRI images. Qualitative assessment of these outputs shows that the proposed model can identify tumor regions with high confidence (often above 0.8). Furthermore, the model did not exhibit sensitivity to tumor size, intensity, or position, and maintained stable performance across different tumor types (small, large, varying contrast). This can also be observed in the dispersion of predicted bounding box coordinates and dimensions in Fig. 10.

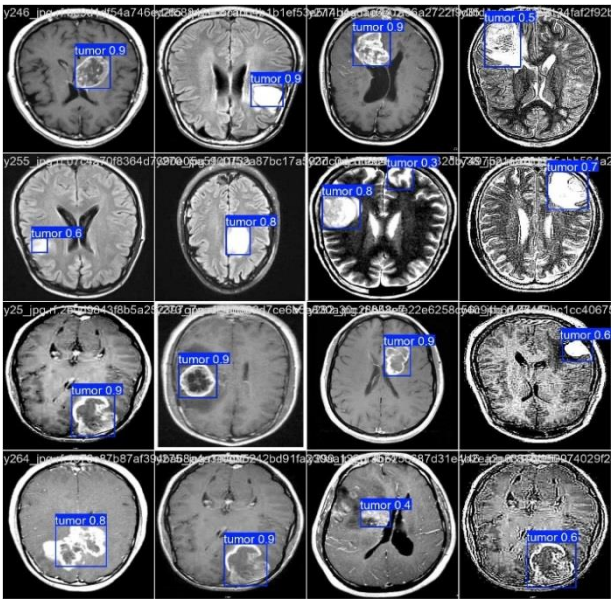

Fig. 8: Predicted tumor regions by proposed model on validation.

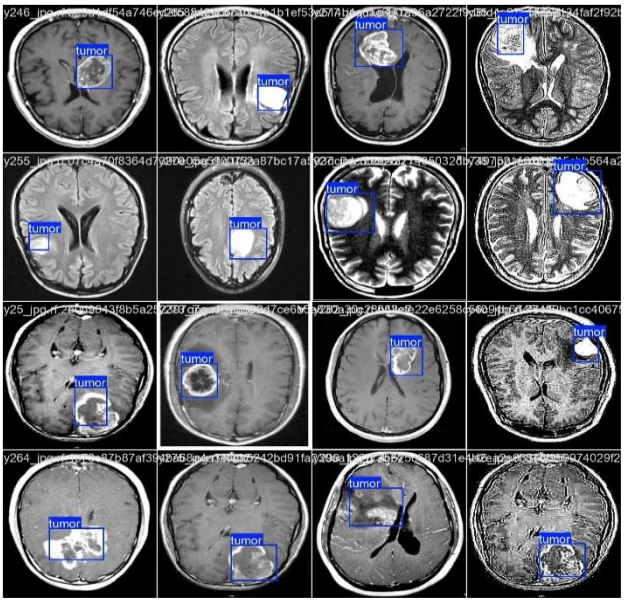

Fig. 9: Ground-truth tumor annotations for validation data.

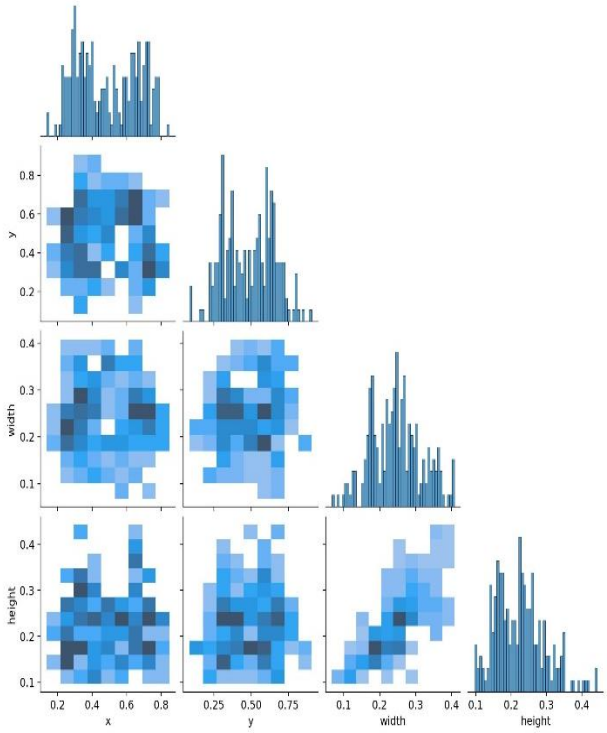

Fig. 10: Correlogram of bounding box predictions for validation data.

