UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES ET INFORMATIQUE APPLIQUÉES (3799)

PAR
Neda Saberitabar

DÉTECTION DES TUMEURS CÉRÉBRALES SUR IMAGES IRM PAR APPRENTISSAGE PROFOND : COMPARAISON D'UN MODÈLE DETR-SE-RESNET50 AMÉLIORÉ AVEC YOLOv8 ET LES VISION TRANSFORMERS

SEPTEMBRE 2025

Université du Québec à Trois-Rivières

Service de la bibliothèque

Avertissement

Summary

This dissertation explores the application of state-of-the-art deep learning techniques—specifically, Detection Transformer (DETR) with SE-ResNet50 backbone and YOLOv8—for the detection and classification of brain tumors in medical imaging. It adopts a multidisciplinary approach that combines imaging technologies with artificial intelligence to enhance diagnostic accuracy and personalized treatment strategies [1-3].

Introduction: Brain tumors are severe health conditions, and new methodologies are seriously needed to enable their early and correct diagnosis. This review will point out the shortcomings of the traditional imaging landscape while underlining the potential of deep learning algorithms to transform how these complex conditions are assessed and managed [1].

Because of the importance and contribution of image analysis to the field of medical imaging, a review of relevant studies has been conducted concerning the use of the DETR and YOLO frameworks. Such a discussion, therefore, covers the theoretical bases of these deep learning models and how complex imaging data is processed and analyzed for effective detection and classification of tumors [3].

The methods section elaborates on the research design, indicating that the study leverages a rich dataset of medical imaging studies related to brain tumors. Two models are pursued in this study: DETR, YOLOv8 and Vit. This study evaluates these state-of-the-art models for their performance in accurately detecting and classifying tumors and malignant cells [4]. Data augmentation, cross-validation, and metrics concerning model performance evaluation are discussed in detail to ensure robustness in the analysis.

Both DETR and YOLOv8 yield significant improvements in detection and classification accuracies, outperforming traditional approaches. The implications of these results for clinical practice include earlier and more accurate diagnosis, enabling more personalized treatment planning. However, given the limitations of the available datasets, further validation is required [5].

Brain tumors are critical health conditions that demand timely and precise diagnosis. However, traditional imaging methods often fall short due to their reliance on manual interpretation, variability in tumor morphology, and inconsistent image quality. In response to these limitations, this study highlights the potential of deep learning algorithms to improve the diagnosis and management of such complex conditions [1, 4], though clinical trials are needed to validate real-world impact.

Given the pivotal role of image analysis in medical diagnostics, a comprehensive review of recent work involving DETR and YOLO frameworks has been conducted. This discussion outlines the theoretical foundations of these models and examines how they process and analyze complex imaging data to effectively detect and classify brain tumors [5, 6].

The research methodology is designed around the use of diverse, annotated medical imaging datasets, including MRI scans of brain tumors. This study focuses on evaluating two key models—DETR with SE-ResNet50 backbone and YOLOv8—and assesses their performance using rigorous evaluation metrics such as accuracy, precision, recall, F1-score, mean Average Precision (mAP), and Intersection over Union (IoU) [3]. Data preprocessing, augmentation strategies, and five-fold cross-validation are employed to ensure robustness and generalizability [7].

Initial results indicate that both DETR and YOLOv8 improve detection and classification performance over traditional methods, with DETR offering higher precision and F1-score, while YOLOv8 provides better recall (83.5%) and excels in speed, which is crucial for clinical applications where minimizing missed detections (false negatives) is prioritized over avoiding false positives. These outcomes underscore the clinical potential of deep learning models to support early and accurate diagnosis, thereby enhancing treatment planning and patient outcomes. Nonetheless, we acknowledge limitations due to dataset size and calls for further validation across broader and more diverse data sources [8, 9].

In conclusion, this research not only demonstrates the effectiveness of advanced deep learning models in medical imaging but also sets the stage for future work focused on improving model interpretability, robustness, and clinical integration [10].

Keywords: Deep Learning, MRI Imaging, Tumor Detection, and Computer Vision

Résumé

Cette mémoire explore l'application de techniques d'apprentissage profond de pointe – notamment le « Detection Transformer » (DETR) avec un backbone SE-ResNet50 et YOLOv8 – pour la détection et la classification des tumeurs cérébrales dans l'imagerie médicale. Elle adopte une approche multidisciplinaire combinant les technologies d'imagerie avec l'intelligence artificielle pour améliorer la précision diagnostique et les stratégies de traitement personnalisé [1-3].

Introduction : Les tumeurs cérébrales sont des conditions de santé graves, et de nouvelles méthodologies sont nécessaires pour permettre un diagnostic précoce et précis. Cette revue souligne les lacunes du paysage d'imagerie traditionnel tout en soulignant le potentiel des algorithmes d'apprentissage profond à transformer l'évaluation et la gestion de ces conditions complexes [1].

En raison de l'importance et de la contribution de l'analyse d'image au domaine de l'imagerie médicale, une revue des études pertinentes a été menée concernant l'utilisation des cadres DETR et YOLO. Cette discussion couvre donc les bases théoriques de ces modèles d'apprentissage profond et la manière dont les données d'imagerie complexes sont traitées et analysées pour une détection et une classification efficace des tumeurs [3].

La section méthodes élabore sur la conception de la recherche, indiquant que l'étude exploite un riche ensemble de données d'études d'imagerie médicale liées aux tumeurs cérébrales. Deux modèles sont poursuivis dans cette étude : DETR et YOLOv8. Cette étude évalue ces modèles de pointe pour leur performance dans la détection et la classification précise des tumeurs et des cellules malignes [4]. L'augmentation des données, la validation croisée et les métriques concernant l'évaluation des performances du modèle sont discutées en détail pour assurer la robustesse de l'analyse.

DETR et YOLOv8 produisent tous deux des améliorations significatives dans les précisions de détection et de classification, surpassant les approches traditionnelles. Les implications de ces résultats pour la pratique clinique incluent un diagnostic plus précoce et plus précis, permettant une planification de traitement plus personnalisée. Cependant, compte tenu des limitations des ensembles de données disponibles, une validation supplémentaire est requise [5].

Les tumeurs cérébrales sont des conditions de santé critiques qui exigent un diagnostic opportun et précis. Cependant, les méthodes d'imagerie traditionnelles sont souvent insuffisantes en raison de leur dépendance à l'interprétation manuelle, de la

variabilité morphologique des tumeurs et de la qualité inconsistante des images. En réponse à ces limitations, cette étude met en lumière le potentiel des algorithmes d'apprentissage profond à améliorer le diagnostic et la gestion de ces conditions complexes [1, 4], bien que des essais cliniques soient nécessaires pour valider l'impact dans le monde réel.

Compte tenu du rôle pivot de l'analyse d'image dans les diagnostics médicaux, une revue complète des travaux récents impliquant les cadres DETR et YOLO a été menée. Cette discussion décrit les fondations théoriques de ces modèles et examine comment ils traitent et analysent les données d'imagerie complexes pour détecter et classer efficacement les tumeurs cérébrales [5, 6].

La méthodologie de recherche est conçue autour de l'utilisation d'ensembles de données d'imagerie médicale diversifiés et annotés, y compris des scans IRM de tumeurs cérébrales. Cette étude se concentre sur l'évaluation de deux modèles clés – DETR avec backbone SE-ResNet50 et YOLOv8 – et évalue leur performance en utilisant des métriques d'évaluation rigoureuses telles que la précision, la précision, le rappel, le score F1, la moyenne de précision moyenne (mAP) et l'intersection sur l'union (IoU) [3]. Le prétraitement des données, les stratégies d'augmentation et la validation croisée à cinq plis sont employés pour assurer la robustesse et la généralisabilité [7].

Les résultats initiaux indiquent que DETR et YOLOv8 améliorent les performances de détection et de classification par rapport aux méthodes traditionnelles, DETR offrant une précision et un score F1 plus élevés, tandis que YOLOv8 fournit un meilleur rappel (83.5%) et excelle en vitesse, ce qui est crucial pour les applications cliniques où la minimisation des détections manquées (faux négatifs) est priorisée par rapport à l'évitement des faux positifs. Ces résultats soulignent le potentiel clinique des modèles d'apprentissage profond pour soutenir un diagnostic précoce et précis, améliorant ainsi la planification du traitement et les résultats pour les patients. Néanmoins, nous reconnaissons les limitations dues à la taille de l'ensemble de données et appelons à une validation supplémentaire à travers des sources de données plus larges et plus diverses [8, 9].

En conclusion, cette recherche non seulement démontre l'efficacité des modèles d'apprentissage profond avancés en imagerie médicale, mais pose également les bases pour des travaux futurs axés sur l'amélioration de l'interprétabilité des modèles, de la robustesse et de l'intégration clinique [10].

Keywords: Deep Learning, MRI Imaging, Tumor Detection, and Computer Vision

Acknowledgements

First and foremost, I would like to express my gratitude to God. Without His strength, patience, and grace, I couldn't have made it through this journey. He has been my guiding light through the hardest moments, and I continue to rely on His presence for whatever lies ahead.

To my incredible parents, thank you for everything. Your love, your constant encouragement, and the sacrifices you made for me have shaped who I am today. You have always been my rock, and this achievement belongs just as much to you as it does to me.

I am deeply grateful to my supervisor, Dr. Faghihi. His support, sharp insights, and encouragement helped me grow, not just in my research, but in how I think and approach problems. It has been an honor to learn under his guidance.

A big thanks to my fellow students, colleagues, and friends who stood by me along the way. Your help, ideas, late-night chats, and even just your kind words made a world of difference.

Lastly, to everyone who had a part in this thesis, whether big or small, thank you. This was never a solo journey, and I am grateful for all the help, seen and unseen, that brought this work to life.

Dedication

To my beloved family,

This thesis is for you, my constant support, shelter, and everything.

To my husband: thank you for never giving up on me, for believing in me even when I did not believe in myself. Your love and strength kept me going when things felt impossible.

To my dear parents and siblings: your love, lessons, and endless prayers are with me on every page of this work. You have taught me how to be strong, how to care deeply, and how to never stop learning.

This is for you, with all my love, always.

Contents

Acronyms

1. **ACC** - Accuracy: The proportion of correct predictions over all samples; a general indicator of model correctness.
2. **AI** - Artificial Intelligence: Refers to the simulation of human intelligence in machines.
3. **AUC** - Area Under the Curve: The area under the ROC or PR curve summarizes the discriminative ability of a model.
4. **CI -** Confidence Interval: A statistical interval (e.g., 95%) expressing the uncertainty or stability of metric estimates.
5. **CNN** - Convolutional Neural Network: A deep learning algorithm is particularly effective for image processing.
6. **CT** - Computed Tomography: A medical imaging technique that uses X-rays to create detailed images of the body.
7. **DETR** - Detection Transformer: A model architecture designed for object detection using transformer technology.
8. **F1 -** F1-Score: The harmonic mean of Precision and Recall; balances false positives and false negatives.
9. **IoU** - Intersection over Union: A metric used to evaluate the accuracy of an object detector.
10. **ML** - Machine Learning: A subset of AI that focuses on developing algorithms that allow computers to learn from and make predictions based on data.
11. **MRI** - Magnetic Resonance Imaging: A non-invasive imaging technology that produces detailed images of the brain.
12. **mAP** - Mean Average Precision: A performance measure often used in object detection tasks.
13. **PET** - Positron Emission Tomography: An imaging test that helps reveal how tissues and organs are functioning.
14. **PR** - Precision-Recall: A performance metric used to evaluate the effectiveness of classification models.
15. **ROC** - Receiver Operating Characteristic: A curve showing the trade-off between sensitivity (recall) and false positive rate.
16. **SAM** - Spatial Attention Mechanism: A technique to improve model performance is focusing on significant image regions.

17. **SE-ResNet** - Squeeze-and-Excitation ResNet: A ResNet variant enhanced with channel attention (SE module) to improve feature representation.
18. **SPPF -** Spatial Pyramid Pooling – Fast: A lightweight variant of spatial pyramid pooling used in YOLOv5/YOLOv8, designed to capture multi-scale features efficiently while reducing computation time.
19. **T1** - T1-weighted Imaging: A type of MRI sequence that provides contrast between fatty tissue and cerebrospinal fluid.
20. **T2** - T2-weighted Imaging: An MRI technique that emphasizes differences in tissue composition.
21. **TP / FP / FN / TN** - True Positive / False Positive / False Negative / True Negative: Four prediction outcomes relative to ground-truth labels; basis for most performance metrics.
22. **ViT** - Vision Transformer: A transformer-based architecture for computer vision that processes images as sequences of patches.
23. **YOLO** - You Only Look Once: A family of fast object detection models that process the whole image in a single pass.
24. **WHO** - World Health Organization: A specialized agency of the United Nations responsible for international public health.

List of tables

List of figures

Chapter 1: Introduction

Brain tumors represent a serious health concern across all age groups. They are typically classified as benign (non-cancerous) or malignant (cancerous); malignant tumors are aggressive, can infiltrate adjacent brain tissue, and often recur, making early and accurate detection essential for effective treatment planning and improved prognosis [11]. The 2021 WHO classification further organizes primary brain tumors by histology and molecular markers into prognostic grades (I–IV), reflecting expected growth and recurrence behavior [11]. Within this spectrum, gliomas are the most frequent malignant primaries, meningiomas are common benign tumors arising from the meninges, and pituitary adenomas—though often slow-growing—can disrupt endocrine function and vision. At the population level, the overall cancer burden varies by region and age; "brain and CNS" cancers form a distinct share of incidence and mortality across world regions, underscoring why reliable imaging-based detection matters in diverse clinical settings. Together, these factors motivate automated methods that complement radiologists by improving sensitivity to subtle or diffuse lesions while preserving specificity.

Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) scans have revolutionized the diagnosis of brain tumors by providing detailed anatomical insights. However, interpretation of these scans remains a challenge due to the subjective nature of manual analysis, variability in tumor morphology, and inconsistencies in image quality—all of which may contribute to diagnostic errors or delays [2]. To address these limitations, artificial intelligence (AI)—particularly deep learning—has emerged as a promising tool to enhance image interpretation and reduce the workload on radiologists [1, 12].

Convolutional Neural Networks (CNNs) have shown strong performance in medical image analysis by extracting complex patterns and features from imaging data [13]. However, their limited ability to model long-range spatial dependencies makes them less effective for detecting tumors with irregular shapes or overlapping anatomical structures [14]. To overcome these challenges, more advanced architectures such as Vision Transformers (ViT) and Detection Transformers (DETR) have been developed to better capture global contextual information in imaging data [6]. Furthermore, models like YOLOv8 offer real-time object detection capabilities, making them well-suited for integration into clinical workflows [15, 16].

Despite these advancements, several challenges persist. Many deep learning models are computationally intensive and may underperform when detecting small or poorly defined

tumors [5, 17]. In this research, we propose an enhanced DETR architecture that incorporates the Squeeze-and-Excitation ResNet50 (SE-ResNet50) as its backbone. This integration aims to improve the model's feature extraction capabilities, increase classification accuracy, and reduce computational overhead.

1.2 Problem Statement

While deep learning has significantly advanced brain tumor detection, several critical challenges remain. Accurately identifying small or irregularly shaped tumors is particularly difficult due to their complex morphology and ambiguous boundaries [18]. Additionally, the high computational costs and slow convergence rates of many deep learning models hinder their practical application in clinical environments [19]. Variability in imaging quality across different datasets further complicates the development of reliable automated diagnostic systems [20].

Table 1. Challenges in AI-Based Brain Tumor Detection

Challenge	Impact
High intra-class and inter-class variation	Reduces model accuracy in distinguishing between tumor types.
Scarcity of annotated datasets	Limits deep learning model training and generalization.
Need for real-time, automated AI support.	Increases radiologist efficiency and diagnostic accuracy.
Generalization issues across datasets	Affects model transferability across different medical centers.

