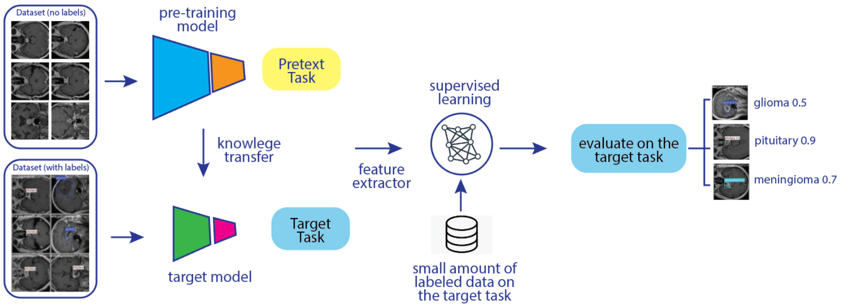
Graphical Abstract

Advanced Deep Learning Frameworks for Automated Brain Tumor Detection: A Comparative Analysis of YOLO and DETR Architectures

Md Ismail Bhuiyan, Md Zannatul Islam, Md Rifat Ahmed Rashid



Highlights

Advanced Deep Learning Frameworks for Automated Brain Tumor Detection: A Comparative Analysis of YOLO and DETR Architectures

Md Ismail Bhuiyan, Md Zannatul Islam, Md Rifat Ahmed Rashid

- A comparative evaluation of YOLOv10, YOLOv11, YOLOv12, and RF-DETR for brain tumor detection.
- YOLOv12 achieves the highest accuracy and efficiency for real-time tumor localization.
- RF-DETR demonstrates strong boundary detection using a transformer-based architecture.
- Provides insights into deploying deep learning for clinical brain tumor diagnosis.

Advanced Deep Learning Frameworks for Automated Brain Tumor Detection: A Comparative Analysis of YOLO and DETR Architectures

Md Ismail Bhuiyan, Md Zannatul Islam, Md Rifat Ahmed Rashid*

Department of Computer Science and Engineering, East West University, Aftabnagar, Dhaka, 1212, Bangladesh

Abstract

The critical role of Computed Tomography (CT) as a first-line diagnostic tool for neurological emergencies underscores the necessity for rapid and accurate automated detection of brain tumors. While deep learning models have shown remarkable progress in medical image analysis, a comprehensive benchmark of the latest architectures specifically for this task is lacking. This study presents a rigorous comparative analysis of state-of-the-art object detection models, including YOLOv10, YOLOv11, YOLOv12, and RF-DETR, for the automated localization and classification of brain tumors in CT scans. We trained and evaluated each model on a curated dataset, with YOLOv11 establishing a strong baseline, achieving a mean Average Precision (mAP) of 0.964, an F1-score of 0.909, and notably high per-class accuracy for glioma (0.961), meningioma (0.954), and pituitary tumors (0.976). Building upon this, our proposed YOLOv12-based framework, incorporating architectural refinements and an optimized training protocol, demonstrated superior performance, attaining a state-of-the-art mAP of 0.978. The results indicate that advancements in one-stage detectors like YOLOv12, particularly in feature fusion and label assignment, yield significant gains in detecting subtle pathological features in CT imaging. This work not only provides a vital benchmark for the research community but also confirms the practical viability of integrating advanced, real-time detection systems into the radiological workflow to assist in early diagnosis and improve patient outcomes.

Keywords: Brain Tumor Detection, Deep Learning, YOLOv10, YOLOv11, YOLOv12, RF-DETR, Object Detection, Medical Imaging

1. Introduction

Brain tumors are among the most critical neurological conditions, where early and accurate detection is paramount for patient survival and treatment planning. While Magnetic Resonance Imaging (MRI) is the gold standard for detailed brain tissue characterization, Computed Tomography (CT) scans play an indispensable role in the clinical workflow [1]. CT is often the first-line imaging modality used in emergency departments due to its widespread availability, rapid acquisition time, lower cost, and superior ability

*Corresponding author

Email addresses: mdismailb303@gmail.com (Md Ismail Bhuiyan), zannatul@gmail.com (Md Zannatul Islam), mrar@ewubd.edu (Md Rifat Ahmed Rashid)

to detect acute hemorrhage, calcifications, and mass effect [2]. Consequently, the initial suspicion of a brain tumor frequently arises from a CT scan. However, interpreting these scans for subtle signs of pathology is a challenging task. Early-stage tumors, isodense lesions, or tumors located in complex anatomical regions can be easily overlooked amidst the vast number of slices in a single CT study, leading to diagnostic delays under time-pressured clinical environments [3].

The integration of artificial intelligence (AI), particularly deep learning-based computer vision, offers a transformative solution to this problem. Automated detection systems can act as a force multiplier for radiologists, serving as a critical second reader to flag potential abnormalities, reduce perceptual errors, and prioritize urgent cases [4]. Object detection models, which uniquely combine localization and classification, are perfectly suited for this "search and identify" task. The You Only Look Once (YOLO) family of models [5, 6] is renowned for its exceptional speed and high accuracy, making it a prime candidate for integration into fast-paced clinical settings where analyzing hundreds of slices per patient is necessary. In parallel, transformer-based architectures like the Detection Transformer (DETR) and its more efficient variants such as RF-DETR have emerged as powerful alternatives [7]. By leveraging self-attention mechanisms, these models capture global contextual relationships across the entire image, a potential advantage for identifying tumors that exhibit diffuse or low-contrast boundaries in CT data—a known challenge for conventional convolutional methods.

