National College of Ireland

Comparative Analysis of Attention – Based and Convolution YOLO Architecture for Small Object Detection in Medical Imaging

MSc Research Project
MSc in Data Analytics

Rahul Koguri
Student ID: x23423561

School of Computing
National College of Ireland

Supervisor: Dr. Abdul Qayum

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Rahul Koguri		
Student ID:	x23423561		
Programme:	MSc in Data Analytics	**Year:**	2025-2026
Module:	MSc Research Practicum Part 2		
Supervisor:	Dr Abdul Qayum		
Submission Due Date:	28-01-2026		
Project Title:	Comparative Analysis of Attention – Based and Convolution YOLO Architectures for Small Object Detection in Medical Imaging		
Word Count:	8340	**Page Count**	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Rahul Koguri

Date: 28-01-2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comparative Analysis of Attention – Based and Convolution YOLO Architecture for Small Object Detection in Medical Imaging

Rahul Koguri
x23423561

Abstract

The detection of small objects is a long-standing issue in the medical domain, since pathological areas tend to be less than one percent of the image size. Though methods in the YOLO family have transformed from convolutional networks to attention-based networks, there has been no comparison regarding their efficiency in the detection of small medical abnormalities. This research presents the first comprehensive analysis comparing attention-based YOLO variants (YOLOv11m with C2PSA, YOLOv12m with Area Attention) against CNN-based models (YOLOv8m, YOLOv10m) in four different medical imaging techniques: brain tumor MRI, liver histopathology, chest radiography, and bone fracture detection.

The work assesses sixteen model and dataset pairs based on standardized training procedures and a set of evaluation metrics such as mean average precision, the trade-off between precision and recall, and inference speed. The results show that architectural supremacy is context-dependent and thus not generic across datasets and tasks. CNN architectures outperform in well-defined pathologies and when the number of training samples is low, and the attention-based architecture has an advantage in identifying diffuse boundaries typical of histopathological images. The efficiency-optimized architectures appear to be enough to handle large pathological features.

The results call into question the assumptions surrounding the superior function of the attention mechanism, instead demonstrating that the characteristics of the dataset, such as the size distribution and definition of the object boundaries, and the amount of available training data, are more informative of the choice than the assumed complexity of the model. The contribution offers evidence-based recommendations for practitioners regarding the choice of architectures in detection tasks for medical imaging applications and the significance of matching architectural capabilities to corresponding medical imaging needs.

Keywords: MRI,CNN,MRI,YOLO,Attention

1 Introduction

1.1 Background and Context

Object detection is a basic application of computer vision and has been a focus area in the last years for the detection and localization of objects in images and videos. Among the numerous methods developed in the last decade, the uniqueness of the YOLO series has been its single-

shot approach to object detection while offering an unparalleled trade-off between the processing speed and the accuracy (Terven et al., 2023). Since the inception of the first model in the YOLO series in the year 2015, there have been twelve architectural updates, each recommending innovative improvements that marked the boundary of real-time object detection. This marks a complete shift in paradigmas from the convolutional neural network-based models to the attention models, thereby defining a new way of processing and analyzing visual data.

The attention mechanism embodied in an object detection architecture resolves the long-standing issues associated with the traditional solutions. The traditional CNN-based object detection methods rely upon the local receptive fields to extract the feature descriptions in the predefined spatial locations, overlooking the global context captured by the distant image areas. Attention mechanisms, a concept first applied to natural language processing tasks, can emphasize certain areas in an image to extract information that a traditional model cannot capture (Khan et al., 2022). This capabilility is an important feature for object detection because the object has to be detected in relation to the context of the image.

The current versions of the YOLO model are examples of this paradigm shift in action. The introduction of the Cross-Stage Partial with Parallel Spatial Attention (C2PSA) module in the YOLOv11 model added spatial attention to the basic convolutional module structure. YOLOv12 advances further with Area Attention and Residual Efficient Layer Aggregation Networks (R-ELAN), focusing attention as the feature extraction mechanism instead of an enhancement layer (Tian et al., 2025). This development brings up a number of important issues regarding application in the area of specific tasks, especially the detection of small objects where the application of contextual understanding has the greatest value.