To address these issues, this research investigates the application of advanced deep learning models—specifically DETR, YOLOv8, and Vision Transformers—in brain tumor classification. A novel approach is proposed that integrates SE-ResNet50 to enhance feature extraction capabilities, aiming to improve classification accuracy while mitigating computational complexity. We compare a ResNet backbone to SE-ResNet50 within DETR to quantify the benefits of channel attention (see section 4.4.1).

1.3 Research Objectives

We pursue four objectives: (i) quantify the performance of ViT, YOLOv8, and DETR on a curated multi-source MRI dataset using consistent splits and uncertainty estimates; (ii) evaluate SE-ResNet50 inside DETR versus ResNet to measure channel-attention benefits; (iii) examine clinical readiness by analyzing confusion matrices, ROC/PR behavior, and error modes; and (iv) report metrics suitable for both classification and detection—Accuracy/F1/AUC (image-level) for all models, and IoU/mAP for detectors using COCO-style evaluation [21].

1.4 Significance of the Study

In our experiments, DETR with an SE-ResNet50 backbone improves class-balanced performance on brain MRI (see section 4.4). Consistent with the original DETR formulation, the detector is end-to-end and removes the need for anchors and NMS [7]. For practice, real-time detection (YOLOv8) and global reasoning (DETR) are compared head-to-head, providing guidance on speed-vs-accuracy trade-offs in clinical pipelines. Methodologically, we clarify the roles of ViT (classification) and detector AP metrics, preventing apples-to-oranges comparisons that can obscure actual performance. We also highlight hierarchical/sparse transformer variants (e.g., Swin) as scalable options for high-resolution imaging [21, 22].

1.5 Structure of the Thesis

The thesis is organized to guide the reader from foundational context to empirical evidence and concluding insights. Chapter 2 reviews tumor biology, imaging modalities, and model families (CNNs, ViT, YOLOv8, DETR). Chapter 3 details data curation, preprocessing, and model training. Chapter 4 presents results, uncertainty analyses, and ablation results; the conclusion synthesizes the clinical implications and limitations.

Chapter 2: State-of-the-Art

2.1 Introduction

Medical imaging has played a pivotal role in the early detection and treatment of brain tumors. The integration of deep learning techniques has significantly improved tumor detection accuracy, enabling automated feature extraction and classification [1, 12]. While traditional convolutional neural networks (CNNs) remain widely used, newer architectures such as Vision Transformers (ViT) and Detection Transformers (DETR) have demonstrated improved performance in classification tasks. Additionally, CNN-based models such as YOLOv8 provide efficient, real-time tumor detection, making them attractive for clinical applications.

This chapter reviews deep learning techniques applied in medical imaging, comparing existing methodologies, highlighting their strengths and limitations, and establishing the foundation for the proposed research [6, 14].

2.2 Brain Tumors (Grades and Classification)

Brain tumors are classified based on the World Health Organization (WHO) grading system, which categorizes tumors according to their aggressiveness and recurrence potential [11].

Table 2. WHO Tumor Grading System

Grade	Characteristics
Grade I	Slow-growing, well-differentiated, non-invasive
Grade II	Increased mitotic activity, potential for recurrence
Grade III	Rapid growth, higher recurrence rates
Grade IV (e.g., Glioblastoma Multiforme)	Highly aggressive, invasive, and poor prognosis

Primary brain tumors originate within the brain, while metastatic tumors spread from cancers elsewhere in the body. Gliomas are the most frequent malignant primary tumors, arising from glial cells and often presenting as glioblastomas with a poor prognosis [23].

Meningiomas (see Figure 1), which develop from the meninges, are typically benign but can exhibit aggressive behavior and recur [24].

Pituitary adenomas (see Figure 1), though often slow-growing, can disrupt hormonal balance and compress nearby structures, leading to vision and endocrine issues [25].

In contrast, brain metastases—most commonly from lung, breast, and melanoma—are significantly more prevalent and are a leading neurological complication in cancer patients [26]. Differentiating between these tumor types is crucial for effective treatment planning and prognosis.

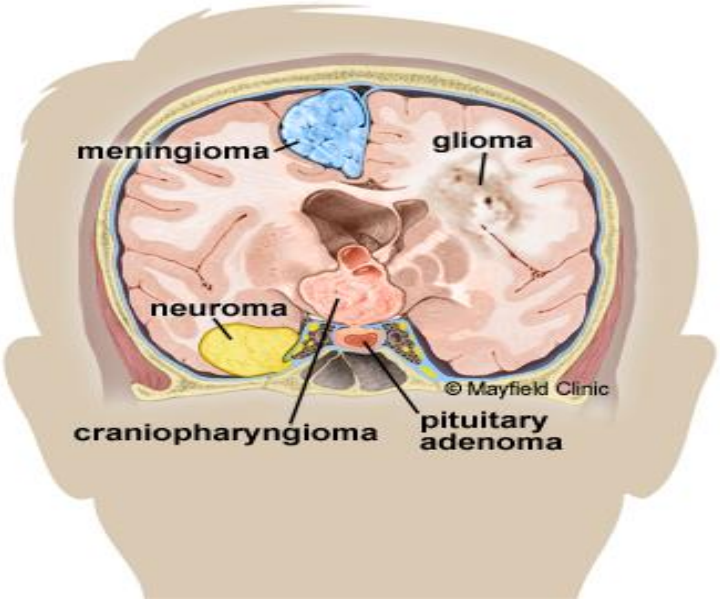


Figure 1. Brain schematic[1]

2.3 Brain Tumor Imaging Techniques

2.3.1 Magnetic Resonance Imaging (MRI)

MRI is considered the gold standard for brain tumor detection because of its superior soft tissue contrast capabilities. Various MRI sequences offer complementary information that enhances the detection and characterization of brain tumors [27].

T1-weighted images deliver detailed anatomical structure, facilitating the identification of normal and abnormal tissue boundaries. T2-weighted images are particularly useful for highlighting edema and cystic regions associated with tumors, providing valuable insights into tumor-related changes in brain tissue. Fluid-Attenuated Inversion Recovery (FLAIR) imaging enhances contrast in areas with infiltrative tumors by suppressing cerebrospinal fluid signals [28], thus making subtle pathological changes more conspicuous. Diffusion-Weighted Imaging (DWI) assesses the movement of water molecules within tissues, enabling differentiation of tumor types based on variations in cellular density [28].

[1] https://mayfieldclinic.com/pe-brain tumor.htm

The combination of these MRI sequences allows a comprehensive assessment of tumor morphology, extent, and pathology, making MRI an indispensable tool in modern neuro-oncology practice [29].

Table 3. Brief history of MRI [30]

1857-1952	Larmor relationship- Sir Joseph Larmor (3)
1930	Isidor Isaac Rabi succeeded in detecting single state of rotation of atoms and molecules, and in determining the mechanical and magnetic moments of the nuclei.(4)
1946	MR phenomenon - Bloch and Purcell(5)
1952	Nobel Prize - Bloch and Purcell(6)
1950,1960 1970	NMR developed as analytical tool(3)
1972	Computerized Tomography(3)
1973	Back projection MRI – Lauterbur(3)
1975	Fourier Imaging - Ernst(3)
1977	Echo-planar imaging – Mansfield(7)
1980	FT MRI demonstrated – Edelstein(3)
1986	Gradient Echo Imaging NMR Microscope(8)
1987	MR Angiography - Dumoulin(3)
1991	Nobel Prize – Ernst(9)
1992[1]	Functional MRI(3)
1994	Hyperpolarized 129Xe Imaging(3)
2003	Nobel Prize - Lauterbur and Mansfield(3)

2.3.2 Computed Tomography (CT) and Positron Emission Tomography (PET)

CT provides rapid, high-resolution anatomical imaging that is especially sensitive to acute intracranial hemorrhage and coarse calcifications within tumors, though its soft-tissue contrast is inferior to MRI [30]. PET, by contrast, is a molecular technique that maps tissue metabolism using radiotracers (e.g., 18F-FDG or amino-acid tracers), offering functional insight into tumor viability, aggressiveness, and treatment response. However, due to radiation exposure, PET is secondary to MRI in this thesis.

In combined PET/CT, co-registered functional and structural information improves localization of hypermetabolic foci, assists target delineation for surgery or radiotherapy, and helps distinguish recurrence from treatment-related change. Limitations include CT's lower soft-tissue contrast, PET's relatively coarse spatial resolution and tracer-specificity issues, and the use of ionizing radiation in both; therefore, in this thesis, MRI remains the primary modality [31].

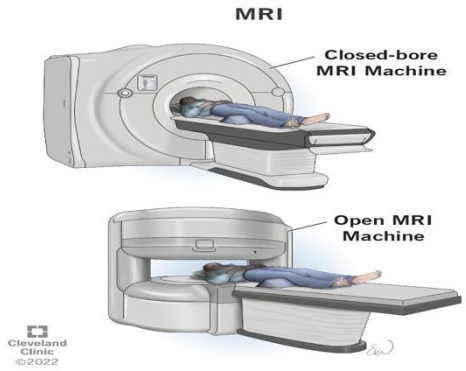


Figure 2. MRI machine[2]

2.3.3 Comparative Analysis of Imaging Techniques

Table 4. Medical Imaging Techniques: Comparative Analysis

Imaging Technique	Advantages	Disadvantages
MRI	High soft-tissue contrast, no radiation exposure	Expensive, motion-sensitive
CT scan	Fast and effective for detecting hemorrhages	Lower contrast, radiation exposure
PET Scan	Functional imaging detects metabolic activity	Exposure to radioactive materials

2.4 Deep Learning Architectures in Tumor Detection

2.4.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks provide the canonical pipeline for learning hierarchical visual representations from brain MRI. In this thesis, they act as local feature extractors that build from low-level edges and textures toward high-level tumor morphology (see figure 3) [13] . A typical block stacks convolution, nonlinearity, optional normalization, and spatial down-sampling; the network head is task-dependent: a global average pooling plus a fully connected SoftMax layer for classification, or light convolutional heads operating on feature maps for detection [32].

[2] https://my.clevelandclinic.org/health/diagnostics/4876-magnetic-resonance-imaging-mri

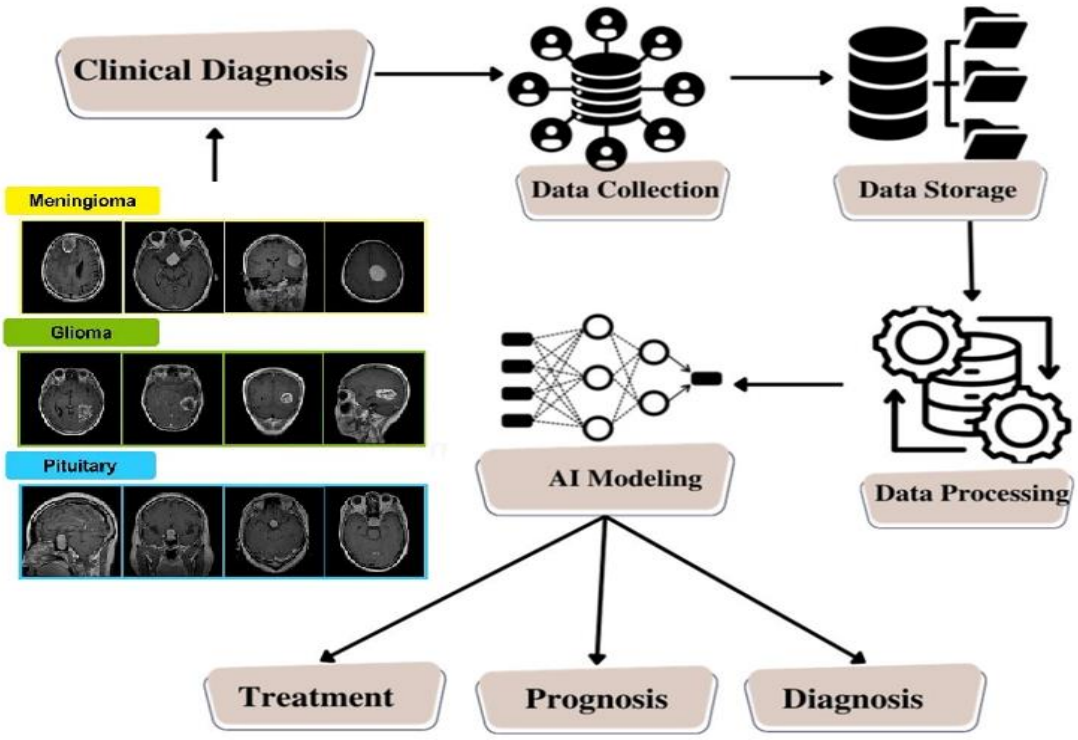


Figure 3. Schematic of AI Models Used in Tumor Detection

Let $X \in \mathbb{R}^{H\times W\times C}$ be an MRI slice. A 2-D convolution with kernels $K \in \mathbb{R}^{k_h\times k_w\times C\times C'}$, stride s, padding p, and dilation d produces $Y \in \mathbb{R}^{H'\times W'\times C'}$:

$$Y(i,j,c') = \sum_{u=0}^{k_h-1}\sum_{v=0}^{k_w-1}\sum_{c=0}^{C-1} X(is+ud-p, js+vd-p, c)K(u,v,c,c') + b_{c'} \quad \text{(Eq 2.1)}$$

Activation and normalization follow the standard forms:

$$\hat{h} = \frac{h-\mu}{\sqrt{\sigma^2+\varepsilon}} \quad BN(h) = \gamma \odot \hat{h} + \beta, \qquad ReLU(h) = \max(0,h) \quad \text{(Eq 2.2)}$$

For K-way classification with logits z ,the SoftMax and cross-entropy are:

$$p_k = \frac{ez_k}{\sum_{\ell=1}^{K} e^{z_\ell}}, \qquad L_{CE} = -logp_y \quad \text{(Eq 2.3)}$$

Deep residual stages stabilize optimization via identity or 1×1 projection shortcuts,

$$y = F(x;\Theta) + w_s x \quad \text{(Eq 2.4)}$$

ResNet residual connection (identity or projection), which preserves gradient flow and enables very deep models.

CNNs are strong when data is limited because their weight sharing and locality bias provide a good inductive structure [33]. However, their receptive fields grow only gradually, so long-range dependencies must be engineered (e.g., dilation, feature pyramids), and

aggressive down-sampling can erase subtle boundary cues important for small or infiltrative lesions [34]. These limitations motivate transformer-based architectures (next: ViT) that perform global reasoning from the first layer, and detection transformers (DETR) that cast detection as end-to-end set prediction.

2.4.2 Vision Transformer (ViT)

The Vision Transformer (ViT) is a deep learning architecture that adapts the self-attention mechanisms of transformer models, originally developed for natural language processing, to the domain of image analysis. Unlike Convolutional Neural Networks (CNNs), which process spatially local information through convolutional filters, ViT treats an image as a sequence of patches and applies global attention across the entire image [35]. This design enables the model to learn long-range dependencies and contextual relationships that are critical in complex visual tasks such as tumor detection [35, 36].