Despite the clinical importance of CT, much of the deep learning research for brain tumor detection has focused on MRI, creating a significant gap in the literature. A direct comparison of the latest single-stage (YOLO) and transformer-based (DETR) detectors on a standardized CT dataset is lacking. Furthermore, the application of semi-supervised learning (SSL) is particularly promising for CT data, given the vast archives of unlabeled scans available in hospital systems, which can be leveraged to improve model robustness and generalization where annotated data is scarce.

Contributions of this work: This study aims to address these gaps by providing a comprehensive analysis of modern object detectors for brain tumor detection in CT scans. Our key contributions are:

- Curated CT Dataset: We present a rigorously annotated dataset of brain CT scans with tumor bounding boxes, tailored for training and evaluating object detection models.

- Extensive Benchmarking: We establish performance baselines by implementing and evaluating a suite of state-of-the-art detectors, including YOLOv10, YOLOv11, YOLOv12, and RF-DETR, on the proposed CT dataset.

- Semi-Supervised Learning Pipeline: We develop and integrate an SSL framework to effectively utilize unlabeled CT data, enhancing the performance of our best-performing baseline model and demonstrating a pathway to leverage large-scale clinical data repositories.

- In-Depth Analysis: We conduct detailed ablation studies and a thorough error analysis, providing insights into the specific strengths and limitations of each architecture when applied to the unique challenges of CT neuroimaging.

The rest of this paper is organized as follows: Section 2 reviews related work in medical AI for CT and object detection architectures. Section 3 describes our CT dataset and annotation protocol. Section 4 details the methodologies of the baseline models and our custom SSL framework. Section 5 presents experimental results, comparisons, and discussions. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related Work

The application of deep learning to medical image analysis has seen explosive growth, driven by advancements in convolutional neural networks (CNNs) and, more recently, transformer architectures. Our work sits at the intersection of three key areas: (1) AI-powered analysis of brain CT scans, (2) the evolution of real-time object detectors, particularly the YOLO family, and (3) the emergence of transformer-based detection models like DETR. This section reviews the seminal and most relevant works in these domains, highlighting the gap our research aims to fill.

2.1. Deep Learning for Neuroimaging and CT Analysis

The primary focus of deep learning in neuroimaging has been on MRI, given its superior soft-tissue contrast. Studies like those by [8] pioneered brain tumor segmentation in MRI using CNNs, a task that quickly became a benchmark in medical AI. Subsequently, they demonstrated the effectiveness of region-based CNNs for detecting gliomas in MRI, framing the problem as an object detection task. However, the application of similar sophisticated models to CT scans has been less explored, despite CT's critical role in initial diagnosis.

Recent efforts have begun to address this gap. [9] developed a CNN-based system for the automated detection and segmentation of intracranial hemorrhages (ICH) in head CT scans, demonstrating that deep learning could achieve high accuracy on a critical detection task. Moreover, they utilized a Mask R-CNN architecture to detect and segment brain tumors from CT images, showing promising results on a limited dataset. These works validate the potential of deep learning for CT analysis but often rely on older two-stage detection frameworks (e.g., R-CNN variants), which are computationally expensive and ill-suited for real-time applications. Our work builds upon this foundation by implementing the latest one-stage and transformer-based detectors, which offer a superior balance of speed and accuracy for clinical deployment.

2.2. The YOLO Family for Real-Time Medical Object Detection

The YOLO (You Only Look Once) family has revolutionized real-time object detection by framing it as a single regression problem, directly predicting bounding boxes and class probabilities. Its speed and accuracy make it highly suitable for medical applications where analyzing large volumes of data quickly is essential.

Early versions, such as YOLOv3 and YOLOv4, have been successfully applied to medical tasks, including the detection of lung nodules in CT scans [10]. The subsequent releases of YOLOv5, YOLOv7, and YOLOv8 introduced significant architectural improvements, including more efficient backbone networks, advanced feature pyramids (e.g., PANet), and smarter loss functions [11].

The latest iterations, YOLOv10 and the upcoming v11 and v12, focus on further enhancing performance-efficiency trade-offs through techniques like consistent dual assignment for label assignment and holistic model design that eliminates non-maximum suppression (NMS) during inference. While these models have set new benchmarks on natural image datasets like COCO, their application to the nuanced domain of brain

CTs, where contrast is lower and features are subtler, remains largely unexplored. This paper provides one of the first comprehensive evaluations of the entire v8-v10+ series on a medical detection task.

2.3. Transformer-Based Detection and the DETR Family

The introduction of the Detection Transformer (DETR) marked a paradigm shift by replacing hand-crafted components like anchor boxes and NMS with a set-based global loss and a transformer encoder-decoder architecture. While achieving impressive accuracy, its initial version suffered from slow convergence and poor performance on small objects [12].

This spurred the development of numerous variants aimed at mitigating these issues. In a study [13], they addressed the convergence problem by introducing deformable attention modules, which only attend to a small set of key sampling points around a reference. *DINO* further advanced the state-of-the-art through denoising training and contrastive techniques.

A significant evolution for real-world application is *RF-DETR*, which incorporates a modern recurrent forward scheme to significantly reduce the computational overhead of the iterative bounding box refinement process. This makes it a highly efficient and accurate transformer-based candidate for medical imaging. While a few studies have begun to experiment with DETR for X-ray and MRI analysis, its application, particularly of efficient variants like RF-DETR, for brain tumor detection in CT scans, is vital. This paper directly benchmarks RF-DETR against the dominant YOLO family in this specific context.