E. Strengths, Limitations, and Future Directions

From a technical standpoint, the proposed model, by leveraging a Transformer architecture and attention mechanisms, was able to extract deeper and more complex feature representations from MRI images. The use of advanced loss functions such as Generalized IoU Loss and L1 Loss enabled more precise bounding box regression and improved convergence during training. This advantage was especially evident in detecting small tumors or low-contrast cases, where the convolution-based YOLOv8s exhibited lower accuracy and a higher false negative rate.

While the proposed model achieves superior precision and specificity, its slightly lower recall compared to the YOLOv8s baseline indicates a remaining challenge in detecting all subtle lesions. To address this, future work will focus on strategies specifically designed to improve sensitivity without compromising the model's excellent false-positive rate. Techniques such as confidence threshold tuning, hybrid loss functions (e.g., incorporating focal loss components [36]), and hard example mining during training could direct the model's capacity toward challenging, low-contrast, or small tumors. Furthermore, a promising avenue is the development of a hybrid ensemble framework that leverages the high-precision nature of the RT-DETR-L model alongside a complementary high-recall detector. The final prediction could be derived through a weighted or meta-learning strategy, optimizing for a clinical operating point that balances both critical metrics effectively.

In conclusion, the proposed model, by leveraging the Transformer architecture and modern loss functions, achieved higher accuracy and robustness than YOLOv8s in detecting and localizing brain tumors in MRI images. The values reported in the confusion matrix and results table reflect different IoU and confidence thresholds (IoU@0.5 and confidence@0.25 for the confusion matrix; aggregated IoU thresholds@0.5–0.95 and confidence@0.001 for mAP@0.5:0.95), ensuring a fair and comprehensive comparison. Nevertheless, to further enhance performance and clinical applicability, employing higher-quality data, regularization techniques, and investigating the impact of pre-training on related datasets are advised. Moreover, more detailed error analysis and improving annotation quality can play a crucial role in increasing model accuracy under real-world conditions.

Conclusion

This study set out to rigorously evaluate whether a recent Transformer-based real-time detector (RT-DETR-L) can surpass a widely used convolutional baseline (YOLOv8s) for brain MRI tumor detection while retaining near real-time practicality, albeit with a higher memory footprint. Focusing on a single-class lesion detection task, the findings demonstrate that the architectural shift from purely convolutional feature pyramids to a hybrid encoder–decoder attention mechanism yields a measurable gain in both localization fidelity and classification reliability. Quantitatively, the proposed model (RT-DETR-L-based model) improved mAP@0.5:0.95 from 0.421 to 0.493 (≈7% relative increase) and mAP@0.5 from 0.941 to 0.963 (≈2% relative increase), evidencing not only tighter bounding boxes across stricter IoU thresholds but also more consistent high-IoU performance. Precision and Specificity reached 1.000, eliminating false positives in the evaluated test set—an attribute of high clinical utility because it minimizes unnecessary patient anxiety, repeat imaging, and downstream cost.

The tradeoff was a marginal decline in Recall (0.932 to 0.925), highlighting a residual risk of missed subtle lesions. While proposed model eliminates false positives, its slightly lower recall underscores the need for further refinement to minimize false negatives, which are critical in medical imaging applications. False negatives can lead to delayed diagnosis and treatment, particularly for subtle or low-contrast lesions. To address this limitation, techniques such as ensemble modeling, hard example mining, and advanced data augmentation should be explored to improve sensitivity without compromising precision. By implementing these strategies, the

proposed model can achieve a better balance between precision and recall, ensuring reliable and comprehensive tumor detection.

Qualitative inspection corroborated the numerical metrics: the proposed model demonstrated promising performance across lesion size variability, non-central anatomical locations, and diminished lesion-to-parenchyma contrast.

However, these findings are based on a limited dataset of 300 MRI images, and further validation on larger, more diverse datasets is necessary to confirm the model's robustness across different imaging conditions and tumor types.

This advantage is plausibly attributable to global attention facilitating long-range contextual reasoning and improved integration of multi-scale cues, especially for small or low-intensity anomalies that can evade purely local receptive fields.

Assuming (as hypothesized but to be explicitly confirmed in deployment studies) that measured inference latency remains close to that of YOLOv8s, these improvements suggest that accuracy gains need not come at the expense of practical clinical throughput, though increased GPU memory consumption may constrain deployment on lower-resource systems without further optimization.