Small object detection has proved to be a long-standing challenge in computer vision, and its performance degrades drastically when objects occupy fewer pixels of the image (Diwan et al., 2022). In medical imaging, the pathological abnormalities represent less than one percent of the entire area in an image. The correct identification of tumors in their various stages, microcalcifications in mammography, and cellular abnormalities notably affects the diagnostic processes and results in medical practices.

1.2 Problem Statement and Motivation

The issues in medical imaging object detection add to the complexities involved in micro-object detection problems. Tumors demonstrate a subtle appearance in the image, low contrast to the tissue, and an ill-defined margin, making it difficult even for the best human observers to identify. Current literature reports a significant loss in detection performance for lesions less than 10 mm in diameter, reducing the detection rate from a possible 95% in large cancers to less than 60% in micro-lesions (Maurya et al., 2022). This discrepancy is a crucial limitation because the abnormalities in the early stages of development, and thus those that are most accessible to a positive intervention outcome, will necessarily be the smallest and most challenging.

Despite considerable work done in the area of YOLO architectures and medical imaging applications, there exist large gaps in the understanding of the implications of the latest architectural advancements in the area of small medical object detection. Current benchmarks test YOLO on general datasets (COCO, PASCAL VOC) where object size distributions differ greatly from medical imaging. Previous comparison studies conducted for CNN-based and

attention-based models did not test the detection rate for objects that occupy less than one percent of the total area in the images. Lack of specific testing in this area makes it difficult for practitioners to make informed decisions based upon available evidence concerning the choice of architecture in medical applications.

The clinical applications of superior detection of small objects relate to various domains. For instance, in chest radiography, the detection of early signs of a lung nodule translates to a better prognosis and survival rate in cancer patients. Another application is mammography, where the microcalcifications representative of early stages of breast cancer need to be detected. Histopathological analyses involve the identification of cells in tissue sections where each cell can occupy a fraction of one percent of the area to be imaged. Each application requires superior systems for the discrimination of small objects, while the optimal design has not been adequately explained.

1.3 Research Question and Objectives

The above discussion regarding the current state-of-the-art object detection models and their application to medical imaging begets the following central research question:

How do attention-based YOLO architectures (YOLOv11 and YOLOv12) quantitatively compare to CNN-based models (YOLOv8 and YOLOv10) in detecting small medical abnormalities, as measured by mean Average Precision, precision, recall, and inference time?

Research Objectives:

1. **To implement and evaluate** four YOLO models—medium variants of versions eight, ten, eleven, and twelve—on four different medical imaging datasets corresponding to different imaging modalities and pathologies.
2. **To quantitatively analyze** performance differences between CNN-based and attention-based YOLO models using standardized metrics (mAP@0.5, mAP@0.5-95, precision, recall, inference time) under controlled experimental conditions.
3. **To identify and characterize** medical imaging task qualities that affect the performance of the model differently include object size distributions, the clarity of boundary definitions, and the amount of data available for training.
4. **To verify empirically** whether attention-based models outperform the optimized convolutional models in the specific task of small object detection tasks across diverse medical contexts.
5. **To establish practical guidelines** for model selection based on the characteristics of the dataset and computational and clinical constraints, and enabling the provision of evidence-based recommendations to practitioners.

Success will be measured by complete training and evaluation of all sixteen model-dataset combinations; statistically significant performance differences enabling clear architectural comparisons; identification of distinguishable patterns relating dataset characteristics to model performance; formulation of actionable deployment guidelines.