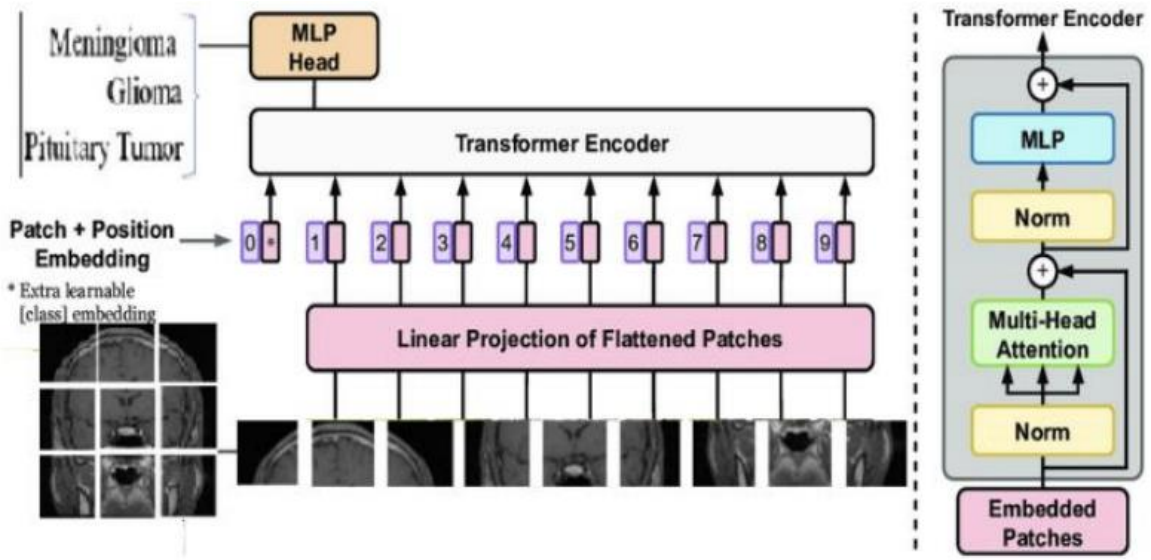


Figure 4. Overview of Vision Transformer (ViT) Architecture

Figure 4 shows that ViT adapts the transformer encoder to images by operating on patch tokens rather than pixels.

Given $X \in \mathbb{R}^{H\times W\times C}$ patch size P, we form N=(H/P)·(W/P) non-overlapping patches. Each P^2 Patch is flattened into a vector of dimension P^2C, forming the patch matrix $X_p \in \mathbb{R}^{N\times(P^2C)}$ With a learnable projection $E \in \mathbb{R}^{(P^2C)\times D}$, a class token $x_{cls} \in \mathbb{R}^D$, and positional embeddings $E_{pos} \in \mathbb{R}^{(N+1)\times D}$, the input sequence is:

$$Z_0 = [x_{cls}; X_p E] + E_{pos} \in \mathbb{R}^{(N+1)\times D} \qquad \text{(Eq 2.5)}$$

Each encoder layer uses pre-norm residual blocks:

$$Z'_\ell = Z_{\ell-1} + MSA\big(LN(Z_{\ell-1})\big), \qquad \text{(Eq 2.6)}$$

$$Z_\ell = Z'_\ell + MLP\big(LN(Z'_\ell)\big), \qquad \ell = 1, \dots, L \qquad \text{(Eq 2.7)}$$

where multi-head self-attention is:

$$MSA(Z) = Concat(h_1, \dots, h_H)W_O, \quad h_h = softmax\left(\frac{ZW_Q^{(h)}(ZW_K^{(h)})^\top}{\sqrt{d_k}}\right)ZW_V^{(h)}. \qquad \text{(Eq 2.8)}$$

Classification uses the final class token: $\hat{y} = softmax(W_{cls}z_L^{(cls)} + b_{cls})$.

After L layers, the class token is fed to a linear head; class probabilities are obtained via SoftMax (Eq. 2.3). ViT captures long-range context directly and performs strongly with large-scale pretraining and careful regularization. Its attention cost scales quadratically with the number of tokens, and it is more data-hungry than CNNs; these are the primary trade-offs. Given the clinical need for both efficiency and accuracy, we next study a real-time detector.

A Vision Transformer (ViT) tokenizes an image into patches and applies global self-attention, enabling modeling of long-range dependencies; when pretrained at scale, ViT can match or surpass CNNs and transfer effectively—including to medical imaging—as shown in the original ViT work and recent surveys [2, 14]. At the same time, ViT lacks CNNs' locality and translation-equivariance inductive biases and is more data-hungry without large-scale pretraining, as analyzed in comparative studies of ViT vs. CNNs [37]. Self-attention has $O(N^2)$ cost in the number of tokens N, making high-resolution MRI expensive unless one increases patch size or uses hierarchical/sparse variants (e.g., Swin or deformable attention) [14]. Given our need for reliable training under limited data and real-time inference, we next examine a real-time detector (YOLOv8) before returning to transformer-based detection with DETR.

2.4.3 YOLOv8: Real-Time Object Detection in Medical Imaging

Although YOLOv10 (2024) is the newest model in the YOLO family, YOLOv8 remains a mature, well-supported baseline that is widely used in production and research, including medical imaging [15]. YOLOv8 (You Only Look Once version 8), which is one of the advancements in the YOLO family of object detection models, was developed by Ultralytics in 2023. It introduces several architectural and performance improvements over its predecessors (YOLOv5, YOLOv7), making it particularly well-suited for real-time and high-accuracy medical image analysis, including tumor detection in brain MRI. Unlike traditional object detectors that rely on region proposal networks and two-stage processing, YOLOv8 follows a single-stage anchor-free architecture that directly predicts bounding

boxes and class probabilities, resulting in both computational efficiency and strong detection accuracy [38, 39].

Figure 5 illustrates the YOLOv8 architecture used for object detection in this thesis. On the left is the backbone, a stack of convolutional and C2f blocks that progressively down-sample the input image and extract features at multiple resolutions; a spatial pyramid pooling–fast (SPPF) layer expands the effective receptive field with minimal cost. The middle portion shows the neck, which fuses features across scales using a top-down and bottom-up pathway with up-sampling and concatenation, producing rich, multi-level representations suitable for detecting small, medium, and large lesions. On the right is the decoupled detection head applied at three output scales (commonly denoted P3–P5). For each scale, separate lightweight branches predict bounding-box offsets, objectness scores, and class probabilities, an arrangement that improves convergence and accuracy compared with coupled heads. The diagram also indicates that YOLOv8 adopts an anchor-free design and is available in multiple model sizes by adjusting depth and width multipliers, enabling a trade-off between speed and accuracy for different compute budgets. Overall, the figure emphasizes YOLOv8's end-to-end flow from hierarchical feature extraction, through multi-scale feature fusion, to fast, scale-aware detection suitable for real-time medical imaging workflows.

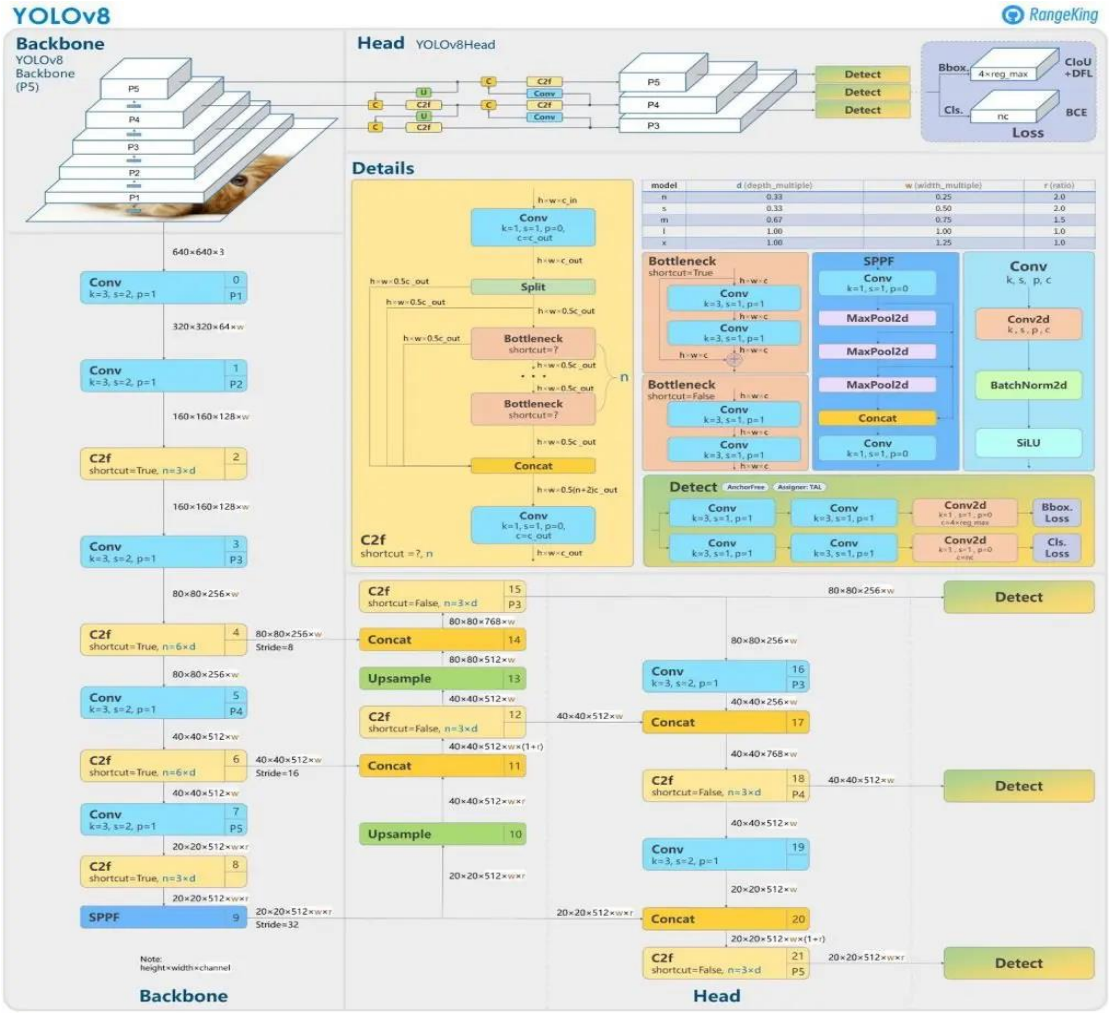


Figure 5. Overview of YOLOv8 Architecture Source: RangeKing (GitHub)

$$L_{YOLO} = \lambda_{box}L_{box} + \lambda_{obj}L_{obj} + \lambda_{cls}L_{cls} \quad \text{(Eq 2.9)}$$

The box term uses Complete-IoU:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2\left((\hat{c}_x, \hat{c}_y), (c_x, c_y)\right)}{c^2} + \alpha v, \qquad v = \frac{4}{\pi^2}(arctan\frac{w}{h} - arctanh\frac{\widehat{w}}{\hat{h}})^2,$$

$$\alpha = \frac{v}{(1-IoU)+v} \quad \text{(Eq. 2.10)}$$

Objectness and class terms use binary cross-entropy per location/class:

$$BCE(p,t) = -[t\ logp + (1-t)log(1-p)] \quad \text{(Eq 2.11)}$$

YOLOv8 delivers low-latency inference and robust multi-scale detection, which suits clinical workflows [15, 16]. Its main limitations stem from a convolutional backbone with limited global context and sensitivity to augmentation/hyperparameters on small heterogeneous datasets [15, 40]. These considerations motivate a detector that integrates global attention and removes heuristic post-processing.

Despite its strengths, YOLOv8 has certain limitations. Firstly, while efficient, its convolutional backbone lacks global context modeling. This restricts its ability to capture long-range dependencies compared to transformer-based models like ViT or DETR, which are better suited for capturing distributed tumor patterns [2, 41]. YOLOv8 also relies heavily on data augmentation and hyperparameter tuning for optimal performance. In medical imaging, where datasets are small and variations in acquisition are high, these dependencies can result in inconsistent generalization [9, 15, 42].

Another concern is its limited interpretability. Like most CNN-based detectors, YOLOv8 does not offer transparent reasoning behind its predictions. Although some visualization tools exist (e.g., Grad-CAM), they do not fully satisfy the clinical demand for model explainability [10, 43].

Finally, YOLOv8, though capable, may struggle with very small tumors or lesions with ambiguous boundaries, particularly in cases with overlapping intensities or artifacts in MRI scans [18, 44]. These challenges necessitate more globally aware models or hybrid architectures.

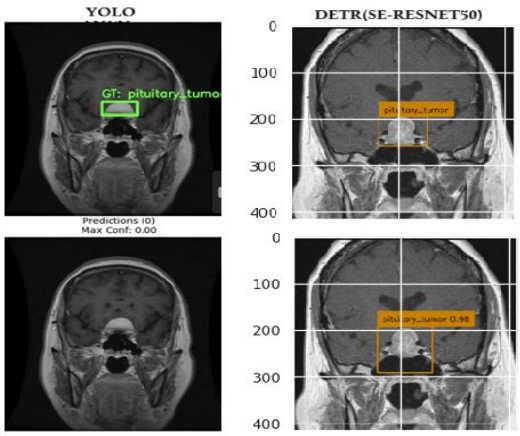


Figure 6. A Sample of DETR and YOLOv8

Given these limitations, the next section introduces Detection Transformers (DETR), which incorporate self-attention mechanisms to address the contextual and interpretability shortcomings of YOLO-based models.

2.4.4. Detection Transformer (DETR)

DETR is a transformer-based detector [6] that removes hand-crafted components such as anchor boxes, region proposals, and non-maximum suppression. Its architecture (Figure 7) consists of a convolutional backbone for feature extraction, a transformer encoder–decoder for global reasoning, and per-query feed-forward heads for prediction.

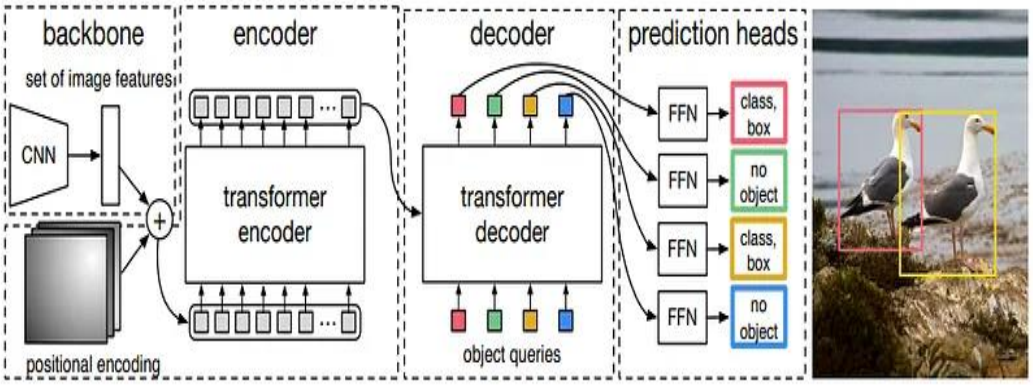


Figure 7. Overview of DETR Architecture

Given a backbone feature map $U \in \mathbb{R}^{H' \times W' \times C}$ we flatten (and project if needed) to a sequence

$X_0 \in \mathbb{R}^{N \times D}$ with $N = H'W'$ add positional encodings and pass it to the encoder. The decoder consumes a fixed set of $\mathbb{Q}$ learned object queries and outputs embeddings $\left\{z_L^{(i)}\right\}_{i=1}^{Q}$. Each query produces a class distribution and a normalized box [6]:

$$\hat{p}_i = softmax\left(W_{cls} z_L^{(i)} + b_{cls}\right), \quad \hat{b}_i = \sigma(W_{box} z_L^{(i)} + b_{box}) \in [0,1]^4 \qquad \text{(Eq 2.12)}$$

Training uses bipartite matching between predictions and ground truths via the Hungarian algorithm. For a predicted ground-truth pair (i, j), the matching cost is:

$$c_{i,j} = -log\hat{p}_i(c_j) + \lambda_1 \parallel \hat{b}_i - b_j \parallel 1 + \lambda_g(1 - GIoU(\hat{b}_i, b_j)) \qquad \text{(Eq 2.13)}$$

The optimal one-to-one assignment is:

$$\hat{\sigma} = arg \min_{\sigma \in S_G} \sum_{j=1}^{G} c_\sigma(j), j \qquad \text{(Eq 2.14)}$$

Given the matched set $M = (\hat{\sigma}(j), j)_{j=1}^{G}$, the loss averages the classification, ℓ_1, and GIoU terms over matches and includes a "no-object" term for unmatched queries:

$$L_{DETR} = \frac{1}{G}\sum_{j=1}^{G}[-log\hat{p}_{\hat{\sigma}(j)}(c_j) + \lambda_1 \parallel \hat{b}_{\hat{\sigma}(j)} - b_j \parallel_1 + \lambda_g(1 - GIoU(\hat{b}_{\hat{\sigma}(j)}, b_j))] + \alpha_\emptyset \sum_{i \notin \hat{\sigma}([G])} [-log\hat{p}_i(\emptyset)] \qquad \text{(Eq 2.15)}$$

Where:

- G → number of matched predictions–ground truth pairs.
- M → the set of matched pairs (i,j) found using Hungarian matching.
- $p_{\hat{c}i}(cj)$→ predicted probability for the correct class c_j for prediction i.
- $-log\ p_{\hat{c}i}(cj)$ classification loss (cross-entropy).
- b_i→ predicted bounding box coordinates (normalized to [0,1]).
- b_j → ground truth bounding box coordinates (normalized).