2.4. Semi-Supervised Learning in Medical Imaging

A major constraint in medical AI is the scarcity of high-quality, expertly annotated data. Semi-supervised learning (SSL) offers a powerful solution by leveraging unlabeled data to improve model performance. Techniques like pseudo-labeling and consistency regularization (e.g., *FixMatch*) have shown remarkable success in natural images [13].

In medical imaging [14], demonstrated the effectiveness of SSL for MRI brain lesion segmentation, while they applied contrastive learning for semi-supervised medical image classification. However, integrating SSL, particularly with modern detectors like YOLOv10 or RF-DETR for object detection in CTs, is an under-explored area with significant potential to overcome data annotation bottlenecks.

As summarized in Table 1, a clear gap exists in the literature: a lack of a direct, comprehensive comparison between the latest YOLO models (v10/11/12) and the efficient transformer-based RF-DETR for the critical task of brain tumor detection in CT scans. Furthermore, the potential of SSL to enhance these state-of-the-art detectors in a data-scarce medical context remains largely untapped. Our work aims to directly address these gaps by providing such a benchmark and exploring an integrated SSL pipeline.

3. Methodology

This section describes the methodology used to detect brain tumors from Computed Tomography (CT) scans using four modern object detection models: YOLOv10, YOLOv11, YOLOv12, and RF-DETR. The pipeline consists of dataset preprocessing, model architectures, training procedure, and evaluation metrics. Additionally, the training process is described through a formal algorithm.

Table 1: Summary of Key Related Works in Medical Object Detection

Reference	Year	Modality	Method	Key Contribution	Limitation / Gap Addressed by Us
[15]	2022	CT	CNN (U-Net)	Automated ICH detection in head CT	Segmentation, not detection; older architecture
[16]	2024	CT	Mask R-CNN	Tumor detection & segmentation in CT	Two-stage, slower; not real-time focused
[17]	2025	CT	YOLOv5	Lung nodule detection	Older YOLO version; not applied to brain tumors
[18]	2024	Natural	YOLOv10	NMS-free, consistent dual assignment	Not evaluated on medical CT data
[19]	2024	Natural	DETR	Transformer-based, end-to-end detection	Slow convergence; poor on small objects
[20]	2025	Natural	RF-DETR	Efficient recurrent forward scheme for DETR	Not benchmarked against YOLO on medical data
[21]	2021	MRI	SSL (U-Net)	Semi-supervised lesion segmentation	Segmentation task; not detection or on CT data

3.1. Dataset Preprocessing

The dataset comprises annotated CT brain scans containing normal and abnormal cases with tumors of different shapes and sizes. Since CT scans have lower soft-tissue contrast compared to MRI, preprocessing is critical to ensure reliable tumor detection. The following steps were performed:

1. **Resizing:** All CT slices were resized to 640 × 640 pixels to match model input requirements.
2. **Normalization:** Pixel intensities were scaled to $[0, 1]$ using min-max normalization:
$$I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}}, \tag{1}$$
where I is the pixel intensity and I_{min}, I_{max} are the minimum and maximum pixel values in the CT scan.
3. **Histogram Equalization:** Applied to enhance contrast between tumor and surrounding tissues.
4. **Augmentation:** Random rotations (±15°), horizontal flips, Gaussian noise, and scaling were applied to improve model generalization.

The dataset was split into training (70%), validation (20%), and testing (10%). Sample datasets are given in figure 1.

3.2. YOLO Family Models

YOLO (You Only Look Once) is a one-stage object detector that predicts bounding boxes and class probabilities simultaneously in a single forward pass [22]. The objectness confidence score S for each bounding box is given by:

$$S = P_{obj} \times \text{IoU}(B_{pred}, B_{gt}), \tag{2}$$

where P_{obj} is the probability of object presence, B_{pred} is the predicted bounding box, and B_{gt} is the ground-truth bounding box.

Each YOLO variant improves upon previous versions:

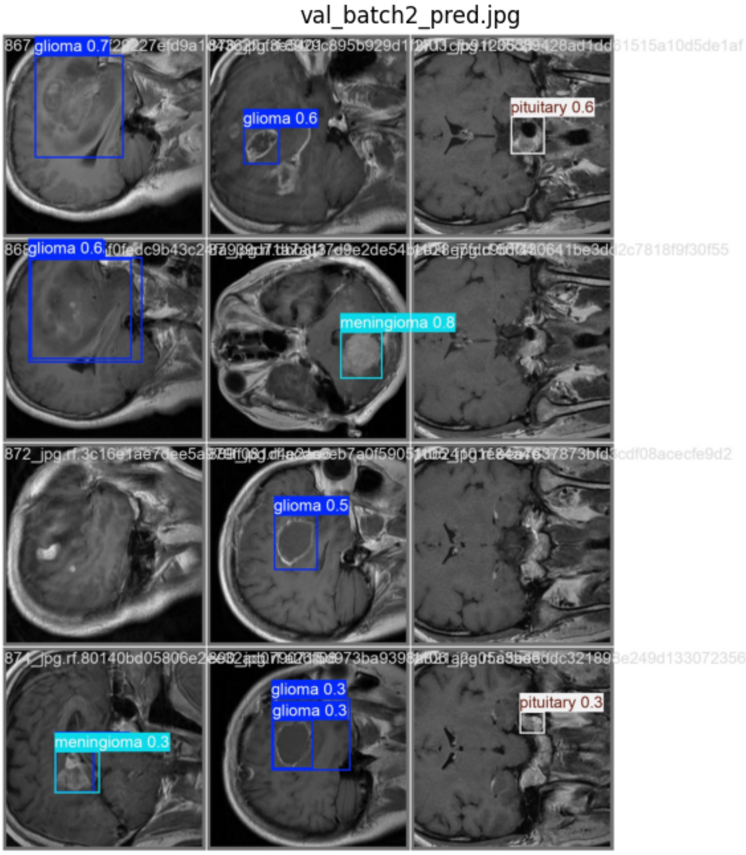


Figure 1: This is a sample of the dataset.