1.4 Report Structure

This report is structured as follows: Section 2 presents a critical review of the current literature available on the topic of YOLO variants, the issue of object detection in smaller objects, and their applications in the medical image domain. Section 3 describes the methodology of the study and contains the choice of datasets, preprocessing steps, design and implementation details, and the metrics to be used for the evaluation process. Section 4 carries the description of design specification and its comparison methodology. Section 5 presents the implementation phase of the work. Section 6 shows the results of the evaluation process performed for each combination. Section 7 offers a detailed discussion of the work in the light of the current literature available in the domain. Section 8 concludes with contributions and future work.

2. Related Work

2.1 Evolution of YOLO Architectures

The YOLO series has a considerable architectural evolution ranging from simple convolutional networks to advanced attention-based models. Terven et al. (2023) offer in-depth analysis of the evolution of YOLOv1 to YOLOv8 regarding the enhancement of the network backbone, feature pyramid, and training methods. The study clearly shows improvements in performance through the application of CNN architectures with greater depth and improved feature fusion, registering an increase in the value of mAP to 41.9% for large objects. Despite this, the task of small object detection is still a challenge with a low value of mAP at 21.8% for objects with a size less than 32×32 pixels. The work of their analysis is prior to the emergence of attention-based models named YOLOv11 with C2PSA and YOLOv12 with Area Attention, leaving a gap in understanding attention mechanism benefits for small object detection.

Xu et al. (2024) explore the basic tenets driving each iteration of YOLO, pointing out the inefficiencies in the detection of small and occluded objects despite a series of improvements. They highlight its uninterpretability and inefficiency as weaknesses against its application in resource-scarce environments. The current survey will emphasize efficiency but will disregard the latest models based on attention mechanisms (YOLOv11, YOLOv12) and the application in medical imaging, ignoring important domain-specific knowledge for its application in a clinical setup.

YOLOv10 brings revolutionary efficiency improvements without compromising accuracy. Jegham et al. (2024) document NMS-free training through dual-assignment and Compact Inverted Blocks (CIB) utilizing depthwise convolution within Efficient Layer Aggregation Networks (ELAN). The first incorporation of an attention mechanism is the Partial Self-Attention (PSA) integration in terms of improving the accuracy with nominal computational costs. The performance assessment of their work focuses on the traditional benchmarks and unrelated to the medical image applications involving the smallest pixel proportions in object detection, limiting insights into clinical applicability where small object detection is paramount.

Alif and Hussain (2024) assess the efficiency of various variants of the YOLO algorithm in agricultural applications, proving the state-of-the-art efficiency of the YOLOv10 algorithm but pointing out the effectiveness depending upon the target and the conditions of the image. The results in the agricultural sector will not be applicable in the medical applications directly due

to the differences in the objectives, contrast ratios, and clinical requirements differ substantially from crop monitoring applications.

Attention-based architectures represent YOLO's next generation. Khanam and Hussain (2025) detail YOLOv11 and YOLOv12 improvements, documenting YOLOv11's C3k2 block replacing YOLOv8's C2f module for optimized processing using dual small convolutions instead of single large operations. The Cross-Stage Partial with Parallel Spatial Attention (C2PSA) module enables dynamic focus on critical image regions, potentially improving small occluded object detectionThe model named YOLOv11 has a reduced number of parameters by 22% against the YOLOv8m model and additionally has enhanced mAP results.The assessment conducted by the authors involves common benchmarks for object detection but doesn't include the evaluation in the medical domain.

Tian et al. (2025) present YOLOv12's paradigm shift toward attention-centric frameworks. Area Attention divides feature maps for optimized processing compared to self-attention, while Residual Efficient Layer Aggregation Networks (R-ELAN) promote block-level residual connections addressing gradient instability. Compared to the YOLOv10 model, the YOLOv12 model shows a gain of 2% in terms of mAP but has the drawback of higher memory footprint and lower CPU speeds similar to CNNs, while there is no evaluation regarding the application of the model to the detection of smaller medical objects where the advantage of attention mechanisms could be more clear-cut.