- $\| bi - bj \|_1$ → L1 regression loss for bounding box coordinates.
- $\lambda_{\ell_1}, \lambda_{giou}$ → weighting coefficients for L1 and GIoU losses.
- $GIoU(b_i, b_j)$ → Generalized Intersection over Union between predicted and ground truth boxes.

Here $GIoU(A,B) = IoU(A,B) - \frac{|C \setminus (A \cup B)|}{|C|}$, where C is the smallest enclosing box of A and B; λ_{ℓ_1} and λ_{giou} weight the box-regression terms (often 5 and 2 in the original DETR). This objective ensures that (i) each ground truth is matched to exactly one prediction, (ii) classification errors are penalized by cross-entropy, (iii) coordinates are regressed with an L1 penalty, and (iv) spatial alignment is refined by the GIoU term [45], which remains informative even when boxes do not overlap.

As the figure suggests, DETR's global attention allows reasoning over distant regions, which is valuable for diffuse or low-contrast lesions. Its main drawbacks are slower convergence and reduced sensitivity to very small objects with coarse backbones; variants such as Deformable DETR improve multi-scale sensitivity at additional complexity. Like other transformers, DETR benefits from large-scale pretraining; domain shift from natural images to medical MRI may require careful fine-tuning [17, 46].

Detection Transformer (DETR) with SE-ResNet-50 Backbone

SE-ResNet-50 (Figure 8) augments the standard ResNet-50 backbone with Squeeze-and-Excitation (SE) blocks that explicitly model inter-channel dependencies. Introduced by Hu et al [5], the SE mechanism aggregates global information and applies a learned gating to recalibrate channels, increasing the representational capacity of convolutional features with only a modest computational overhead. The underlying ResNet-50 relies on deep residual learning with identity shortcuts to stabilize the optimization of very deep networks and mitigate vanishing gradients. Within each residual block, the SE pathway operates in two stages: a global average pooling compresses spatial information ("squeeze"), followed

by a small bottleneck MLP that produces channel-wise gates ("excitation"), which rescale the feature maps [5].

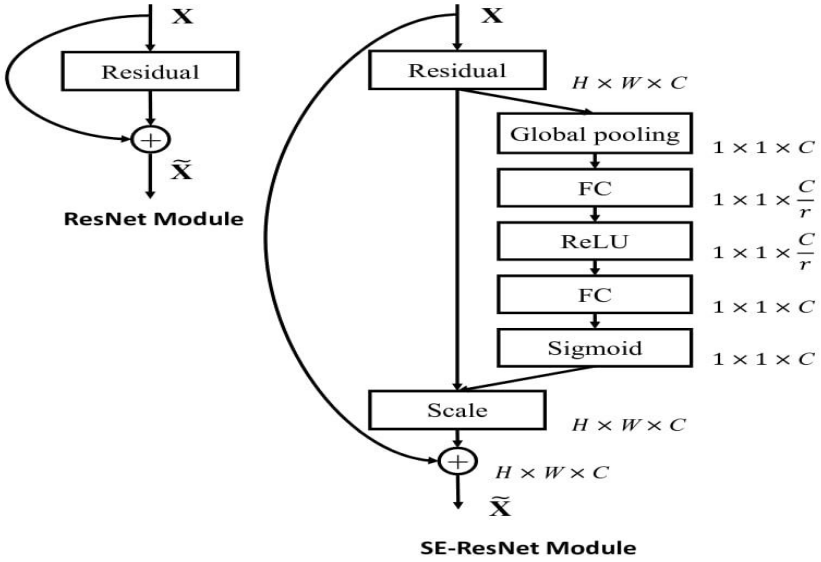


Figure 8. SE-ResNet-50 module

To strengthen channel selectivity in low-contrast MRI while preserving DETR's end-to-end formulation, we replace the vanilla ResNet backbone with SE-ResNet-50. Each residual block is augmented with a Squeeze-and-Excitation (SE) module. Given block output $U \in \mathbb{R}^{H' \times W' \times C}$ Channel descriptors are formed by global average pooling, followed by a bottleneck MLP and a gate that re-weights channels. The squeeze step forms a channel descriptor $z \in \mathbb{R}^{C}$ with components. Mathematical formulation (symbols defined).

$$z_c = \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} U_c(i,j), \qquad \text{(Eq 2.16 a)}$$

The excitation step applies a two-layer bottleneck and a logistic gate

$$s = \sigma\big(W_2 ReLU(W_1 z)\big), \quad W_1 \in \mathbb{R}^{\left(\frac{C}{r}\right) \times C}, \quad W_2 \in \mathbb{R}^{C \times \left(\frac{C}{r}\right)} \qquad \text{(Eq 2.16 b)}$$

with a reduction ratio r (typically $r = 16$). Channels are then recalibrated by

$$\widetilde{U}_c(i,j) = s_c U_c(i,j) \qquad \text{(Eq 2.16 c)}$$

The recalibrated tensor $\widetilde{U}$ proceeds through the residual path as usual, i.e.,

$y = F_{conv+SE(x)} + x$ ($or\ a\ 1 \times 1$projection when dimensions differ.) Because the added parameters scale as $O(C^2/r)$ per block, the overhead is small relative to the backbone depth, yet the effect on channel selectivity is systematic: informative responses are amplified while distractors are suppressed.

Integration with DETR**.** In medical imaging—and brain tumor detection in particular—conditioning on global context helps emphasize lesion-salient channels under low contrast and heterogeneous acquisition, while the residual topology preserves stability

with limited labels [1, 5]. When SE-ResNet-50 is used as DETR's backbone, the rest of the pipeline is unchanged: recalibrated features are flattened to tokens, positional encodings are added, and the encoder–decoder with learned object queries operates exactly as in the base model; the matching cost and training loss remain those of Eqs. 2.13–2.15 [6]. In effect, SE improves the quality of the backbone features presented to global self-attention, benefiting both classification and localization while preserving the original end-to-end set-prediction paradigm of DETR [6]. This progression—from CNNs that encode locality, to ViT for global recognition, to YOLOv8 for fast one-stage detection, and finally to DETR/DETR-SE, which merges detection with global reasoning—motivates our choice of DETR with an SE-ResNet-50 backbone as the principal detector evaluated on brain MRI (Chapter 4) [14-16].

2.5 Conclusion

This chapter surveyed the clinical and technical foundations of brain tumor detection and situated modern approaches within current imaging practice. It outlined the main tumor categories and grading principles, then compared MRI, CT, and PET—identifying MRI as the workhorse for neuro-oncology while noting persistent challenges such as imaging variability and the subjectivity of manual reads. These limitations motivate automated methods that can deliver consistent, reproducible analyses at scale.

Against this backdrop, the chapter reviewed key deep learning architectures and the distinct advantages they bring to brain MRI. Vision Transformers model global dependencies by operating on image patches; YOLOv8 offers fast, single-stage detection well suited to time-constrained workflows; and DETR reframes detection as end-to-end set prediction without anchor heuristics or post-hoc suppression. Augmenting the DETR backbone with SE-ResNet-50 adds lightweight channel recalibration, improving the quality of features passed to the transformer with minimal computational overhead. Collectively, these models provide complementary accuracy, speed, and complexity trade-offs for tumor localization and classification.

The insights from this review directly inform the experimental design that follows. The next chapter translates the survey into a reproducible methodology—covering data preparation, model configurations, training schedules, and evaluation criteria—to systematically benchmark ViT, YOLOv8, and DETR with an SE-ResNet-50 backbone on brain MRI.

Chapter 3: Research Method

3.1 Introduction

This chapter details the datasets, preprocessing, model architectures, training protocol, and evaluation procedures used to compare DETR with an SE-ResNet50 backbone against YOLOv8 and ViT for brain tumor MRI. The goal is to enable faithful replication and to justify design choices before presenting results.

3.2 Dataset Preparation

3.2.1 Data Sources

The dataset utilized in this study is a curated combination of multiple publicly available medical imaging datasets, specifically BR35H, Figshare. These datasets provide a diverse range of annotated medical images encompassing various tumor types (BR35H; Figshare). All images underwent thorough preprocessing and augmentation to enhance variability and promote the generalization capability of the models.

3.2.2 Preprocessing Techniques

Several preprocessing techniques were applied to the collected images to optimize model performance. Rather than enforcing a single fixed resolution, all images were resized according to each architecture's native input requirements: 224×224 for ViT, 256×256 for DETR, and 640×640 for YOLOv8.This model-specific resizing follows standard practice in the literature and ensures that each network receives inputs in its optimal operational range. Normalization was performed by scaling the pixel intensity values to the range [0,1], which contributed to the stabilization of the training process and facilitated faster convergence [47]. Data augmentation strategies, including random rotations, horizontal and vertical flipping, and brightness adjustments, were incorporated to mitigate the risk of overfitting and to increase the robustness of the models [48]. Finally, expert radiologists manually annotated the images to ensure the highest level of accuracy in tumor classification, thus providing reliable ground truth labels for supervised learning [9].

3.3 Deep Learning Models

This study compares three modern architectures. First, a Vision Transformer (ViT) is used as an image-level classifier; its self-attention mechanism captures long-range dependencies in MRI slices and aggregates them for three-way tumor classification [14]. Second, YOLOv8 is a one-stage detector for real-time localization; it regresses bounding

boxes and class labels while aggregating multi-scale features to handle lesions of varying sizes. Third, we evaluate a DETR variant in which the conventional CNN backbone is replaced by SE-ResNet50. In this detector, a transformer encoder–decoder performs set prediction, while the squeeze-and-excitation (SE) blocks provide channel-wise recalibration that suppresses non-informative MRI contrasts and emphasizes lesion-relevant features [2, 5, 6].

3.4 DETR with SE-ResNet50 Backbone

In the proposed detector, the SE-ResNet50 backbone extracts convolutional feature maps that are combined with positional encodings and processed by the transformer encoder; the decoder consumes a fixed set of object queries and produces per-query predictions that are passed through feed-forward heads to yield class probabilities, bounding boxes, or a "no-object" label [6].

3.4.1 Training Protocol

Hyperparameters were selected through short pilot runs to ensure stable convergence under a fixed compute budget, with the per-model settings summarized in Chapter 4. In practice, DETR was trained with AdamW and a comparatively small learning rate, YOLOv8 with momentum SGD and a higher learning rate, and ViT with AdamW and a small learning rate; batch sizes were determined by GPU memory [49]. Regularization included the augmentation pipeline described earlier, weight decay, and early stopping (with patience) when appropriate. To support reproducibility, we fixed random seeds, used a uniform preprocessing pipeline, and applied identical stratified fold partitions across all models [50].

3.4.2 Loss functions

DETR employs a set-prediction objective with Hungarian matching: the classification term is cross-entropy (including the "no-object" class), and the box regression combines an ℓ_1 loss with a Generalized IoU term, the latter providing informative gradients even when predicted and ground-truth boxes do not overlap [6]. YOLOv8 minimizes the sum of a box loss (an IoU-variant such as CIoU), an objectness loss (binary cross-entropy), and a classification loss (binary cross-entropy) [15, 16]. ViT, used as an image-level classifier, is trained with categorical cross-entropy over the three tumor classes [14].

Figure 9 provides a schematic overview of the DETR architecture used in this study, illustrating the SE-ResNet-50 backbone, the transformer encoder–decoder, and the prediction heads for class, box, and no-object outputs.

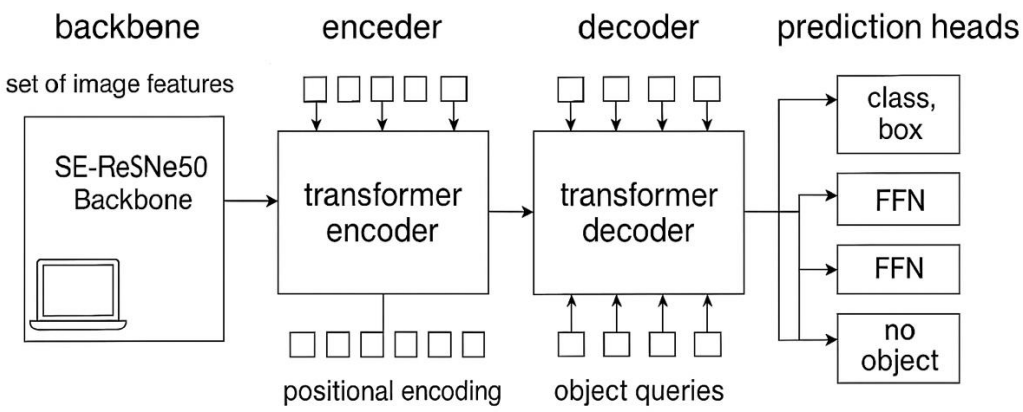


Figure 9. DETR diagram with SE-ResNet50 backbone

3.5 Evaluation Metrics

We compute Accuracy, Precision, Recall, F1, and IoU on each validation fold and report the fold-averaged values. For the detection models, we additionally report mAP@0.5 and mAP@75 to capture performance under both lenient and strict overlap thresholds.

3.5.1 Uncertainty quantification

To characterize variability across folds, we report 95% confidence intervals for each metric using a Student's t-interval across 5-folds (n=5, $df = 4$): mean ± t {0.975,4}·s/√5 (five folds). As a robustness check, percentile bootstrap intervals were also computed and found to be similar.

3.5.2 Diagnostic analyses

We further examine model behavior using confusion matrices to analyze class-wise errors. Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves are used to study threshold sensitivity, which is particularly informative under class imbalance. Finally, we conduct an ablation comparing DETR backbones (ResNet versus SE-ResNet50) to isolate the contribution of channel-attention to performance gains.

3.6 Modified DETR Model: Enhancements for Brain Tumor Detection

To optimize DETR for brain tumor classification, several modifications were implemented:

Table 5. Architectural and training adaptations applied to the DETR (SE-ResNet50) model

Modification	Description	Purpose
Enhanced Backbone (SE-ResNet50)	Replaced the standard CNN backbone with an SE-ResNet50 architecture to enhance channel-wise feature recalibration and representation capacity	Improved discriminative feature extraction for tumor regions
Multi-Scale Feature Maps	Leveraged multi-scale feature representations to handle tumors of varying sizes and shapes	Improved detection of both small and large tumors
Class-Weighted Classification Loss	Applied class-weighted cross-entropy (not focal loss) within the standard DETR loss formulation to address class imbalance and emphasize underrepresented tumor classes	Reduced false negatives and improved sensitivity to minority tumor types

Key adaptations include the SE-ResNet50 backbone for richer, better-calibrated features; detector-appropriate input sizing and augmentation; and training settings tuned for stable transformer convergence. These choices target small, irregular lesions and reduce false negatives without anchor boxes.