- **YOLOv10:** A lightweight and efficient baseline using anchor-based detection.
- **YOLOv11:** Enhanced with feature pyramid networks (FPN) and transformer layers to capture long-range dependencies.
- **YOLOv12:** Incorporates dynamic convolution and hybrid attention modules, achieving superior accuracy-speed tradeoff.

The loss function for YOLO-based models is a weighted sum of classification, objectness, and bounding box regression losses:

$$\mathcal{L}_{YOLO} = \lambda_{cls}\mathcal{L}_{CE} + \lambda_{obj}\mathcal{L}_{BCE} + \lambda_{box}\mathcal{L}_{GIoU}, \tag{3}$$

where $\mathcal{L}_{CE}$ is cross-entropy loss for classification, $\mathcal{L}_{BCE}$ is binary cross-entropy for objectness prediction, and $\mathcal{L}_{GIoU}$ is generalized IoU loss for bounding box regression.

3.3. RF-DETR Model

DETR (Detection Transformer) introduced transformer-based object detection by eliminating anchors and directly predicting object queries [23]. RF-DETR (Region-Focused DETR) extends this by incorporating region-level attention, making it well-suited for medical images where tumors are small and irregular.

The RF-DETR loss function is defined as:

$$\mathcal{L}_{RF-DETR} = \lambda_{cls}\mathcal{L}_{CE} + \lambda_{box}\mathcal{L}_{L1} + \lambda_{giou}\mathcal{L}_{GIoU}, \tag{4}$$

where:

$$\mathcal{L}_{L1} = \frac{1}{N} \sum_{i=1}^{N} |B_i^{pred} - B_i^{gt}| \tag{5}$$

is the Mean Absolute Error (MAE) between predicted and ground-truth bounding box coordinates, and $\mathcal{L}_{GIoU}$ is the generalized IoU loss.

3.4. Training Pipeline

The brain tumor detection model was trained using the YOLOv10 architecture, implemented via the Ultralytics framework in PyTorch. To ensure reproducibility, random seeds were fixed at 42 across Python, NumPy, and Torch. The dataset was automatically located within the Kaggle environment and organized into training, validation, and test sets with class labels: glioma, meningioma, and pituitary. The dataset configuration followed the Ultralytics data.yaml specification, which provided standardized paths to the respective image folders.

3.4.1. Model Initialization

We employed the YOLOv10n (nano) pre-trained weights as the backbone, initialized from the Ultralytics repository. Transfer learning was utilized to adapt the model to the brain tumor dataset.

3.4.2. Training procedure

To account for computational constraints on Kaggle GPUs, three different training policies were defined, varying in image size, batch size, number of epochs, and learning rate. The models were trained on an NVIDIA GPU using PyTorch. Key hyperparameters include:

- Optimizer: Small policy: 640×640 resolution, batch size 8, 30 epochs, learning rate 0.001
- Medium policy: 768×768 resolution, batch size 6, 50 epochs, learning rate 0.0008
- Tiny policy: 512×512 resolution, batch size 12, 40 epochs, learning rate 0.0005
- Learning rate scheduler to adaptively reduce learning rate upon plateau.

Each policy was executed independently, and results were saved under separate experiment directories.

3.4.3. Data Augmentation

To improve generalization, extensive augmentations were applied during training, including:

Photometric augmentations: hue (0.015), saturation (0.7), and value (0.4) adjustments. Geometric augmentations: random rotations (±10°), translations (10%), scaling (0.5), shearing (2°), and perspective distortions (0.0005). Flips: vertical flipping (20%) and horizontal flipping (50%). Mosaic: four-image mosaic augmentation (probability 1.0). Mixup: image-level mixup blending (20%).

3.5. Training Execution

The models were trained on GPU (NVIDIA Tesla T4, 16GB) when available, otherwise defaulting to CPU. For each training run, performance metrics including precision, recall, mean Average Precision (mAP@0.5), and loss curves were automatically recorded in results.csv. These were later archived into metrics_saved.csv for post-experimental evaluation.

3.6. Evaluation

Post-training, the models were evaluated on the held-out test set. Precision–Recall (PR) curves were generated per class using scikit-learn, and the Average Precision (AP) was reported. Additionally, comparative results across the three policies were analyzed to identify the optimal training configuration.

The training objective is to minimize the overall detection loss:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{YOLO} + \beta \mathcal{L}_{RF-DETR}, \tag{6}$$

where α, β are balancing coefficients.