2.2 Small Object Detection Challenges

Small Object Detection is a persistent challenge in the area of Computer Vision. Diwan et al. (2022) to point out this is the ultimate weakness of the YOLO model because of the poor representation of features in the last layers of the network where the presence of smaller objects is lost due to the pooling layers. The detection of objects is reduced by a factor of 50% for any object that is less than 32x32 pixels in size. They highlight the weakness of the traditional CNN model but don't consider how the attention mechanism can be a solution to the basic issues in feature representation across network depth.

Pérez-Hernández et al. (2020) implement ODeBiC based on binary classification to remove FPs, resulting in an improvement of 19.57% in the accuracy of small objects by post-processing instead of architectural modifications. The methods applied through post-processing include additional computational costs and the results might not be generalized to various object characteristics in a variety of medical image modalities, and thus architectural modifications could be more robust.

Zhang et al. (2025) to prove the superiority of transformer architectures over the CNN approaches in the capability to maintain multi-scale feature information. For their cervical cytology image analysis tasks where, pathological cells occupy less than 0.5% of the picture area, they report the average precision of the attention-based Faster R-CNN to be 67.4%. The two-stage model detection process is a radically different model design than the one-stage YOLO model, making a comparison impossible and leaving the question of whether any similar advantages exist in the YOLO model structure unanswered.

He et al. (2025) assesses the RT-DETR transformer architecture in retinopathy detection, reporting a mAP of 97.7% to be contrasted with the 86.0% of the state-of-the-art YOLOv8 in the small cluttered environment, with the relevant competition excluded in their results: those

involving YOLOv11 and YOLOv12 that employ the attention mechanism in the YOLO structure, leaving unknown whether similar benefits exist in these latest single-stage methods.

Computational efficiency complicates architectural decisions. Albuquerque et al. (2024) assesses 311 medical imaging papers and concludes a transformer model is more accurate for smaller objects but costs 2 to 3 times more in computational cost than CNNs. The assessment is before the latest versions of the YOLO model and doesn't consider the current optimizations in attention-based models that might improve the current trade-off between efficiency and accuracy.

Wu et al. (2020) demonstrate how feature pyramid networks and multi-scale training compensate for the problems of CNNs in identifying small objects without additional computational costs. The work is outdated in terms of current attention methods' effectiveness in performing well with less computational costs, especially in the context of efficient architectures such as YOLOv11 and YOLOv12 designed for efficiency.

2.3 Medical Imaging Applications

Medical image object detection brings its own domain-specific problems, adding to the complexities of small object detection. Maurya et al. (2022) thorough analysis of cancer detection across various anatomical locations, pointing out the basic limitation in the detection of small lesions. Overall performance rapidly worsens for cancers smaller than 10mm in diameter, reducing from 95% in large cancers to less than 60% in micro-cancers. The assessment will be confined to basic deep learning architecture and will exclude architectural work and the potential benefits of attention mechanisms offer advantages.

Satushe et al. (2025) systematically review 155 brain tumor diagnosis studies (2010-2025), to demonstrate the effectiveness of hybrid transformer models in outperforming the traditional CNN architecture by a margin of 15% in the detection of small tumors while utilizing the attention modules. The work should focus solely on brain tumors and thus lack applicability to other medical image sources based on the variations in the object to be detected, for instance, histopathology and chest radiography.

Çallı et al. (2021) survey results of chest X-ray analysis, recording the attainment of only 45% mAP in small nodules but 78% in large consolidation by YOLO versions. The ensemble methods are known to perform better than the architectures taken separately. The assessment doesn't consider the latest versions of the models, namely YOLOv11 and YOLOv12, and the benefits of the attention mechanism benefits for small pathology detection.

Saraei et al. (2025) report a YOLOv8 with a resulting 71.6% mAP score on the small category of medical objects, an improvement upon its predecessors. YOLOv11 and YOLOv12 are excluded from the results, while the contribution of architectural components to the results has not been analyzed to understand what ideas are responsible for the advancement.