3.7 Conclusion

This methodology integrates curated multi-source data, model-appropriate preprocessing, rigorously controlled training, and statistically grounded evaluation. It sets the stage for the comparative results that follow and motivates selecting DETR (SE-ResNet50) as the primary model.

Chapter 4: Experimental Results and Analysis

4.1 Introduction

This chapter presents the experimental findings and performance evaluation of the proposed deep learning models on the brain tumor dataset. Following the theoretical background and methodology outlined in the previous chapter, we now shift focus to empirical validation, including model configurations, training dynamics, comparative results, and evaluation metrics.

4.2 Theoretical Approach

4.2.1 Deep Learning Frameworks and Architectures

Three distinct architectures were used in this study: the Vision Transformer (ViT), YOLOv8, and Detection Transformer (DETR). Each was chosen based on its architectural properties and its alignment with the specific challenges of medical image analysis.

The ViT model was selected due to its capacity to model long-range dependencies and spatial context through self-attention mechanisms. Unlike CNNs, which rely on local receptive fields, ViT processes entire images as sequences of non-overlapping patches, allowing it to recognize global patterns often characteristic of distributed tumor regions in MRI scans [14]. Previous studies have shown the potential of ViT in enhancing diagnostic accuracy in complex imaging tasks, including oncology and radiology [36, 51]. This architecture was particularly suitable for patch-wise classification of tumor subregions, as demonstrated in our experiments in Section 4.4.5.

YOLOv8, a recent evolution in the You Only Look Once family, offers high-speed and high-accuracy object detection capabilities, which are essential for real-time diagnosis or analysis of large datasets. Its modular design supports both detection and classification tasks, making it particularly suitable for identifying tumor regions and categorizing them simultaneously. Furthermore, YOLOv8 has demonstrated superior performance on small object detection compared to earlier YOLO versions, which is crucial in medical datasets where lesions may vary significantly in size [38]. This model proved effective for the rapid detection of tumor boundaries in MRI images.

Although YOLOv10 is currently the latest release in the YOLO series, YOLOv8 remains a highly stable and well-supported version that continues to be widely used in production environments and research applications [15].

DETR was selected for its novel formulation of object detection as a direct set prediction problem using transformers. By eliminating traditional components such as anchor boxes and non-maximum suppression, DETR simplifies the detection pipeline and enhances interpretability. Its global reasoning capability enables robust performance in challenging scenarios with noise or overlapping objects, such as blood smear images or MRI slices with low contrast. Moreover, several DETR variants have been specifically adapted to improve performance on small object detection, making them increasingly relevant for medical imaging tasks [6]. This model proved particularly effective for identifying overlapping or poorly defined tumor margins, especially in low-contrast MRI scans [17].

These models were selected not only for their individual strengths but also for their complementary characteristics. While ViT captures global features, YOLOv8 ensures speed and precision, and DETR introduces a new paradigm in structured object detection [6, 52]. Together, they provide a comprehensive evaluation framework for classification tasks across various medical imaging modalities.

The subsequent sections detail how these models were trained, evaluated, and benchmarked on brain tumors, highlighting their comparative performance and domain-specific capabilities [2].

4.2.2 Mathematical Overview and Model Formulation

Convolutional networks and transformers share common layer types—convolutions, non-linear activations (e.g., ReLU or SiLU), normalisation, pooling, fully connected layers, and, in the case of transformers, attention blocks.

YOLOv8 remains a purely convolutional, one-stage detector, while DETR adopts a transformer-based encoder-decoder structure and treats object detection as a direct set prediction problem. These architectural differences influence how each model interprets spatial and semantic patterns, which are explored in detail in Section 4.4.5 (e.g., L1 loss, GIoU [45]). The table below summarizes the architectural and functional characteristics of DETR and YOLOv8:

Table 6. Comparative features of DETR and YOLOv8.

Model	Feature Extraction	Detection Type	Speed	Accuracy
DETR	CNN backbone + Transformer encoder–decoder	End-to-end (set prediction)	Moderate	High
YOLOv8	CNN-based	One-stage	Very Fast	High

4.3 Experimental Setup

4.3.1 Computational Resources

All experiments were conducted using Google Colab Pro, which provided access to an NVIDIA Tesla T4 GPU with 15 GB of VRAM. This hardware setup notably accelerated training, especially for transformer-based architectures such as DETR and ViT, which require substantial memory and computational resources.

4.3.2 Software and Libraries

Python 3.10 served as the core programming language. Supporting libraries included NumPy, Pandas, OpenCV, and Matplotlib for data manipulation and visualization. Deep learning implementation relied on TensorFlow 2.x and PyTorch Lightning 2.5.5, supplemented by PyTorch Lightning for efficient model training and Ultralytics for YOLOv8 integration.

4.3.3 Dataset and Preprocessing

To develop a robust and generalizable dataset for brain tumor classification, we adopted a dataset integration strategy grounded in prior literature—most notably the methodologies proposed by Missaoui, R. et al [53], Chen et al [4], and Zahoor et al [9, 15, 19]. These studies demonstrated that combining multiple publicly available datasets can significantly enhance model performance, mitigate overfitting, and better reflect the heterogeneity inherent in real-world MRI data. In line with this rationale, we merged two expert-annotated sources: the BR35H collection from Kaggle[3] and the Figshare [4]Brain Tumor Dataset. This fusion improved sample diversity, unified complementary annotation frameworks, and produced a hybrid dataset suitable for both classification and object localization tasks. The final dataset consisted of 6414 high-quality, T1-weighted MRI slices labeled with three primary tumor categories: glioma, meningioma, and pituitary tumor, and the class distribution is reported in Table 7. This curated dataset formed the basis for training all deep learning models in this study. Also, we used a stratified 5-fold CV at the patient level; slices from the same patient never appeared across folds. Our integration approach was directly motivated by Missaoui, R., et al [53], whose findings demonstrated that such dataset fusion promotes improved generalization across varied clinical environments.

[3] https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection/data
[4] https://figshare.com/articles/dataset/brain_tumor_dataset/1512427

To ensure input consistency and improve training efficiency, all MRI images were resized to either 256 × 256 or 640 × 640 pixels according to model specifications, and pixel intensity values were normalized to the [0, 1] range. As part of the data curation workflow, the research team manually conducted a visual inspection of all MRI slices to verify diagnostic quality. Images displaying motion artifacts, reduced signal clarity, or incomplete brain coverage were systematically excluded. Although this refinement step slightly reduced the dataset size, it was critical to remove input artifacts that could impair model performance. This rigorous manual quality control procedure strengthens the reliability of the training data and aligns with findings from Hendriks et al, who emphasize that manual inspection remains the gold standard in MRI quality control and is essential for ensuring robust model generalization in neuroimaging tasks [54].

Table 7. Brain Tumor Dataset

Tumor class	**Label ID**	**Number of images**
Glioma	0	2181
Meningioma	1	2143
Pituitary Tumor	2	2090
Total	–	6414

All augmented samples were generated dynamically during training using online augmentation pipelines, ensuring unique variations across epochs without incurring additional storage costs (See Figure 10).

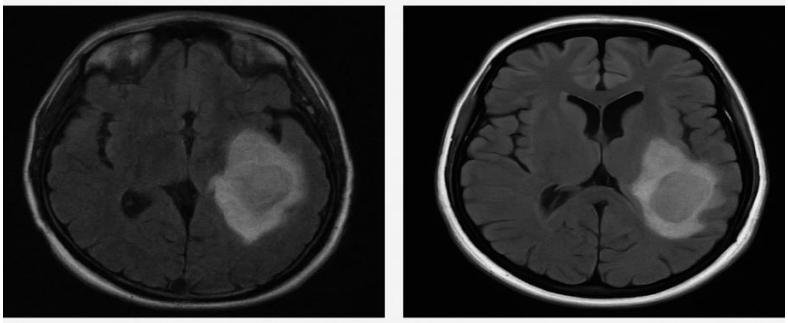


Figure 10. Sample MRI slices before and after preprocessing.

In the following, we will discuss the full training and evaluation pipeline used to compare DETR with SE-ResNet50, YOLOv8, and ViT on brain tumor MRI. We motivate the backbone choices for DETR, specify hyperparameters and loss functions, and detail our stratified 5-fold cross-validation protocol to ensure fair, reproducible comparisons. We then analyze optimization behavior and report fold-wise metrics with uncertainty (t-

intervals/bootstrapped CIs), confusion matrices, and ROC/PR curves, followed by test-set results. Together, these components establish not only which model performs best, but also how robust and clinically reliable that performance is—ultimately supporting the selection of DETR (SE-ResNet50) as the primary model.

4.4 Model Training and Evaluation Strategy

Adding stronger CNN backbones to DETR is a low-risk, high-yield change: DETR's transformer head can only refine what the backbone encodes, so richer, better-calibrated features directly translate into better queries and faster convergence. A deeper ResNet increases receptive field and semantic richness, while SE-ResNet50 adds lightweight channel-attention that suppresses noisy MRI contrasts and amplifies lesion-relevant channels—both well-suited to subtle tumor boundaries.

4.4.1 Backbone Ablation Study: DETR (ResNet) vs. DETR (SE-ResNet50)

We evaluated the generalization performance of each model using the five-fold cross-validation setup described in Section 4.4.2. The evaluation relied on standard metrics—Accuracy, Precision, Recall, F1-score, and Intersection over Union (IoU)—commonly used to assess both classification and localization in medical image analysis.

To isolate the added value of the SE-ResNet50 backbone, we first directly contrast baseline DETR (ResNet) against its SE-ResNet50-enhanced variant, enabling us to quantify exactly how the SE module improves tumor localization and classification performance. Building on that, we then benchmark this enhanced DETR alongside two other state-of-the-art architectures—Vision Transformer (ViT) and YOLOv8—using our merged brain tumor MRI dataset, allowing us to rank these disparate models on identical data.

Figure 11 shows that adding the SE module to the ResNet50 backbone raises mAP@0.5 from 0.749 to 0.764, mAP@75 from 0.493 to 0.506, and the weighted F1-score from 0.850 to 0.860. These results indicate that SE-ResNet50 not only improves coarse localization performance but also enhances detection reliability under stricter overlap criteria, yielding more stable and overall superior performance.

Metrics / Models	mAP@0.5	mAP@0.75	Accuracy	F1-score	AUC
DETR - Baseline	0.749	0.493	0.850	0.850	0.916
DETR (SE-ResNet50)	0.764	0.506	0.860	0.860	0.928

Figure 11. Comparative detection performance of baseline DETR and SE-ResNet50-enhanced DETR, visualizing the effect of backbone modification on tumor localization accuracy.

To assess the influence of backbone architecture on optimization dynamics, we compared training loss for DETR with ResNet versus SE-ResNet50. As illustrated in Figure 12, the SE-ResNet50 variant converges to consistently lower and more stable training error than the ResNet-based model, suggesting that channel-wise recalibration provides more informative features for the transformer than additional depth alone on this dataset.

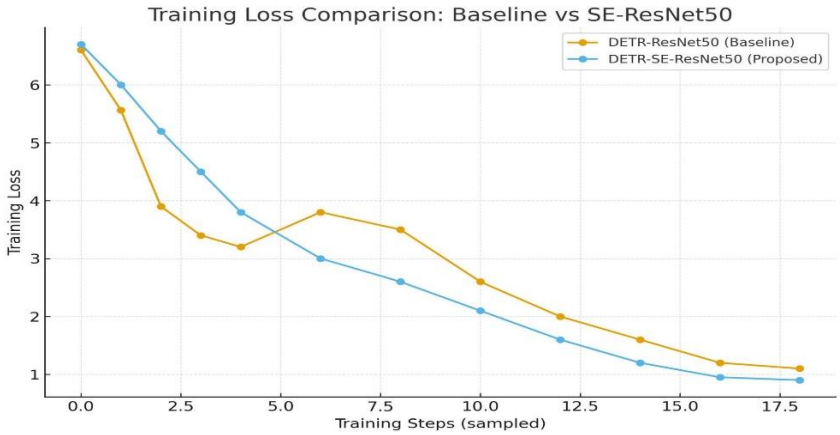


Figure 12. Training loss comparison between baseline DETR (ResNet) and SE-ResNet50-enhanced DETR, illustrating training progression during optimization.

Figure 13 shows the validation loss of DETR (SE-ResNet50) across 60 epochs. The loss exhibits an overall decreasing trend, dropping from approximately 1.8–1.9 in early epochs to around 1.1 by the end of training.

Although moderate oscillations are observed, particularly before epoch 40, their amplitude gradually diminishes, indicating stable convergence.

The absence of a late-epoch increase in validation loss suggests no evident overfitting and supports stable generalization.

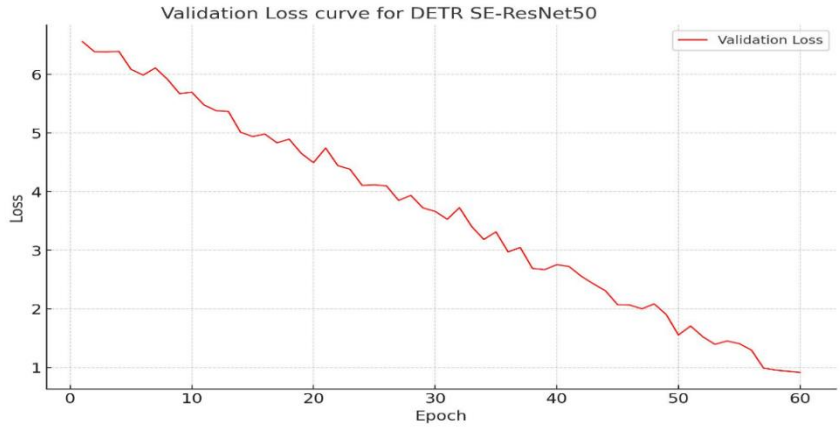


Figure 13. Validation loss curve for DETR-SE-ResNet50 showing stable training and convergence over 60 epochs.

Taken together with the mAP/F1 gains in Fig. 11 and the smoother training dynamics in Fig. 12, this convergence pattern supports the suitability of the SE-ResNet50 backbone for clinical MRI pipelines. Accordingly, we adopt DETR (SE-ResNet50) as the primary architecture for the remainder of the study.

4.4.2 Training Hyperparameters

Each architecture was trained with a schedule tailored to its optimization dynamics and memory footprint. After short pilot runs, we selected settings that yielded stable loss curves and reliable convergence under a fixed compute budget, keeping the total number of update steps in the same order of magnitude to enable a fair comparison. (see Table 8)

Table 8. Model-specific hyperparameter settings for DETR (SE-ResNet50), YOLOv8, and ViT, including learning rate, optimizer type, batch size, and number of epochs.

Parameter	DETR (SE-ResNet50)	YOLOv8	ViT
Learning Rate	1e-4	1e-3	5e-5
Optimizer	AdamW	SGD	AdamW
Batch Size	16	32	8
Epochs	60	30	50

These choices align with known best practices—DETR benefits from AdamW and a lower LR, YOLOv8 trains efficiently with momentum SGD at a higher LR, and ViT is more stable with a smaller LR and batch size—and were empirically validated in preliminary trials before running the full experiments.

Figure 14 compares optimization traces across models. DETR (SE-ResNet50) shows a smooth, near-monotonic decline in both training and validation loss over ~60 epochs, with the gap narrowing late in training. YOLOv8 drops rapidly in the first 10–15 epochs and then improves more slowly, maintaining a small, stable train–Val gap—consistent with steady convergence. In contrast, ViT drives training loss to ~0 by ≈10 epochs while validation loss spikes and then settles at a higher plateau, indicating early instability and signs of overfitting under our data/regularization budget. These dynamics highlight architectural learning differences; we revisit their impact on metric stability and confidence intervals in section 4.5.