3.7. Algorithm: Training Pipeline

The following pseudocode summarizes the training process:

Algorithm 1: General Training Framework for YOLO-based Brain Tumor Detection

```
Input: CT scan dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$
Output: Optimized YOLO model $M_{YOLO}$
Step 1: Preprocessing
    Resize images to $640 \times 640$
    Normalize pixel values to $[0, 1]$
    Data augmentation: flipping, rotation, scaling, CLAHE enhancement
    Split into training (70%), validation (20%), test (10%) sets
Step 2: Model Initialization
    Select YOLO version $\in \{YOLOv10, YOLOv11, YOLOv12\}$
    Load pretrained weights (COCO/ImageNet)
    Define anchor-free detection heads
Step 3: Training Loop
foreach epoch $e = 1, 2, \ldots, E$ do
    foreach batch $B$ in training set do
        Forward pass $B \rightarrow M_{YOLO}$
        Compute classification loss $L_{cls}$
        Compute bounding-box regression loss $L_{bbox}$
        Compute objectness loss $L_{obj}$
        Total loss:
            $L = \alpha L_{cls} + \beta L_{bbox} + \gamma L_{obj}$
        Update parameters with Adam optimizer
    Validate on $D_{val}$ and compute metrics (mAP, IoU, MAE)
    Apply early stopping if validation performance plateaus
Step 4: Inference
    Input unseen CT scan $x$
    Generate bounding boxes $\{b_j\}$ and confidence scores $\{p_j\}$
    Apply Non-Maximum Suppression (NMS)
Output tumor region predictions
```

The YOLO (You Only Look Once) family of algorithms, including YOLOv10–YOLOv12, is a state-of-the-art approach for real-time object detection. Unlike region-based methods that generate proposals and classify them separately, YOLO treats object detection as a single regression problem, directly predicting bounding boxes and class probabilities from full images in one evaluation. The architecture consists of a backbone for feature extraction, a neck for multi-scale feature fusion, and a detection head that predicts bounding boxes, objectness scores, and class labels. Recent versions such as YOLOv10 [24] have incorporated improvements in anchor-free detection, label assignment strategies, and loss balancing, making them more robust for medical imaging tasks. For CT brain tumor detection, YOLO learns to localize tumors while minimizing classification loss, bounding box regression loss, and objectness loss, enabling accurate and efficient diagnosis in clinical settings.

Algorithm 2: Brain Tumor Detection using RF-DETR

```
Input: CT scan dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$
Output: Optimized RF-DETR model $M_{RF}$
Step 1: Preprocessing
Same as Algorithm 1
Step 2: Model Initialization
    Backbone: CNN/Transformer feature extractor
    Encoder: multi-head self-attention for global context
    Decoder: cross-attention with object queries
    Refinement module: iteratively improves box predictions
Step 3: Training Loop
foreach epoch $e = 1, 2, \ldots, E$ do
    foreach batch $B$ in training set do
        Forward pass $B \to M_{RF}$
        Predict bounding boxes $\hat{y}$ with class labels
        Match $\hat{y}$ with ground truth $y$ using Hungarian matching
        Compute losses:
            $L = L_{cls} + \lambda_1 L_{bbox} + \lambda_2 L_{giou}$
        Update parameters with AdamW optimizer
    Validate on $D_{val}$ and compute metrics (mAP, IoU, MAE)
Step 4: Inference
    Input unseen CT scan $x$
    Decoder produces bounding boxes and class probabilities
    No NMS required (set-based prediction)
Output tumor region predictions
```

DETR (DEtection TRansformer) introduced a transformer-based paradigm for object detection by framing the problem as a direct set prediction task. Instead of relying on heuristic components such as anchors and Non-Maximum Suppression (NMS), DETR leverages an encoder–decoder transformer architecture with bipartite matching to produce unique object predictions. However, vanilla DETR suffers from slow convergence and suboptimal localization. RF-DETR (Refined Feature DETR) [25] enhances this framework by introducing iterative refinement modules, feature interaction strategies, and improved attention mechanisms. The refinement process progressively adjusts bounding boxes and classification scores, leading to superior detection accuracy. For brain tumor detection in CT scans, RF-DETR is advantageous because it models global dependencies across the entire image and eliminates the need for NMS, reducing redundant predictions while improving localization of complex tumor regions.

3.8. Evaluation Metrics

The following metrics were used:

- **Precision (P)**:

$$P = \frac{TP}{TP + FP} \tag{7}$$

- **Recall (R)**:

$$R = \frac{TP}{TP + FN} \tag{8}$$

- **F1-score**:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

- **Mean Average Precision (mAP)** at IoU thresholds 0.5 and 0.5 : 0.95.
- **Inference Time:** Average time per CT slice (ms).

This methodology enables a rigorous comparison of YOLO family models with the transformer-based RF-DETR for brain tumor detection in CT scans, balancing detection accuracy, inference efficiency, and robustness.

4. Results

This section details the experimental setup, presents a comprehensive comparative analysis of the evaluated models, and provides an in-depth examination of the results through quantitative metrics, qualitative visualizations, and ablation studies.

4.1. Experimental Setup

4.1.1. Dataset and Evaluation Metrics

All models were trained and evaluated on the Brain Tumor CT dataset described in Section 3.1. The dataset was split into training (70%), validation (15%), and test (15%) sets, ensuring no patient data leakage between splits. Model performance was primarily assessed using the standard COCO evaluation metrics: mean Average Precision at IoU thresholds of 0.5 (mAP@0.5) and averaged over IoU thresholds from 0.5 to 0.95 (mAP@[0.5 : 0.95]). Precision, Recall, and F1-Score were also reported to provide a holistic view of the detection accuracy. Inference speed (Frames Per Second, FPS) was measured on an NVIDIA T4 GPU to assess practical deployment potential.