Clinically, the configuration that yields zero false positives can streamline radiologist workflow by prioritizing review of high-confidence detections rather than filtering spurious marks; nevertheless, even a modest number of false negatives—particularly involving very small, infiltrative, or noise-obscured tumors—underscores the necessity for continued refinement before independent diagnostic use. The present scope introduces several limitations: (1) restriction to a single tumor class precludes assessing performance in multi-pathology differential contexts; (2) potential dataset size, class imbalance, and scanner heterogeneity constraints may limit generalizability; (3) reliance on existing annotations could cap attainable upper bounds if label noise or boundary imprecision is present; (4) external validation and prospective testing were not yet performed, leaving domain shift resilience unproven; and (5) the higher computational demand of RT-DETR-L compared to lightweight CNNs.

This presents a challenge for clinical deployment on edge devices or portable MRI scanners with limited hardware resources.

While high-end hospital workstations can easily handle RT-DETR-L, widespread deployment in resource-constrained environments will require future work on model compression techniques, such as quantization, pruning, or knowledge distillation, to balance the superior accuracy of Transformers with the efficiency required for embedded clinical systems.

The translational pathway should therefore prioritize: expanding and diversifying cohorts across institutions, vendors, field strengths, and acquisition protocols; employing advanced augmentation (intensity harmonization, simulation of low-SNR or motion artifacts) and curriculum or hard-example mining to emphasize rare small lesions; leveraging large-scale pre-training (medical self-supervision or mixed natural/medical corpora) followed by domain adaptation; calibrating confidence scores and modeling epistemic/aleatoric uncertainty for risk-aware triage; applying efficiency optimizations (quantization, pruning, knowledge distillation, TensorRT acceleration) to secure latency headroom and reduce memory footprint; exploring ensemble or two-stage refinement (coarse global detection plus local high-resolution verification); generating interpretable attention and saliency visualizations to bolster clinician trust; and conducting external, prospective, and, ultimately, regulatory-aligned evaluations (including reproducibility, drift monitoring, and failure case taxonomies).

To address the recall limitation, future work should explore techniques such as ensemble modeling, hard example mining, or post-processing methods to capture subtle tumor instances that may currently be missed. Additionally, expanding the dataset and improving annotation quality can help the model learn to detect challenging cases more effectively.

In summary, this work serves as a preliminary engineering validation and a controlled architectural comparison.

We have demonstrated that on a constrained single-class detection task, the proposed RT-DETR-L model delivers superior localization accuracy and eliminates false positives compared to the YOLOv8s baseline, particularly in reducing false positives and achieving higher precision and specificity. However, its slightly lower recall (0.925 vs. 0.932 for YOLOv8s) highlights the need for further refinement to minimize false negatives, which are critical in medical imaging applications. This tradeoff underscores the importance of balancing precision and recall to ensure comprehensive and reliable tumor detection.

It is crucial to emphasize that the limited scale and scope of the dataset restrict the immediate clinical generalizability of these findings. Therefore, the results should be interpreted as a promising proof-of-concept. Future work must prioritize expansion to larger, multi-class, and multi-institutional datasets to assess robustness across diverse clinical scenarios and confirm the translational potential of this architectural approach.

Author Contributions

A. M. Sedghi performed all primary scientific and operational tasks: initial ideation (together with the supervisor), data preparation and organization, coding and implementation, execution of experiments, result analysis, creation of figures and visualizations, and drafting the original manuscript. Dr. Shahla Nemati provided overall supervision, contributed guidance to the conceptualization and methodological design, critically reviewed and scientifically edited the manuscript, validated the findings, supplied resources and scholarly support, and managed project progression (project administration).

Acknowledgment

The authors gratefully acknowledge Shahrekord University.

Funding

This research received no external funding.

Conflict of 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.

Abbreviations

AP	Average Precision
CNN	Convolutional Neural Network
COCO	Common Objects in Context (Dataset)
DFL	Distribution Focal Loss
IoU	Intersection over Union
mAP	mean Average Precision
MRI	Magnetic Resonance Imaging
NMS	Non-Maximum Suppression
RT-DETR-L	Real-Time Detection Transformer (Large variant)
SGD	Stochastic Gradient Descent
YOLO	You Only Look Once (Object Detection Family)
YOLOv8s	YOLO Version 8 – Small Variant
Glou	Generalized IoU

References

[1] Q. T. Ostrom et al., "CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2015–2019," Neuro-Oncology, 24, (Supplement_5): v1–v95, 2022.

[2] D. N. Louis et al., "The 2021 WHO classification of tumors of the central nervous system: A summary," (in eng), Neuro Oncol, 23(8): 1231–1251, 2021.