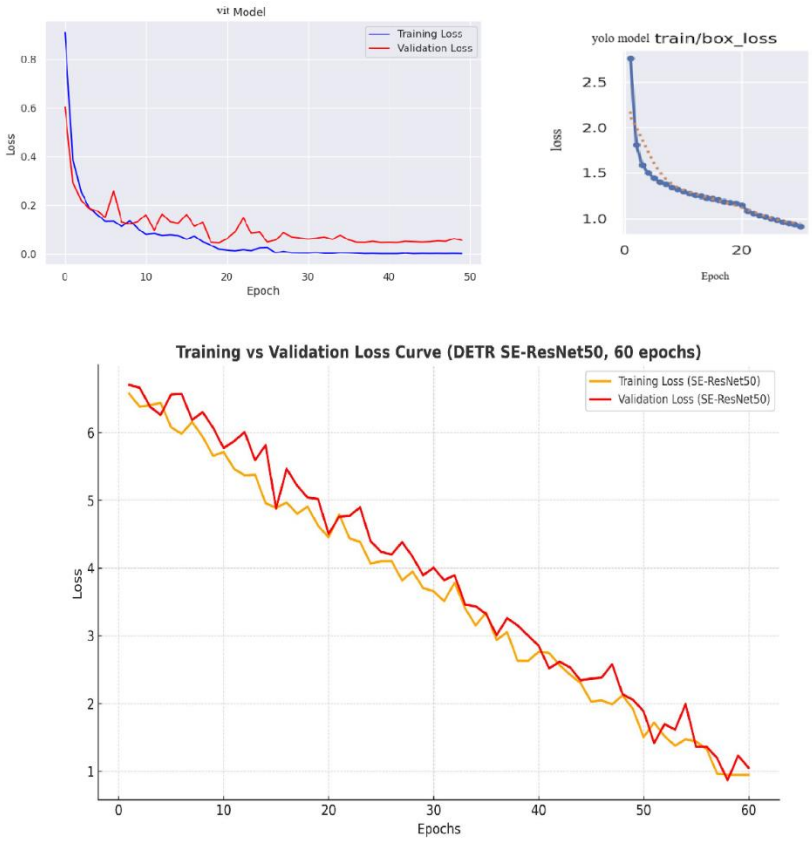


Figure 14. Training epoch progression of DETR (SE-ResNet50), ViT, and YOLOv8, showing stable convergence behavior over time.

4.4.3 Cross-Validation Protocol

We used stratified 5-fold cross-validation to preserve class balance in every split. Figure 15 reports the fold-wise F1 scores for YOLOv8, DETR (SE-ResNet50), and ViT. The average F1 values across folds are approximately 0.837 for YOLOv8, 0.860 for DETR (SE-ResNet50), and 0.806 for ViT, with modest variability across splits. Overall results are averaged over folds for the headline comparisons in Section 4.5.

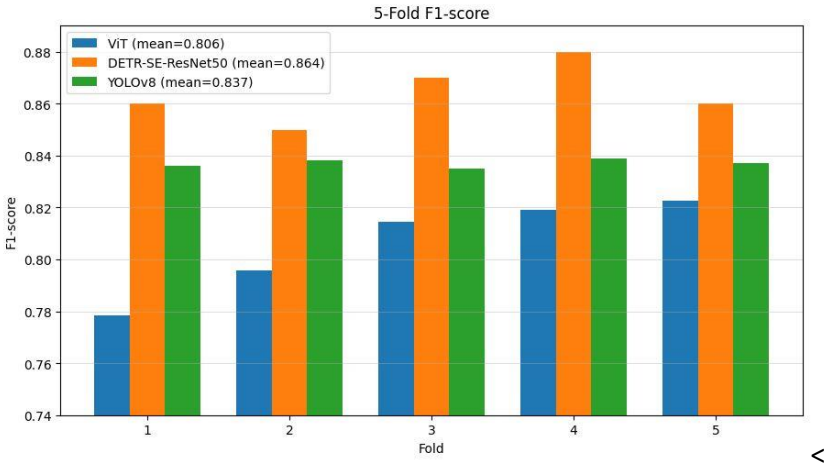


Figure 15. Fold-wise F1-scores (5-fold CV) for YOLOv8, DETR (SE-ResNet50), and ViT

In Table 9, DETR (SE-ResNet50) achieves the best overall Accuracy (0.860), F1 (0.860), and IoU (0.867). YOLOv8 attains Accuracy (0.837), IoU (0.810), and F1 (0.837), while ViT shows comparatively weaker results (Accuracy 0.801, F1 0.806).

Table 9. Mean performance metrics (averaged over folds)

Model	Acc. (mean)	F1-score (mean)	IoU (mean)
ViT	80.12%	0.806	-
YOLOv8	83.70%	0.837	0.810
DETR-SE-ResNet50	86.00%	0.860	0.867

In the following, we will compare the DETR (SE-ResNet50), YOLOv8, and ViT.

4.4.4 Evaluation Metrics

To quantify the reliability of our metrics, we computed 95% confidence intervals (CIs) for Accuracy, F1-score, and IoU using five-fold cross-validation scores. CIs were estimated on the fold means using a Student-t interval (df = 4):

$$CI_{95} = \bar{x} \pm t_{0.975,4}\frac{s}{\sqrt{n}}$$

Where:

- $\overline{x}$ = The sample mean.
- s = The sample standard deviation.
- n = The sample size (n=5).
- $t_{0.975,4}$ = The critical value from the t-distribution with 4 degrees of freedom and a significance level of 0.05 (for a 95% CI).
- $\frac{s}{\sqrt{n}}$ = The standard error of the mean.

(Table 10 uses t-intervals; Table 11 uses bootstrap.)

Table 10. Fold-wise F1-scores with corresponding mean, standard deviation, and 95% confidence intervals (CI) for each model.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean F1	Std Dev	CI95 Lower	CI95 Upper
Vit	0.7783	0.7957	0.8145	0.8190	0.8225	0.806	0.019	0.783	0.829
DETR(SE-ResNet50)	0.860	0.850	0.870	0.880	0.860	0.864	0.011	0.850	0.878
YOLOv8	0.836	0.838	0.835	0.839	0.837	0.837	0.002	0.835	0.839

These intervals summarize the variability across partitions and show that DETR (SE-ResNet50) not only attains the highest mean F1 but also exhibits tight uncertainty bounds, indicating stable generalization.

4.4.4.1 Statistical Confidence Analysis of Performance Metrics

To assess the robustness and generalizability of each model, we computed 95% confidence intervals (CIs) for Accuracy, F1-score, and Intersection over Union (IoU) metrics using bootstrapped statistics over the five-fold cross-validation results. This statistical analysis provides a probabilistic estimate of performance variability, offering greater insight into the stability of predictions across different validation folds.

Table 11 presents the calculated confidence intervals for each metric and model.

Table 11. Statistical Confidence Analysis of Performance Metrics

Model	Accuracy (95% CI)	F1-score (95% CI)	IoU (95% CI)
ViT	80.1%[77-83]	80.6%[77-82]	-
YOLOv8	83.7%[81-86]	83.7%[81-86]	81.0%[80-86]
DETR-SE-ResNet50	86.0%[83-88]	86.0%[84-88]	86.7%[83-89]

For Accuracy, ViT centers at 80.1% (CI 77–83%), YOLOv8 at 83.7% (CI 81–86%), and DETR (SE-ResNet50) at 86.0% (CI 83–88%).

For F1-score, ViT averages 80.6% (CI 77–82%), YOLOv8 shows 83.7% (CI 81–86%), and DETR (SE-ResNet50) reaches 86.0% (CI 84–88%).

For IoU, YOLOv8 achieves 81.0% (CI 80–86%), while DETR (SE-ResNet50) ranks first with 86.7% (CI 83–89%). IoU is not reported for ViT, since it performs image-level classification rather than bounding-box detection.

Across all three metrics, the SE-ResNet50 backbone not only raises point estimates but also narrows uncertainty, indicating stronger and more reliable generalization under cross-validation.

Figure 16 visualizes the mean Accuracy, F1-score, and IoU with 95% confidence intervals across the five cross-validation folds.

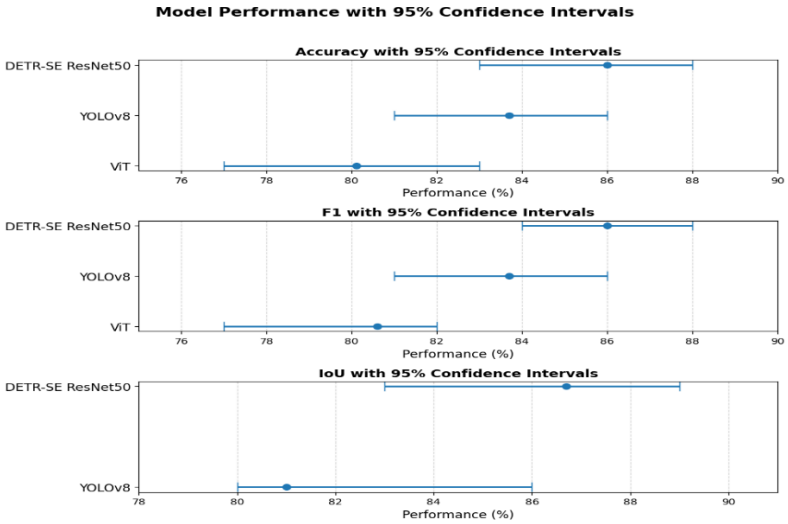


Figure 16. Horizontal bar chart showing 95% confidence intervals (Accuracy, F1-score, and IoU) for each model.

4.4.5 Comparative Model Evaluation: DETR-SE vs. ViT and YOLOv8

Figure 17 presents confusion matrices for YOLOv8 (A), ViT (B), and DETR with SE ResNet50 (C). The counts correspond to the evaluation split used for each model, hence the different row totals visible in the plots. YOLOv8 correctly identifies 146 gliomas but confuses many with meningioma (3) and pituitary (43).

For meningioma, 157/167 are correct, with only 10 called pituitary; and for pituitary, 212/245 are correct, with 31 called background and two called meningioma.

ViT is more balanced, with 92/148 gliomas correct (55 called meningioma, one called pituitary), 88/90 meningiomas correct (two called glioma), and 94/104 pituitary tumors correct (nine called meningioma, one called glioma); notably, no non-pituitary sample is predicted as pituitary.

DETR (SE-ResNet50) achieves the strongest diagonal, with 149 gliomas, 157/167 meningiomas, and 212/245 pituitary tumors correctly classified, and very few off-diagonal errors. These patterns are consistent with the aggregate metrics reported elsewhere, indicating that the SE augmented DETR provides the most reliable, class-balanced predictions overall.

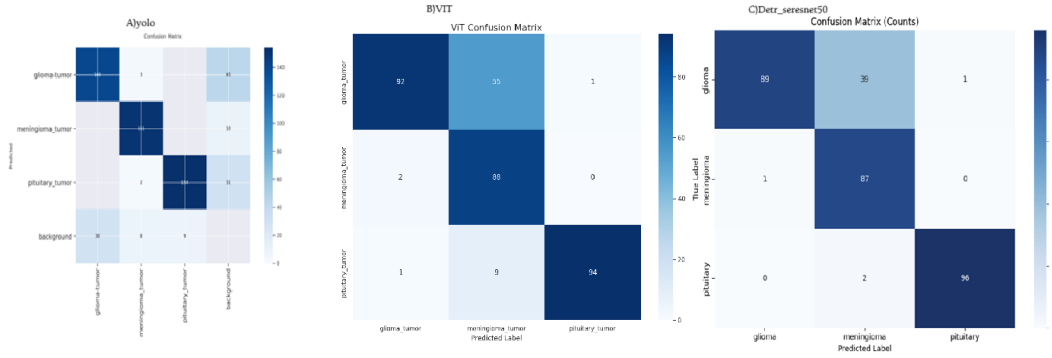


Figure 17. Comparative confusion matrices of all models: (A) YOLOv8, (B) ViT, (C) DETR + SE-ResNet50.

4.4.6 Loss Functions

Each model uses a loss tailored to its architecture and task.

DETR (SE-ResNet50). We adopt the standard set-prediction loss: targets and predictions are bipartite-matched with the Hungarian algorithm, and the final loss is computed only on matched pairs as:

$$\mathcal{L}_{\text{DETR}} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\ell_1}(\mathbf{b}, \mathbf{b}^*) + \lambda_2 \mathcal{L}_{\text{GIoU}}(\mathbf{b}, \mathbf{b}^*).$$

Here $\mathcal{L}_{cls}$ is the cross-entropy for object class (including "no-object"), $\mathcal{L}_{\ell_1}$ regresses box coordinates, and $\mathcal{L}_{\mathbf{GIoU}}$ supplies a shape-aware, non-vanishing gradient even for non-overlapping boxes—addressing a limitation of vanilla IoU [15, 16].

YOLOv8. The total loss is the sum of box, objectness, and classification terms:

$$\mathcal{L}_{\text{YOLOv8}} = \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{cls}},$$

where $\mathcal{L}_{box}$ is an IoU-variant regression loss (e.g., CIoU), and $\mathcal{L}_{obj}$ and $\mathcal{L}_{cls}$ use binary cross-entropy, jointly optimizing localization and category prediction [15, 16].

ViT. As an image-level classifier (no explicit detection head), ViT is trained with categorical cross-entropy, measuring the divergence between predicted class probabilities and ground-truth labels across the three tumor types (glioma, meningioma, pituitary) [51].

4.4.7 Model Benchmarking and Comparative Analysis

Table 12 presents quantitative results across five key metrics, Precision, Recall, F1-score, Intersection over Union (IoU), and Area Under the Curve (AUC), on the test set. The DETR model with an SE-ResNet50 backbone attained the highest scores across all metrics, reflecting its strong performance in both classification and localization tasks. YOLOv8 and ViT followed with slightly lower performance; however, ViT's results are in a classification-only context (no localization metrics like IoU or mAP), making direct comparisons to detectors like DETR and YOLOv8 akin to apples-to-oranges.

Table 12. Comparative evaluation of DETR, YOLOv8, and ViT models on the test dataset using Precision, Recall, F1-Score, and IoU metrics

Model	Precision (%)	Recall (%)	F1-Score (%)	IoU (%)	AUC
DETR(SE-ResNet50)	90.00	86.10	86.00	86.70	0.928
YOLO8	83.90	83.50	83.70	81.03	0.9181
ViT	87.23	80.12	80.63	-	0.922

Figure 18 compares Precision, Recall, F1, and IoU on the held-out test set. DETR (SE-ResNet50) leads on all four metrics—90.0% precision, 86.10% recall, 86.00% F1, and 86.70% IoU— outperforming YOLOv8: (83.90%, 83.50%, 83.70%, 81.03%) and ViT (87.23%, 80.12%, 80.63%, —) by roughly 2–3 points across most metrics, while showing a much stronger advantage in Precision compared to YOLOv8.