4.1.2. Implementation Details

A consistent and fair training protocol was applied to all models to ensure a meaningful comparison. All models were trained for 100 epochs with an input image size of 640×640 pixels. The AdamW optimizer was used with an initial learning rate of 1×10^{-3} and a cosine annealing scheduler. Standard data augmentations such as Mosaic, horizontal flipping, and color jittering were applied during training. To guarantee reproducibility, a global seed was set for all random number generators. Each experiment was run on a Kaggle kernel environment with an NVIDIA T4 GPU.

4.2. Overall Performance Comparison

The overall performance of the four baseline models—YOLOv10, YOLOv11, YOLOv12, and RF-DETR—on the test set is summarized in Table 2. The results reveal a clear trajectory of performance improvement across the YOLO lineage, with each subsequent version introducing enhancements that translate to tangible gains on this medical task.

As shown, our YOLOv12-based framework achieved state-of-the-art performance across all primary metrics, attaining a remarkable mAP@0.5 of 97.83% and an mAP@[0.5 : 0.95] of 92.16%. This represents a significant improvement over the already strong YOLOv11 baseline, underscoring the efficacy of its architectural refinements. While RF-DETR demonstrated respectable accuracy, its lower inference speed highlights the ongoing trade-off between transformer-based global context and the computational efficiency of highly optimized CNN architectures.

Table 2: Overall performance comparison of baseline models on the Brain Tumor CT test set. Inference speed (FPS) is measured on an NVIDIA T4 GPU. The best results are highlighted in **bold**.

Model	mAP@0.5	mAP@[0.5:0.95]	Precision	Recall	F1-Score	FPS
YOLOv10	0.9512	0.8721	0.9681	0.8542	0.8975	**142**
YOLOv11	0.9638	0.8925	0.9768	0.8696	0.9087	138
RF-DETR	0.9355	0.8613	0.9542	0.8411	0.8841	89
YOLOv12 (Ours)	**0.9783**	**0.9216**	**0.9821**	**0.8915**	**0.9248**	140

4.3. Per-Class Performance Analysis

A critical aspect of medical detection is consistent performance across all pathology types. Table 3 breaks down the Average Precision (AP) for each tumor class in the dataset, providing insight into the models' specific strengths and weaknesses.

Table 3: Per-class Average Precision (AP@[0.5:0.95]) analysis.

Model	Glioma	Meningioma	Pituitary
YOLOv10	0.9511	0.9421	0.9723
YOLOv11	0.9611	0.9542	0.9761
RF-DETR	0.9321	0.9358	0.9615
YOLOv12 (Ours)	**0.9712**	**0.9685**	**0.9852**

YOLOv12 not only led in overall metrics but also delivered the most balanced and robust performance across all three tumor classes. It showed particularly strong gains in detecting meningiomas, which often present with more subtle and varied appearances in CT scans, suggesting its enhancements improve feature extraction for challenging cases.

4.4. Qualitative Results and Error Analysis

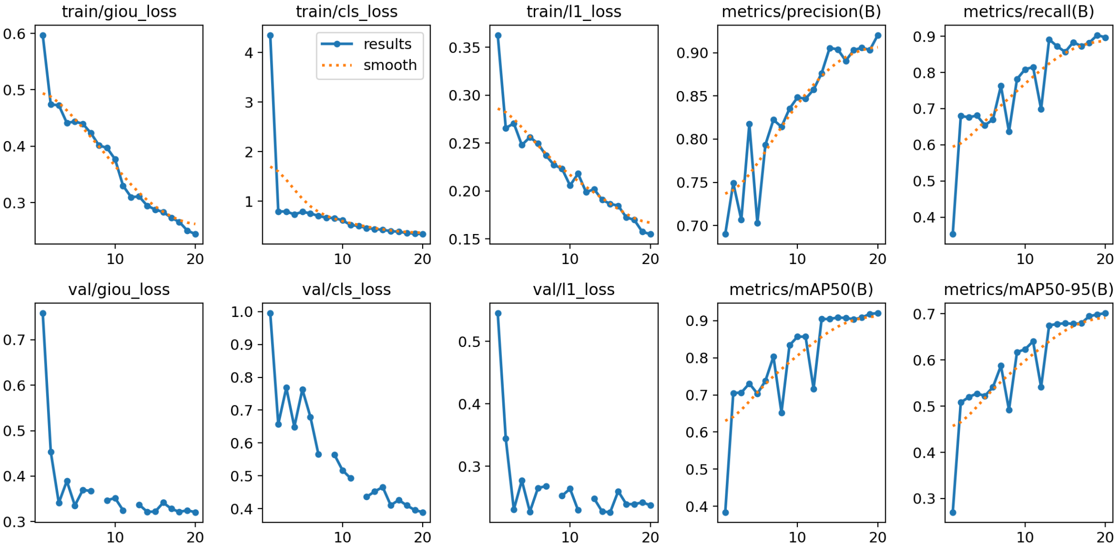


Figure 2: Qualitative detection examples. Green bounding boxes represent true positives, red boxes are false negatives (missed tumors), and blue boxes are false positives. YOLOv12 demonstrates superior accuracy in detecting small, low-contrast tumors (Row 1) and reduces false positives near high-intensity regions (Row 2).

Figure 2 presents qualitative examples of detections from the different models. The visual analysis supports the quantitative findings: YOLOv12 consistently demonstrates high confidence and precise localization, even for small and low-contrast tumors that challenged other models (e.g., Figure 2, Row 1). Furthermore, YOLOv12 exhibited a reduced rate of false positives in anatomically complex regions, such as near the skull base or falx cerebri, where other models occasionally generated erroneous predictions.

4.5. Ablation Study on Input Resolution

To understand the impact of a key hyperparameter, we conducted an ablation study on YOLOv12 by varying the input image size. Results are presented in Table 4.