[3] S. Bauer, R. Wiest, L. P. Nolte, M. Reyes, "A survey of MRI-based medical image analysis for brain tumor studies," Phys. Med. Biol., 58(13): R97, 2013.

[4] S. Cha, "Update on brain tumor imaging: From anatomy to physiology," Am. J. Neuroradiol., 27(3): 475, 2006.

[5] B. H. Menze et al., "The multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," IEEE Trans. Med. Imaging, 34(10): 1993–2024, 2015.

[6] G. Litjens et al., "A survey on deep learning in medical image analysis," Med. Image Anal., 42: 60–88, 2017.

[7] A. S. Lundervold, A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," Zeitschrift für Medizinische Physik, 29(2): 102–127, 2019.

[8] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012.

[9] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[10] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Proc. Advances in Neural Information Processing Systems 28 (NIPS 2015) 2015.

[11] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You only look once: unified, real-time object detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June, 2016.

[12] A. Bochkovskiy, C. Y. Wang, H. Y. Liao, "yolov4: optimal speed and accuracy of object detection," arXiv:2004.10934, 2020.

[13] K. He, G. Gkioxari, P. Dollar, R. Girshick, "Mask R-CNN," presented at the IEEE International Conference on Computer Vision (ICCV), Oct, 2017.

[14] T. Y. Lin et al., "Microsoft COCO: Common objects in context," in Proc. Computer Vision – ECCV 2014, Cham, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., 2014// 2014: Springer International Publishing, pp. 740–755.

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, "The Pascal Visual Object Classes (VOC) challenge," Int. J. Comput. Vision, 88(2): 303–338, 2010.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, "End-to-end object detection with transformers," in Computer Vision – ECCV 2020, Cham, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., 2020// 2020: Springer International Publishing, pp. 213–229.

[17] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[18] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," arXiv preprint arXiv:2010.04159, 2020.

[19] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, C. Cui, Y. Du, Q. Dang, Y. Liu, "DETRs beat YOLOs on real-time object detection," arXiv:2304.08069, 2023.

[20] F. Mercaldo, L. Brunese, F. Martinelli, A. Santone, M. Cesarelli, "Object detection for brain cancer detection and localization," Appl. Sci., 13(16), 2023.

[21] J. Huang et al., "Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July, 2017.

[22] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nat. Med., 25(1): 44–56, 2019.

[23] F. Pesapane, M. Codari, F. Sardanelli, "Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine," Eur. Radio. Exp., 2(1): 35, 2018.

[24] A. Esteva et al., "Deep learning-enabled medical computer vision," npj Digital Med., 4(1): 5, 2021.

[25] K. He et al., "Transformers in medical image analysis," Intell. Med., 3(1): 59–78, 2023.

[26] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Cham, A. Crimi and S. Bakas, Eds., 2022// 2022: Springer International Publishing, pp. 272–284.

[27] R. Mutukuru, A. Rajesh, V. Ponduri, J. Ahammed, L. P. Kothala, "Transformer-based RT-DETR framework for accurate chest X-Ray disease detection," IRBM, 46(6): 100912, 2025.

[28] W. He, Y. Zhang, T. Xu, T. An, Y. Liang, B. Zhang, "Object Detection for Medical Image Analysis: Insights from the RT-DETR Model," arXiv:2501.16469, 2025.

[29] N. B. Bahadure, A. K. Ray, H. P. Thethi, "Image analysis for MRI based brain tumor detection and feature extraction using biologically inspired BWT and SVM," Int. J. Biomed. Imaging, 2017(1): 9749108, 2017.

[30] R. Farnoosh, H. Noushkaran, "Application of a modified combinational approach to brain tumor detection in MR images," J. Digital Imag., 35(6): 1421–1432, 2022.

[31] S. R. Gunasekara, H. N. T. K. Kaldera, M. B. Dissanayake, "A systematic approach for MRI brain tumor localization and segmentation using deep learning and active contouring," J. Healthcare Eng., 2021(1): 6695108.

[32] A. Veeramuthu et al., "MRI brain tumor image classification using a combined feature and image-based classifier," Front. Psychol. Original Res., 13, 2022.

[33] J. Walsh, A. Othmani, M. Jain, S. Dev, "Using U-Net network for efficient brain tumor segmentation in MRI images," Healthcare Anal., 2: 100098, 2022.

[34] A. A. Asiri et al., "Advancing brain tumor detection: harnessing the Swin Transformer's power for accurate classification and performance analysis," PeerJ Comput. Sci., 10: e1867, 2024.