Iqbal et al. (2023) stressing the importance of preprocessing pipelines and augmentation techniques has a large effect on the performance of detection, and proper augmentation helps to partially overcome the issue with the size of the training data. The issue of the choice of the best model remains unaddressed when comparing preprocessing methods alone and architectures are not taken into account.

2.4 Research Gap and Summary

The comprehensive review of existing literature reveals significant gaps across three critical dimensions. In YOLO architectural evolution, studies by Terven et al. (2023) comprehensively analyze YOLOv1 through YOLOv8, documenting CNN-based improvements but achieving only 21.8% mAP for small objects, yet their analysis excludes attention-based models YOLOv11 and YOLOv12. Similarly, Xu et al. (2024) investigate YOLO efficiency and identify computational barriers but miss medical contexts and latest versions. While Khanam and Hussain (2025) document YOLOv11/v12 architectures achieving 22% parameter reduction through C2PSA, they provide no medical imaging evaluation. Tian et al. (2025) demonstrate YOLOv12's Area Attention achieving 2% improvement over YOLOv10, but without small object specific testing critical for medical applications.

In small object detection research, significant limitations persist across methodologies. Zhang et al. (2025) achieve 67.4% average precision for tiny pathological cells using medical transformers, but employ two-stage detectors rather than YOLO framework. He et al. (2025) reporting the RT-DETR model to achieve a 97.7% mAP score against the YOLOv8 model's 86% in the medical application scenario, but without the comparison to YOLOv11 and YOLOv12, leaving unknown whether attention-enhanced YOLO variants offer similar benefits. Albuquerque et al. (2024) review 311 papers published in the medical imaging literature and find that transformers cost 2-3x more computationally than CNNs, but the results are before the latest versions of YOLO with optimized attention mechanisms.

The literature relating to applications of medical imaging also suffers from a lack of thorough architectural comparisons. Maurya et al. (2022) document performance decrease from 95% to 60% in tumors less than 10mm but do not provide architectural comparison. Satushe et al. (2025) show the effectiveness of transformers with a 15% improvement in the case of brain tumors, but the domain specificity to the brain prevents generalized results. Çallı et al. (2021) report that the best results for YOLO are only 45% mAP for small nodules and 78% for large objects in the case of chest X-rays, but are predated by attention-enhanced versions of YOLOs. All the cited studies taken together point out the lack of thorough comparison of attention-based and CNN-based variants of YOLO in the area of small object detection.

Table 1: Literature Review Summary

Study	**Focus Area**	**Key Findings**	**Critical Gap**
Terven et al. 2023	YOLO evolution (v1-v8)	CNN improvements; 21.8% mAP small objects	No attention models (v11, v12) analysis
Xu et al. 2024	YOLO efficiency	Computational barriers identified	Missing medical context and latest versions
Khanam & Hussain 2025	YOLOv11/v12 architecture	22% parameter reduction with C2PSA	No medical imaging evaluation
Tian et al. 2025	YOLOv12 Area Attention	2% improvement over v10	No small object specific testing
Zhang et al. 2025	Medical transformers	67.4% AP for tiny cells	Two-stage detector, not YOLO
He et al. 2025	RT-DETR medical	97.7% mAP vs YOLOv8 86%	Excludes v11, v12 comparison

Maurya et al. 2022	Cancer detection	95%→60% degradation <10mm	No architectural comparison
Satushe et al. 2025	Brain tumor diagnosis	Transformers +15% improvement	Brain-specific, limited scope
Çallı et al. 2021	Chest X-ray	45% small vs 78% large objects	Predates attention YOLOs
Albuquerque et al. 2024	Medical detection review	Transformers 2-3x computational cost	Missing latest YOLO versions

The literature reveals a critical gap: no systematic comparison exists between attention-based YOLO architectures (YOLOv11, YOLOv12) and CNN-based predecessors (YOLOv8, YOLOv10) for small medical object detection. Though the benefits of attention mechanisms in theory in related works are proved, empirical proof in the context of the YOLO framework has been lacking in the medical context and will be filled by the current study through our benchmarks in the medical domain, providing evidence-based guidance for architecture selection in clinical deployment scenarios.