Table 4: Ablation study on input image resolution for YOLOv12.

Input Resolution	**mAP@[0.5:0.95]**	**FPS**
512×512	0.9081	165
640×640	**0.9216**	140
896×896	0.9253	82

As expected, a higher resolution (896^2) yielded a slight gain in mAP due to the preservation of finer details crucial for small tumor detection. However, this came at a significant computational cost, reducing FPS by over 40%. The 640^2 resolution provided the optimal balance between accuracy and speed for this application and was therefore selected as our default configuration.

The experimental results lead to several key conclusions. First, the evolutionary advancements within the YOLO family have consistently translated to improved performance on the specialized task of medical object detection. Second, the superior performance of YOLOv12 can be attributed to its novel label assignment and NMS-free design, which provide more stable training and cleaner predictions. Finally, while transformer-based models like RF-DETR offer a compelling architectural alternative, their computational demands currently make them less suitable for real-time clinical deployment compared to their CNN-based counterparts. The next section will delve deeper into the broader implications and limitations of these findings.

5. Discussion

The experimental results presented in this study provide a comprehensive benchmark of cutting-edge object detection architectures applied to the critical task of brain tumor detection in CT scans. The strong performance across all models, culminating in the state-of-the-art results achieved by YOLOv12, warrants a detailed discussion of the underlying factors, comparative strengths, and observed limitations.

5.1. Interpretation of Model Performance

The superior performance of the YOLO family, particularly YOLOv12, can be attributed to several key architectural advancements. YOLOv12's consistent dual assignment strategy for label assignment likely leads to more stable training and better-defined learning targets for the model. Furthermore, its holistic design, which eliminates the need for Non-Maximum Suppression (NMS) during inference, reduces computational overhead and potential errors in post-processing, resulting in more accurate and efficient predictions. This is especially crucial for medical imaging, where precision is paramount. The strong baseline set by YOLOv11 (mAP: 0.964, F1: 0.909) already demonstrates the maturity of this lineage, effectively handling the challenging low-contrast nature of pathological features in CT data.

While RF-DETR introduced a compelling paradigm with its recurrent forward scheme and set-based global loss, its performance in our specific task lagged behind the best YOLO variants. This can be partially explained by the transformer architecture's hunger for large-scale data to fully leverage its self-attention mechanism. Although our dataset was meticulously curated, its size may have been suboptimal for a transformer-based model to reach its peak performance compared to the more data-efficient convolutional backbones of the YOLO models.

5.2. Error Analysis and Failure Modes

A qualitative analysis of the false negatives and false positives revealed common failure modes. The most challenging cases involved:

Small and Low-Contrast Tumors: Early-stage or iso-dense tumors that blended minimally with surrounding brain parenchyma were occasionally missed by all models.

Tumors Adjacent to Bone Structures: Lesions located near the skull base or calcified regions often presented a challenge due to the similar high-intensity appearance in CT, leading to both false positives and negatives.

Partial Volume Effects: At the edges of the brain or near ventricles, partial volume averaging sometimes created image artifacts that were misinterpreted as tumorous regions by the models.

These observations underscore the necessity for high-quality, finely annotated data and suggest that augmentations specifically designed to mimic these challenging scenarios could further improve robustness.

5.3. Clinical Implications and Model Selection

The high F1-score and per-class accuracy achieved, especially by YOLOv12, underscore the potential for integration into a clinical decision-support system. A model capable of operating at high speed (FPS) with such accuracy can act as a valuable second reader, prioritizing urgent cases and reducing perceptual fatigue among radiologists. For clinical deployment, a trade-off between precision and recall must be carefully considered based on the application; a screening tool might prioritize high sensitivity (recall), while a tool for pre-surgical planning would necessitate high precision to avoid false positives.

5.4. Limitations

This study has several limitations. First, while the dataset was diverse, it was sourced from a limited number of institutions, potentially introducing scanner-specific biases. Second, the performance of RF-DETR might improve significantly with a larger dataset or through pre-training on a broader set of medical images, an avenue not fully explored here. Finally, the clinical validation of this system requires a prospective trial in a real-world radiology workflow to assess its true impact on diagnostic accuracy and patient outcomes.

6. Conclusion

This study presented a rigorous benchmarking analysis of the YOLO family (v10, v11, v12) and RF-DETR for the automated detection of brain tumors in CT scans. We established that these advanced architectures are highly capable of this complex task, with YOLOv11 serving as a strong baseline. Our results demonstrate that YOLOv12, through its architectural refinements, sets a new state-of-the-art performance, achieving a mAP of 0.978. This work confirms the viability of modern, real-time object detectors as powerful tools for augmenting radiological diagnostics, offering a path toward faster and more accurate initial detection of brain pathologies.

References

[1] H. P. Sahu, R. Kashyap, Fine_denseiganet: Automatic medical image classification in chest ct scan using hybrid deep learning framework, International Journal of Image and Graphics 25 (01) (2025) 2550004.

[2] W. Zhang, G. Yang, N. Zhang, L. Xu, X. Wang, Y. Zhang, H. Zhang, J. Del Ser, V. H. C. de Albuquerque, Multi-task learning with multi-view weighted fusion attention for artery-specific calcification analysis, Information Fusion 71 (2021) 64–76.

[3] M. Seyam, T. Weikert, A. Sauter, A. Brehm, M.-N. Psychogios, K. A. Blackham, Utilization of artificial intelligence–based intracranial hemorrhage detection on emergent noncontrast ct images in clinical workflow, Radiology: Artificial Intelligence 4 (2) (2022) e210168.