[35] M. Zhang, "A brain tumor segmentation method based on CLIP and 3D U-Net with cross-modal semantic guidance and multi-level feature fusion," arXiv:2507.09966v1, 2025.

[36] A. Chen, D. Lin, Q. Gao, "Enhancing brain tumor detection in MRI images using YOLO-NeuroBoost model," Frontiers Neurol., Original Res., 15, 2024.

[37] S. Muksimova, S. Umirzakova, S. Mardieva, N. Iskhakova, M. Sultanov, Y. I. Cho, "A lightweight attention-driven YOLOv5m model for improved brain tumor detection," Comput. Biol. Med., 188: 109893, 2025.

[38] N. F. Hikmah, A. D. Hajjanto, A. F. A. Surbakti, N. A. Prakosa, T. Asmaria, T. A. Sardjono, "Brain tumor detection using a MobileNetV2-SSD model with modified feature pyramid network levels," Int. J. Electr. Comput. Eng. (IJECE), 14(4): 10, 2024.

[39] A. Abdusalomov et al., "Accessible AI diagnostics and lightweight brain tumor detection on medical edge devices," Bioengineering, 12(1): 62, 2025.

[40] X. Zhang et al., "CarveMix: A simple data augmentation method for brain lesion segmentation," NeuroImage, 271: 120041, 2023.

[41] M. M. E. Yurtsever, Y. Atay, B. Arslan, S. Sagiroglu, "Development of brain tumor radiogenomic classification using GAN-based augmentation of MRI slices in the newly released gazi brains dataset," BMC Med. Inf. Decis. Making, 24(1): 285, 2024.

[42] Brain Tumor Dataset. [Online].

[43] J. Nalepa, M. Myller, M. Kawulok, "Training- and test-time data augmentation for hyperspectral image segmentation," IEEE Geosci. Remote Sens. Lett., 17: 292–296.

[44] A. B. Abdusalomov, M. Mukhiddinov, T. K. Whangbo, "Brain tumor detection based on deep learning approaches and magnetic resonance imaging," Cancers, 15(16): 4172, 2023.

[45] M. Wu, Y. Qiu, W. Wang, X. Su, Y. Cao, Y. Bai, "Improved RT-DETR and its application to fruit ripeness detection," Frontiers Plant Sci., 16, 2025.

[46] E. L. T. Jun, M. L. Tham, B. H. Kwan, "A comparative analysis of RT-DETR and YOLOv8 for urban zone aerial object detection," in Proc. 2024 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS): 340–345.

[47] Ultralytics. "YOLOv8 Models Documentation," [Online].

[48] W. Lv, Y. Zhao, Q. Chang, K. Huang, G. Wang, Y. Liu, "RT-DETRv2: improved baseline with bag-of-freebies for real-time detection transformer," arXiv:2407.17140, 2024.

Biographies

Amir Mahdi Sedghi was born in Isfahan, Iran, on 24 February 2002. He received the B.Sc. degree in Computer Engineering from Shahid Chamran University of Ahvaz, Ahvaz, Iran, in September 2024. He is currently pursuing the M.Sc. degree in Computer Engineering (Artificial Intelligence) at Shahrekord University, Shahrekord, Iran, since October 2024. His academic interests span medical data processing computational neuroscience, natural language processing, and computer vision. He is particularly interested in applying AI techniques to analyze biomedical and neural data. His research interests include medical data analysis, neuroscience-inspired AI, natural language processing, and computer vision.

- Email: asedghi1380@gmail.com
- ORCID: 0009-0001-3532-1395
- Web of Science Researcher ID: OCL-7029-2025
- Scopus Author ID: N/A
- Homepage: N/A

Shahla Nemati was born in Shiraz, Iran, in 1982. She received the B.Sc. degree in Hardware Engineering from Shiraz University, Shiraz, in 2005, the M.Sc. degree from the Isfahan University of Technology, Isfahan, Iran, in 2008, and the Ph.D. degree in Computer Engineering from Isfahan University, Isfahan, in 2016. Since 2017, she has been an Assistant Professor with the Computer Engineering Department, Shahrekord University, Shahrekord, Iran. She has written several articles in the fields of data fusion, emotion recognition, affective computing, and audio processing. Her research interests include data fusion, affective computing, and data mining.

- Email: s.nemati@sku.ac.ir
- ORCID: 0000-0003-2906-5871
- Web of Science Researcher ID: AAA-3341-2019
- Scopus Author ID: 24512475100
- Homepage: https://www.sku.ac.ir/~snemati#