3 Research Methodology

This section introduces the framework to compare the attention-based model and the CNN-based model for the YOLO algorithm in the detection of small medical objects, encompassing experimental setup, dataset preparation, model configuration, and evaluation procedures.

3.1 Research Design

The experimental comparative method assesses four YOLO models against four medical image datasets, resulting in a total of sixteen model and dataset pairs for thorough comparison. The controlled-variable comparative method promotes equal training conditions and focuses the comparison solely on model performance variation. Medium-sized variants (denoted by 'm' suffix) balance computational efficiency with detection accuracy, enabling fair comparison between CNN-based models (YOLOv8m, YOLOv10m) and attention-enhanced variants (YOLOv11m with C2PSA, YOLOv12m with Area Attention).

3.2 Dataset Description

Four publicly available datasets of medical images from Roboflow Universe include a variety of types of pathology, modes of imaging, and distributions of object sizes. The properties of the datasets are listed in Table 1.

Table 1: Dataset Characteristics

Dataset	Domain	Images	Classes	Train/Val/Test	Avg Objects/Image	Object Size	Source
Brain Tumor MRI	Neuroimaging	5,833	4	5,085/502/246	1.01	Medium-Large	https://universe.roboflow.com/thesiswork-o8h4e/labeled-mri-brain-tumor-dataset-yocrm/dataset/1

Liver Pathology	Histopathology	9,512	4	8,322/790/400	2.47	Small	https://universe.roboflow.com/thesiswork-o8h4e/liver-disease-agmvl-kjyir/dataset/2
Chest X-ray	Radiography	10,064	5	8,805/832/427	1.08	Large	https://universe.roboflow.com/thesiswork-o8h4e/chest-x-ray-images-put19-lxxby/dataset/1
Bone Fracture	Radiography	1,110	4	978/88/44	1.33	Small-Medium	https://universe.roboflow.com/thesiswork-o8h4e/bone-fracture-7fylg-jpa10/dataset/1

Brain Tumor MRI contains well-defined tumors (Glioma, Meningioma, Pituitary, No Tumor) with Single Object Detection (SOD) characteristics meaning each image typically contains one annotated object, simplifying detection compared to multi-object scenarios. Liver Pathology presents challenging small, diffuse pathological regions with higher annotation density (2.47 objects/image) reflecting multi-focal disease patterns. The Chest X-ray supplies large pathological regions for the baseline evaluation of the attention mechanism. Bone Fracture, with limited training data (1,110 images) and small fracture lines, tests model robustness under data scarcity.

3.3 Data Preprocessing and Augmentation

The standardized preprocessing guarantees the same input conditions for all datasets. The same pipeline was applied to each of the four datasets to remove any confounding effects due to preprocessing.

Table 2: Preprocessing Pipeline (Applied to All Datasets)

Step	Specification
Auto-orientation	EXIF metadata removal
Spatial Resizing	640×640 pixels (stretch)
Annotation Format	YOLO normalized bounding boxes
Augmentation	3× training set expansion

Data augmentation generates three variants per source image through geometric transformations (horizontal/vertical flips at 50% probability, rotation ±15°, shear ±10°) and photometric adjustments (brightness ±15%, exposure ±10%, Gaussian blur 0-2.5 pixels). This augmentation technique, applied to each of the datasets in a uniform manner, will solve the issue of limited medical training data and will simulate the variability that is present in real-world images.