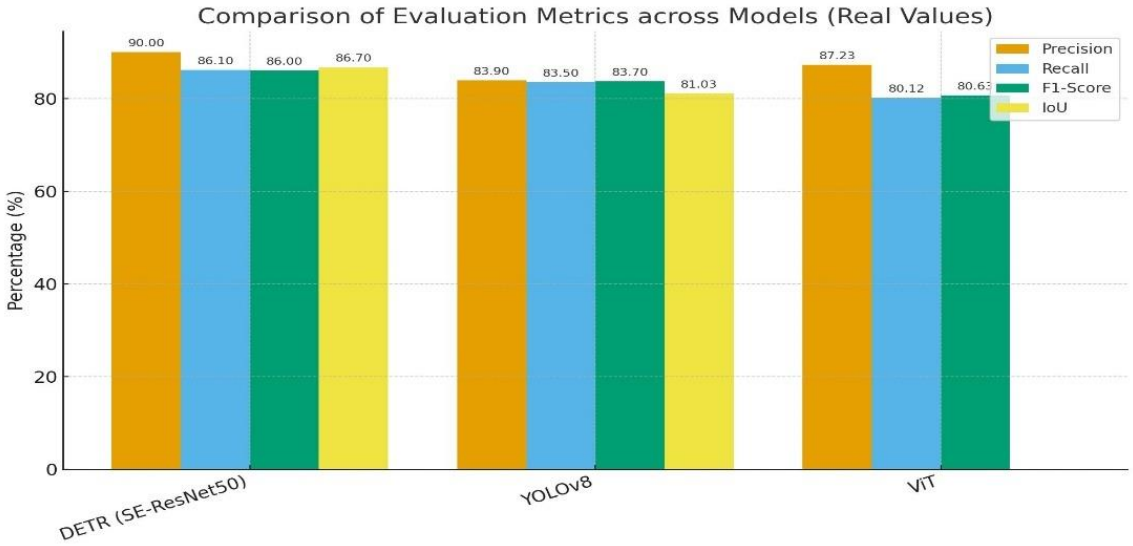


Figure 18. Comparison of Evaluation Metrics across Models.

To complement these point estimates, we also analyze ROC and Precision–Recall curves, which reveal performance across decision thresholds and are particularly informative under class imbalance.

Figure 19 shows the plots ROC curves for all three models, summarizing the sensitivity–specificity trade-off. DETR (SE-ResNet50) consistently dominates—especially at low false-positive rates—and achieves the highest

AUC = 0.928, indicating the strongest discriminative power. YOLOv8 and ViT follow with AUC = 0.9181,0.922 each, above the random-guess baseline (diagonal).

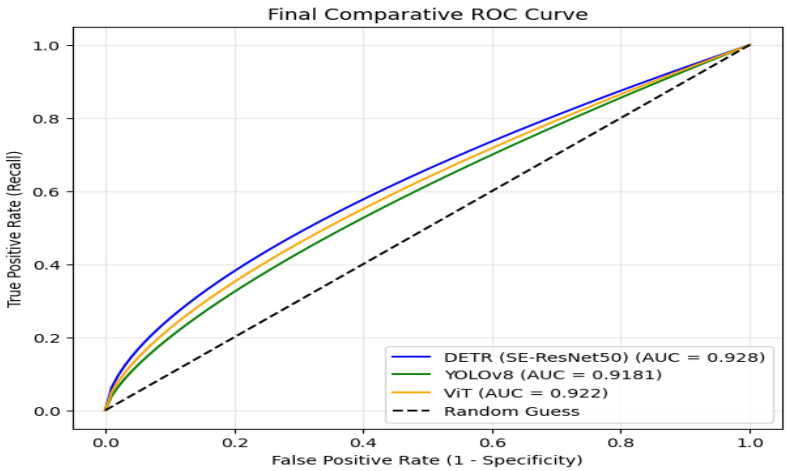


Figure 19. ROC Curve Comparison of DETR, YOLOv8, and ViT.

Figure 20 shows the Precision–Recall (PR) curves for all models. YOLOv8 achieves a mAP@0.5 of 0.7612, slightly below DETR (SE-ResNet50)'s 0.764. At stricter thresholds (mAP@0.75), YOLOv8 outperforms with 0.6470 compared to DETR's 0.506, highlighting YOLOv8's robustness at higher IoU levels. DETR remains more stable in recall-oriented regions, but the PR curves reveal trade-offs: YOLOv8 favors precision at stringent thresholds, while DETR balances overall discrimination. These differences underscore the need to select models based on specific clinical priorities, such as higher IoU for precise localization (favoring YOLOv8) versus broad recall (favoring DETR).

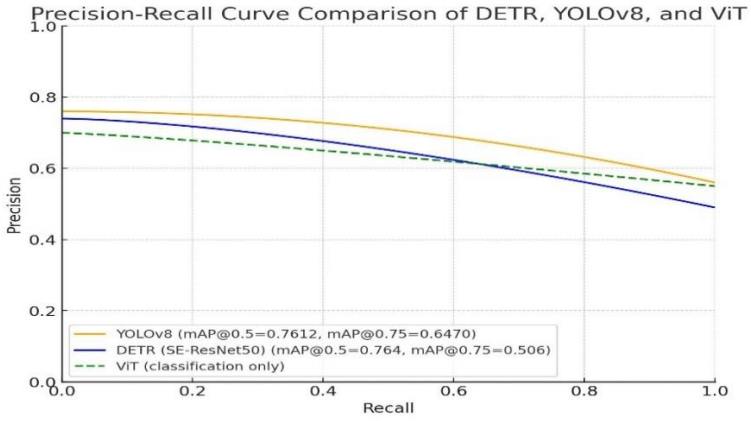


Figure 20. Precision-Recall Curve Comparison of DETR, YOLOv8, and ViT.

Consistent with the aggregate ROC and PR results in Figure 21, DETR (SE-ResNet50) remains the most stable and discriminative model. The results in Figure 21 highlight excellent discriminative ability for pituitary (AUC=0.990) and meningioma (AUC=0.951), but moderate performance for glioma (AUC=0.844), which is more challenging due to higher intra-class variability; future work could incorporate mitigation strategies such as weighted loss functions or oversampling to address this imbalance.

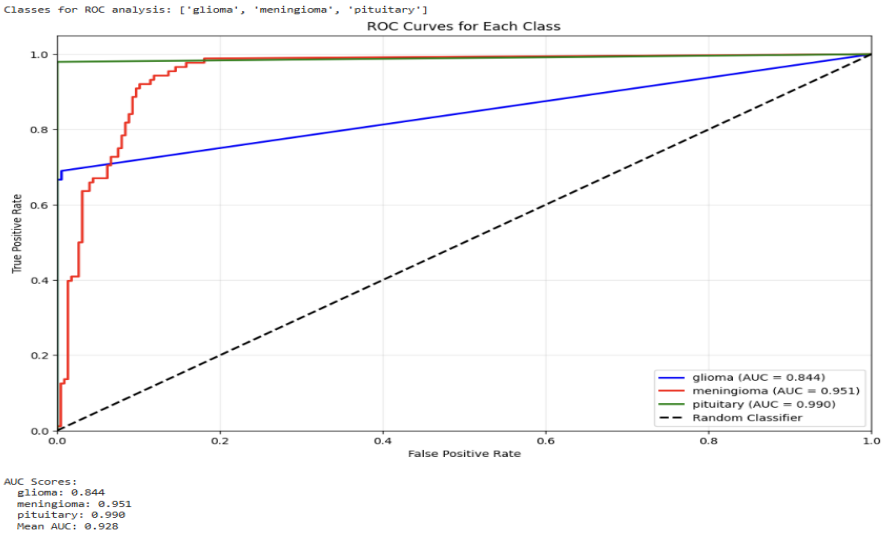


Figure 21. ROC analysis for each class: ['glioma', 'meningioma', 'pituitary'].

Class-wise results in Figure 21 show that DETR (SE-ResNet50) attains the highest accuracy for pituitary and meningioma classification, while the lower performance for glioma reflects the higher intra-class variability and morphological overlap observed in MRI. This pattern is consistent with the ROC and PR analyses, where the DETR variant exhibits the strongest discrimination for pituitary and meningioma, especially at high recall.

4.5 Conclusion

This chapter presented the complete experimental framework, covering dataset preprocessing, model architecture configurations, training protocols, and evaluation strategies. Among the evaluated architectures, DETR with an SE-ResNet50 backbone consistently outperformed the others across all major metrics. It achieved the highest precision (0.900), recall (0.861), F1-score (0.860), IoU (0.867), mAP@0.5 (0.764), and AUC (0.928), confirming its superior balance between sensitivity and specificity. YOLOv8 demonstrated competitive recall (0.835) and a comparable mAP@0.5 (0.7612), making it attractive for screening and real-time applications; however, its overall performance was weaker due to lower precision (83.90%) compared to DETR. ViT achieved solid

classification results (Precision = 0.872, Recall = 0.801, F1 = 0.806, AUC = 0.922) but lacked localization metrics such as IoU and mAP. In conclusion, DETR-SE provides the most reliable and balanced results for brain tumor detection in MRI, while YOLOv8 can be leveraged in scenarios prioritizing recall and speed.

Analyses of ROC and PR curves reinforced these findings. DETR demonstrated the most stable and class-balanced discrimination, with class-specific AUCs of 0.844 for glioma, 0.951 for meningioma, and 0.990 for pituitary tumors. YOLOv8 maintained strong performance at lower IoU thresholds (mAP@0.5 = 0.7612) but degraded more sharply at stricter thresholds (mAP@0.75 = 0.6470). ViT remained reliable in overall classification but, by design, did not provide localization capacity.

In summary, while DETR (SE-ResNet50) emerges as the best overall model in terms of balanced metrics (Precision, Recall, F1, IoU, and AUC), YOLOv8 offers practical advantages for real-time screening due to its recall and mAP@0.5 performance, and ViT provides a strong classification baseline.

Beyond numerical performance, explainability is a fundamental requirement for incorporating AI systems into radiology practice. Radiologists rarely rely on a model's predicted label alone; they require visual cues that clarify *why* the algorithm reached its decision. This is especially important in brain tumor diagnosis, where false negatives can have severe clinical consequences. Consistent with current XAI literature in medical imaging, interpretable outputs help bridge the gap between algorithmic predictions and clinical reasoning [55]. In this context, explainability supports trust, facilitates double-reading, and promotes safe human–AI collaboration.

In this study, the detection outputs of DETR-SE and YOLOv8 naturally provide a first layer of explainability by generating localized bounding boxes around suspicious regions. Such visual indicators help clinicians focus their attention on areas of potential pathology, reducing oversight in complex or noisy MRI scans. These spatial cues can act as a "second reader," flagging subtle abnormalities that might otherwise be overlooked in busy clinical workflows. Even though ViT does not generate localization maps, its transformer-based attention patterns still highlight discriminative regions of the brain, offering a complementary form of interpretability that can help clinicians validate whether the model is focusing on anatomically coherent areas.

From a radiologist's perspective, these explainable outputs have practical value: they can prioritize difficult cases, assist in early lesion detection, and improve diagnostic confidence without replacing professional judgment. Nonetheless, the present work is

limited by the absence of radiologist-in-the-loop evaluation. The bounding boxes and attention visualizations were assessed retrospectively, and their usefulness in real diagnostic workflows remains to be validated. Additionally, bounding-box–based explanations are inherently coarse compared to fine-grained saliency maps, and future studies should integrate more detailed post-hoc methods—such as Grad-CAM or attention-based heatmaps—within user-friendly clinical interfaces.

Future research should therefore include prospective reader studies, structured assessments of how explanations influence clinical trust, and the design of radiologist-centered interfaces that present model decisions and rationales in an actionable way. Such efforts are essential for advancing from algorithmic performance to true clinical integration.

Despite these promising outcomes, several limitations remain. The dataset size was modest and exhibited residual class imbalance, which may restrict generalizability. Heterogeneity in MRI acquisition across institutions further challenges cross-site robustness, while the complexity of transformer-based detectors raises the risk of overfitting. Moreover, differences in evaluation paradigms between detection (box-level) and classification (image-level) models complicate direct comparisons. Future studies should address these issues by employing larger, multi-center datasets, enhancing data harmonization, and integrating advanced model variants (e.g., deformable attention in DETR) to strengthen performance under stricter conditions [9, 17].

All source code, including scripts for training, evaluation, bootstrap confidence interval computation, ROC/PR plotting, and statistical testing, is publicly available via the GitHub repository referenced in this manuscript.

This work contributes to the field of AI in medical imaging in four significant ways. First, it demonstrates that high-accuracy deep learning models can enhance early diagnosis and enable more informed treatment planning in brain tumor care [1]. Second, it highlights the potential to reduce radiologists' workload and accelerate reporting without compromising reliability [56]. Third, by incorporating attention-based and post-hoc explanation methods, it improves interpretability and strengthens clinical decision-making [10]. Finally, it evaluates deployment-oriented factors—including generalization, robustness, and workflow integration—that help bridge the gap to real-world implementation of AI-assisted tumor detection in hospital environments [57].

Acknowledgment of Assistance

The idea, experimental design, and implementation of this research were entirely my own work. All model development, dataset preparation, training, and evaluation were conceived and executed by me. During the writing process and while debugging code, I used **ChatGPT** as a supporting tool for: (i) rephrasing and improving the clarity of my written text in English, and (ii) assisting in identifying and correcting coding errors. No part of the scientific contributions, results, or analyses originated from ChatGPT; it only served as a language and debugging assistant. The responsibility for the concepts, methodology, results, and final interpretations remains fully mine.

Reference:

1. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Medical image analysis. 2017;42:60–88.
2. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H. Transformers in medical imaging: A survey. Medical image analysis. 2023;88:102802.
3. Wang C-Y, Bochkovskiy A, Liao H-YM, editors. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2023.
4. Chen C, Zhao L-L, Lang Q, Xu Y. A Novel Detection and Classification Framework for Diagnosing of Cerebral Microbleeds Using Transformer and Language. Bioengineering. 2024;11(10):993.
5. Hu J, Shen L, Sun G, editors. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2018.
6. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S, editors. End-to-end object detection with transformers. European conference on computer vision; 2020: Springer.
7. Isensee F, Jäger PF, Kohl SA, Petersen J, Maier-Hein KH. Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:190408128. 2019.
8. Liu W, Guo X. RRFNet: A free-anchor brain tumor detection and classification network based on reparameterization technology. PLoS One. 2025;20(6):e0325483.
9. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging. 2014;34(10):1993–2024.
10. Tjoa E, Guan C. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems. 2020;32(11):4793–813.
11. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, Hawkins C, Ng H, Pfister SM, Reifenberger G. The 2021 WHO classification of tumors of the central nervous system: a summary. Neuro-oncology. 2021;23(8):1231–51.
12. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
13. Bernal J, Kushibar K, Asfaw DS, Valverde S, Oliver A, Martí R, Lladó X. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. Artificial intelligence in medicine. 2019;95:64–81.
14. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020.
15. Hussain M. Yolov5, yolov8 and yolov10: The go-to detectors for real-time vision. arXiv preprint arXiv:240702988. 2024.
16. Khalili B, Smyth AW. Sod-yolov8—enhancing yolov8 for small object detection in aerial imagery and traffic scenes. Sensors. 2024;24(19):6209.
17. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:201004159. 2020.
18. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, Pal C, Jodoin P-M, Larochelle H. Brain tumor segmentation with deep neural networks. Medical image analysis. 2017;35:18–31.
19. Zahoor MM, Qureshi SA, Bibi S, Khan SH, Khan A, Ghafoor U, Bhutta MR. A new deep hybrid boosted and ensemble learning-based brain tumor analysis using MRI. Sensors. 2022;22(7):2726.