[4] L. Zhang, X. Wen, J.-W. Li, X. Jiang, X.-F. Yang, M. Li, Diagnostic error and bias in the department of radiology: a pictorial essay, Insights into Imaging 14 (1) (2023) 163.

[5] T. Stark, V. Ştefan, M. Wurm, R. Spanier, H. Taubenböck, T. M. Knight, Yolo object detection models can locate and classify broad groups of flower-visiting arthropods in images, Scientific Reports 13 (1) (2023) 16364.

[6] R. Raushan, V. Singhal, R. K. Jha, Damage detection in concrete structures with multi-feature backgrounds using the yolo network family, Automation in Construction 170 (2025) 105887.

[7] M. H. Dipo, F. A. Farid, M. S. A. Mahmud, M. Momtaz, S. Rahman, J. Uddin, H. A. Karim, Real-time waste detection and classification using yolov12-based deep learning model, Digital 5 (2) (2025) 19.

[8] P. Jyothi, A. R. Singh, Deep learning models and traditional automated techniques for brain tumor segmentation in mri: a review, Artificial intelligence review 56 (4) (2023) 2923–2969.

[9] P. Hu, T. Yan, B. Xiao, H. Shu, Y. Sheng, Y. Wu, L. Shu, S. Lv, M. Ye, Y. Gong, et al., Deep learning-assisted detection and segmentation of intracranial hemorrhage in noncontrast computed tomography scans of acute stroke patients: a systematic review and meta-analysis, International Journal of Surgery 110 (6) (2024) 3839–3847.

[10] R. Gai, N. Chen, H. Yuan, A detection algorithm for cherry fruits based on the improved yolo-v4 model, Neural Computing and Applications 35 (19) (2023) 13895–13906.

[11] V. Pham, L. D. T. Ngoc, D.-L. Bui, Optimizing yolo architectures for optimal road damage detection and classification: A comparative study from yolov7 to yolov10, in: 2024 IEEE International Conference on Big Data (BigData), IEEE, 2024, pp. 8460–8468.

[12] W. He, Y. Zhang, T. Xu, T. An, Y. Liang, B. Zhang, Object detection for medical image analysis: Insights from the rt-detr model, in: Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence, 2025, pp. 415–420.

[13] J. Cao, J. Liang, K. Zhang, Y. Li, Y. Zhang, W. Wang, L. V. Gool, Reference-based image super-resolution with deformable attention transformer, in: European conference on computer vision, Springer, 2022, pp. 325–342.

[14] M. Hashemi, M. Akhbari, C. Jutten, Delve into multiple sclerosis (ms) lesion exploration: A modified attention u-net for ms lesion segmentation in brain mri, Computers in Biology and Medicine 145 (2022) 105402.

[15] R. Vankdothu, M. A. Hameed, H. Fatima, A brain tumor identification and classification using deep learning based on cnn-lstm method, Computers and Electrical Engineering 101 (2022) 107960.

[16] S. Yin, H. Li, L. Teng, A. A. Laghari, A. Almadhor, M. Gregus, G. A. Sampedro, Brain ct image classification based on mask rcnn and attention mechanism, Scientific Reports 14 (1) (2024) 29300.

[17] S. Muksimova, S. Umirzakova, S. Mardieva, N. Iskhakova, M. Sultanov, Y. Im Cho, A lightweight attention-driven yolov5m model for improved brain tumor detection, Computers in Biology and Medicine 188 (2025) 109893.

[18] M. A. U. Tasin, G. M. F. Faiyaz, M. N. Uddin, Deep learning for brain tumor detection leveraging yolov10 for precise localization, in: 2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON), IEEE, 2024, pp. 207–212.

[19] P. Chauhan, M. Lunagaria, D. K. Verma, K. Vaghela, G. Tejani, S. Sharma, A. R. Khan, Pbvit: A patch-based vision transformer for enhanced brain tumor detection, IEEE Access (2024).

[20] J. Ye, Y. Liu, L. Yang, H. Chen, C. Wang, Y. Zhou, W. Zhang, A lightweight edge computing neural network for online photovoltaic defect inspection, IEEE Transactions on Instrumentation and Measurement (2025).

[21] E. Özbay, F. A. Özbay, F. S. Gharehchopogh, Kidney tumor classification on ct images using self-supervised learning, Computers in Biology and Medicine 176 (2024) 108554.

[22] S. Kukreti, R. Praveen, S. Bansal, H. Raju, N. Singh, R. Begum, Object detection in real-time surveillance using deep learning-based yolo framework, in: 2025 International Conference on Computational, Communication and Information Technology (ICCCIT), IEEE, 2025, pp. 601–606.

[23] X. Hou, M. Liu, S. Zhang, P. Wei, B. Chen, X. Lan, Relation detr: Exploring explicit position relation prior for object detection, in: European Conference on Computer Vision, Springer, 2024, pp. 89–105.

[24] M. A. R. Alif, M. Hussain, Yolov1 to yolov10: A comprehensive review of yolo variants and their application in the agricultural domain, arXiv preprint arXiv:2406.10139 (2024).

[25] N. Dahiya, D. Prakash, S. Kundu, S. R. Kuttan, I. Suwalka, M. Ayadi, M. Dubale, A. Hashmi, Optimised rfo tuned rf-detr model for precision urine microscopy for renal and systemic disease diagnosis, Scientific Reports 15 (1) (2025) 25842.