3.4 Model Architecture Summary

Table 3: YOLO Architecture Specifications

Model	Key Components	Attention Type	Parameters	GFLOPs
YOLOv8m	C2f	None (CNN)	25.84M	78.7
YOLOv10m	CIB + PSA	Partial Self-Attention	15.32M	58.9
YOLOv11m	C3k2 + C2PSA	Spatial Attention	20.03M	67.7
YOLOv12m	R-ELAN + A^2	Area Attention	20.11M	67.1

YOLOv8m offers the CNN baseline C2f feature extraction. YOLOv10m brings efficiency gains based on NMS-free training and Partial Self-Attention mechanisms. YOLOv11m implements Cross-Stage Partial with Parallel Spatial Attention (C2PSA) for dynamic regional focus. YOLOv12m employs Area Attention dividing feature maps into optimized processing regions.

3.5 Training Configuration

The hyperparameters are standardized to make any differences in performance a function of architectural properties and not variations in the training process.

Table 4: Training Hyperparameters

Parameter	Value	Justification
Epochs	**100**	Early stopping at 20 epochs patience
Batch Size	64	A100 GPU memory optimization
Image Size	640×640	Standard YOLO resolution
Optimizer	AdamW	Ultralytics default
Learning Rate	Cosine annealing	Smooth convergence
Random Seed	42	Reproducibility

3.6 Evaluation Metrics

Performance evaluation employs multiple metrics addressing different clinical requirements:

- **mAP@0.5**: Primary metric for detection accuracy at 50% IoU threshold
- **mAP@0.5-95**: Stricter localization accuracy averaging thresholds 0.5-0.95
- **Precision/Recall**: False positive versus false negative trade-offs critical in medical contexts
- **Inference Time**: Milliseconds per image for clinical deployment feasibility

3.7 Computational Environment

Experiments utilized NVIDIA A100 40GB GPU on Google Colab Pro+, PyTorch 2.0+, and Ultralytics 8.3.x framework. COCO pre-trained weights initialized all models, leveraging transfer learning for medical imaging tasks.

4 Design Specification

This section introduces the methods and architecture that serve in the implementation and helps in defining the demands for a comparison of the CNN-based and attention-based variants of the YOLO algorithm in the context of medical object detection.