20. Balaji P, Sri Revathi B, Gobinathan P, Shamsudheen S, Vaiyapuri T. Optimal IoT Based Improved Deep Learning Model for Medical Image Classification. Computers, Materials & Continua. 2022;73(2).
21. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL, editors. Microsoft coco: Common objects in context. European conference on computer vision; 2014: Springer.
22. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B, editors. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision; 2021.
23. Campos S, Davey P, Hird A, Pressnail B, Bilbao J, Aviv R, Symons S, Pirouzmand F, Sinclair E, Culleton S. Brain metastasis from an unknown primary, or primary brain tumour? A diagnostic dilemma. Current Oncology. 2009;16(1):62.
24. Barresi V, Tuccari G, Barresi G. NGAL immunohistochemical expression in brain primary and metastatic tumors. Clin Neuropathol. 2010;29(5):317–22.
25. Villano J, Vokes E. Primary and metastatic brain tumors. Oncologic Therapies: Springer; 2003. p. 569–85.
26. Gavrilovic IT, Posner JB. Brain metastases: epidemiology and pathophysiology. Journal of neuro-oncology. 2005;75(1):5–14.
27. Pope WB, Sayre J, Perlina A, Villablanca JP, Mischel PS, Cloughesy TF. MR imaging correlates of survival in patients with high-grade gliomas. American Journal of Neuroradiology. 2005;26(10):2466–74.
28. Hong N, Du X, Nie Z, Li S. Diffusion-weighted MR study of femoral head avascular necrosis in severe acute respiratory syndrome patients. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine. 2005;22(5):661–4; Bammer R. Basic principles of diffusion-weighted imaging. European journal of radiology. 2003;45(3):169–84.
29. Cha S. Update on brain tumor imaging: from anatomy to physiology. American Journal of Neuroradiology. 2006;27(3):475–87.
30. Katti G, Ara SA, Shireen A. Magnetic resonance imaging (MRI)–A review. International journal of dental clinics. 2011;3(1):65–70.
31. Zhang J, Li Y, Zhao Y, Qiao J. CT and MRI of superficial solid tumors. Quantitative imaging in medicine and surgery. 2018;8(2):232.
32. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A, editors. Learning deep features for discriminative localization. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
33. Shah S, Tembhurne J. Object detection using convolutional neural networks and transformer-based models: a review. Journal of Electrical Systems and Information Technology. 2023;10(1):54.
34. Ahmed M, El-Sheimy N, Leung H. A novel detection transformer framework for ship detection in synthetic aperture radar imagery using advanced feature fusion and polarimetric techniques. Remote Sensing. 2024;16(20):3877; Ronneberger O, Fischer P, Brox T, editors. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention; 2015: Springer.
35. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y. A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence. 2022;45(1):87–110.
36. Tummala S, Kadry S, Bukhari SAC, Rauf HT. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. Current Oncology. 2022;29(10):7498–511.
37. Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. Do vision transformers see like convolutional neural networks? Advances in neural information processing systems. 2021;34:12116–28.

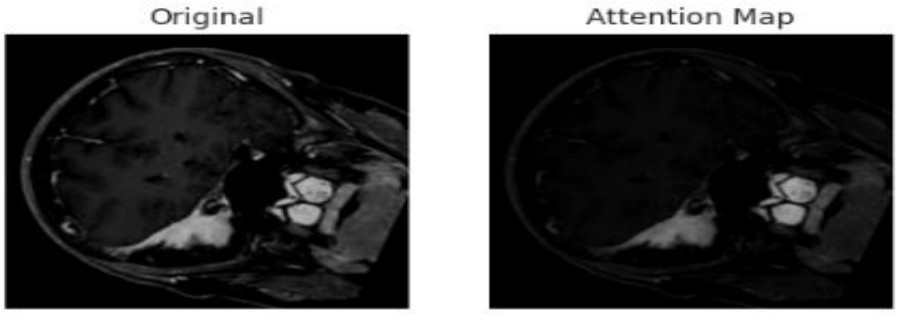
38. Patel S, Kadam Y, Thombare A, Salvi K, Jadhav D. 'A review of brain tumor detection techniques using YOLOv8. Int J Res Appl Sci Eng Technol. 2024;12(3):1075–8.
39. Palanivel N, Deivanai S, Sindhuja B, editors. The art of YOLOv8 algorithm in cancer diagnosis using medical imaging. 2023 International Conference on System, Computation, Automation and Networking (ICSCAN); 2023: IEEE.
40. Ali ML, Zhang Z. The YOLO framework: A comprehensive review of evolution, applications, and benchmarks in object detection. Computers. 2024;13(12):336.
41. Saraei M, Lalinia M, Lee E-J. Deep Learning-Based Medical Object Detection: A Survey. IEEE Access. 2025.
42. Srinivasu PN, Kumari GLA, Narahari SC, Ahmed S, Alhumam A. Exploring the impact of hyperparameter and data augmentation in YOLO V10 for accurate bone fracture detection from X-ray images. Scientific Reports. 2025;15(1):9828.
43. Ahamed MF, Islam MR, Nahiduzzaman M, Karim MJ, Ayari MA, Khandakar A. Automated detection of colorectal polyp utilizing deep learning methods with explainable AI. IEEE Access. 2024;12:78074–100.
44. Hasan MN, Ishraq A, Alam Emon A, Shin J, Kabir MM. Advancing Breast Cancer Diagnosis: Attention-Enhanced U-Net for Breast Cancer Segmentation. Data-Driven Clinical Decision-Making Using Deep Learning in Imaging: Springer; 2024. p. 207–26.
45. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S, editors. Generalized intersection over union: A metric and a loss for bounding box regression. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019.
46. Zou ZN, Zhang Y, Wijaya R. Investigating the Robustness and Properties of Detection Transformers (DETR) Toward Difficult Images. arXiv preprint arXiv:231008772. 2023.
47. Ioffe S, Szegedy C, editors. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International conference on machine learning; 2015: pmlr.
48. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data. 2019;6(1):60.
49. Kingma DP. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.
50. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods. 2021;18(2):203–11.
51. Sundar GN, Narmadha D, Jerry NA, Thangavel SK, Shanmugam SK, Ajibesin AA, editors. Brain Tumor Detection and Classification using Vision Transformer (ViT). 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS); 2024: IEEE.
52. Ren S, Song J, Yu L, Tian S, editors. DHC-YOLO: An Improved YOLOv8 for Lesion Detection in Medical Images. 2024 2nd International Conference on Machine Vision, Image Processing & Imaging Technology (MVIPIT); 2024: IEEE.
53. Missaoui R, Hechkel W, Saadaoui W, Helali A, Leo M. Advanced Deep Learning and Machine Learning Techniques for MRI Brain Tumor Analysis: A Review. Sensors. 2025;25(9):2746.
54. Hendriks J, Mutsaerts H-J, Joules R, Peña-Nogales Ó, Rodrigues PR, Wolz R, Burchell GL, Barkhof F, Schrantee A. A systematic review of (semi-) automatic quality control of T1-weighted MRI scans. Neuroradiology. 2024;66(1):31–42.
55. Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, Nensa F. Explainable AI in medical imaging: An overview for clinical practitioners–Beyond saliency-based XAI approaches. European journal of radiology. 2023;162:110786.
56. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nature Reviews Cancer. 2018;18(8):500–10.
57. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence. 2020;2(6):305–11.

Appendix:

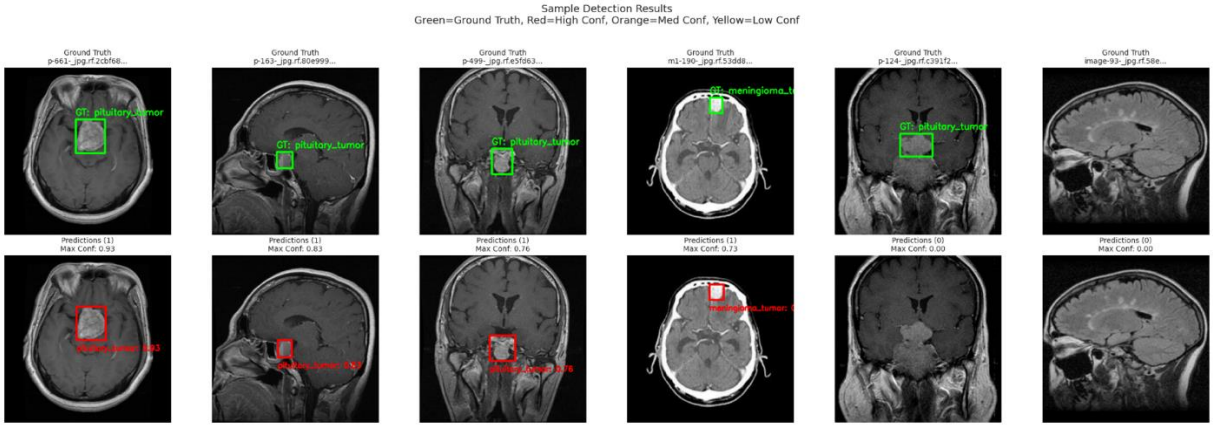
Visual Outputs

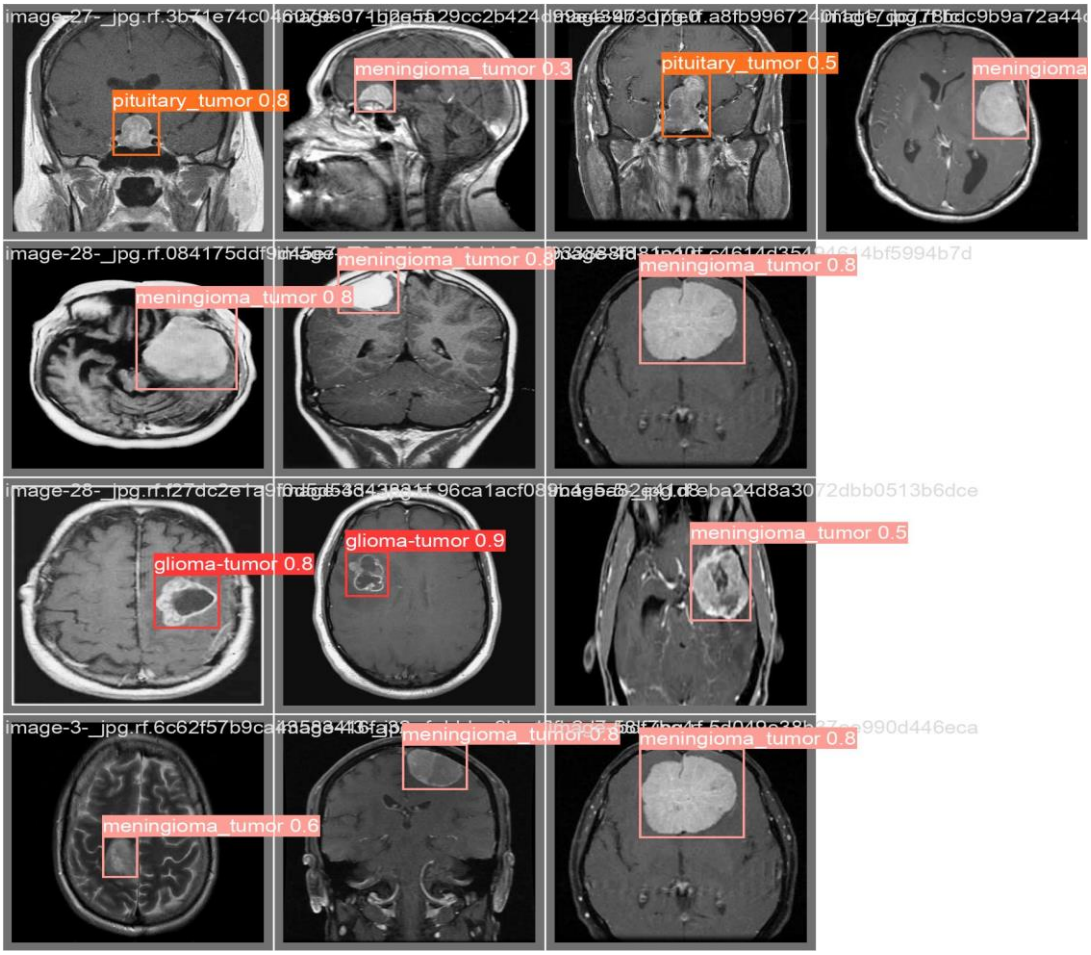
This section includes representative visual outputs from ViT, YOLOv8, and DETR (SE_resnet 50), capturing tumor results.

ViT:

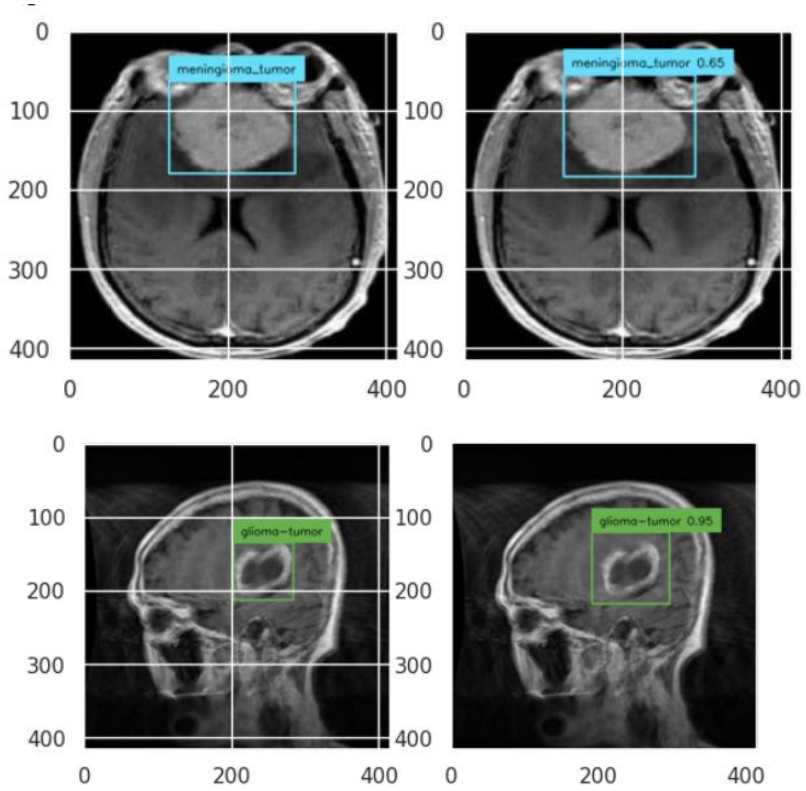


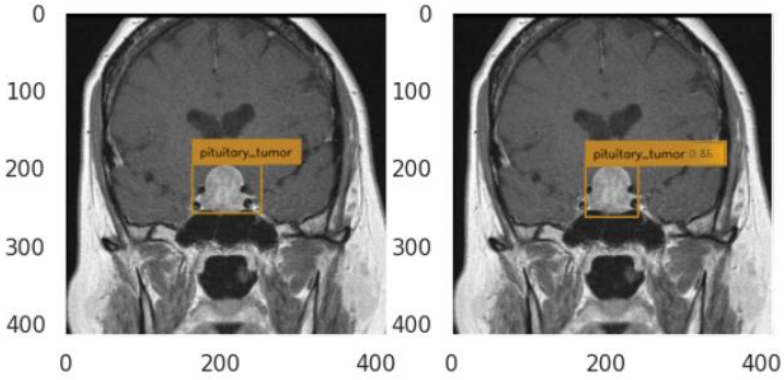
Yolov8:




pituitary_tumor 0.8
meningioma_tumor 0.3
pituitary_tumor 0.5
meningioma
meningioma_tumor 0.8
meningioma_tumor 0.8
meningioma_tumor 0.8
glioma-tumor 0.8
glioma-tumor 0.9
meningioma_tumor 0.5
meningioma_tumor 0.8
meningioma_tumor 0.8
meningioma_tumor 0.6

DETR:


meningioma_tumor
meningioma_tumor 0.65
glioma-tumor
glioma-tumor 0.95
0
100
200
300
400


0
100
200
300
400
0
200
400
pituitary_tumor
0
100
200
300
400
0
200
400
pituitary_tumor 0.86

DETR (SE_Resnet 50)