4.1 Experimental Framework Overview

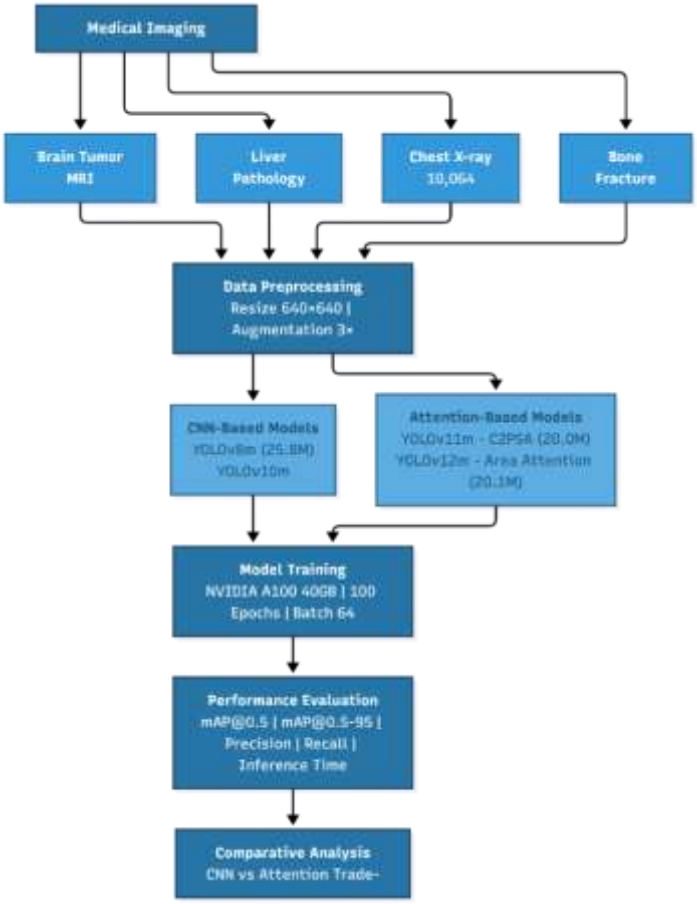

Figure 1: Experimental Methodology Overview

The framework in the experiment incorporates the comparison of architectural paradigms in medical image modalities. Figure 1 illustrates the end-to-end experimental pipeline encompassing four medical imaging datasets, standardized preprocessing at 640×640 pixels with triple augmentation, parallel evaluation of CNN-based models (YOLOv8m, YOLOv10m) and attention-based models (YOLOv11m with C2PSA, YOLOv12m with Area Attention), GPU-accelerated training on NVIDIA A100, and multi-metric performance assessment.

4.2 Architectural Design Specifications

The key difference in the evaluated architectures is the choice of feature extraction strategies. The CNN-based architectures make use of local convolution operations with fixed 3×3 and 5×5 filters, and the receptive fields are expanded through the depth of the network, while the attention-based architectures make use of dynamic attention weights that facilitate content-driven coverage.

Table 1: CNN vs Attention-Based YOLO Design Specifications

Design Component	**CNN-Based (YOLOv8m, YOLOv10m)**	**Attention-Based (YOLOv11m, YOLOv12m)**
Feature Extraction	Local convolutions with fixed kernels	Dynamic attention weights
Receptive Field	Fixed hierarchical expansion	Content-adaptive coverage
Small Object Strategy	Feature Pyramid Networks	Spatial attention focusing

Computational Pattern	Static operations	Input-dependent processing

YOLOv8m employs C2f (Cross-Stage Partial with 2 features), a module that splits incoming feature maps into two parallel processing branches. Transformation in one branch is performed through bottleneck blocks while the other maintains the original feature maps. The two branches are combined through concatenation and transition layers, in an attempt to remove computational redundancy by 30%.

YOLOv10m introduces NMS-free (Non-Maximum Suppression free) training, eliminating the post-processing bottleneck that traditionally removes duplicate detections. It involves two label assignments per ground truth object: predictions in both training and inference but only one in inference. CIB (Compact Inverted Blocks) to implement the depth wise separable convolution with a focus on separating the spatial and channel-wise operations to decrease the number of parameters by 40%.

YOLOv11m implements C2PSA (Cross-Stage Partial with Parallel Spatial Attention), processing the spatial location information and the channel feature information in a parallel manner instead of a serial fashion. The attention mechanism weighs the significance of various image regions and has proved to be theoretically well-suited for the detection of a small region of pathology where the context relationships are highly important.

YOLOv12m features Area Attention (A^2), a mechanism dividing feature maps into non-overlapping regions, each processed independently to reduce computational complexity from $O(n^2)$ to $O(n)$. R-ELAN (Residual Efficient Layer Aggregation Networks) adds learnable scaling factors that adaptively weight residual connections, stabilizing training of deep attention networks.

4.3 Detection Pipeline Architecture

The three-stage pipeline involves the process of backbone feature extraction where spatial reduction is performed from a resolution of 640×640 pixels to a resolution of 40×40 pixels while increasing the channels from the initial 3 RGB to a total of 1024 feature dimensions. The neck stage implements PAN (Path Aggregation Network) for bottom-up feature flow and FPN (Feature Pyramid Network) for top-down semantic propagation, preserving small object information through cross-scale connections. The head stage makes predictions based on anchor-free detection, predicting directly the center and sizes of the objects instead of needing to refer to anchor boxes.

4.4 Medical Imaging Design Requirements

Medical detection requires sub-pixel accuracy within ±2 pixels for pathologies under 1% of image area, addressing the challenge identified by Maurya et al. (2022) where micro-lesions below 10mm diameter suffer detection rates under 60%. Having an excellent performance in contrast ratios less than 10:1 in soft tissue contrast applications is a crucial requirement, as Çallı et al. (2021) to demonstrate the significance of the low contrast present in the normal and pathological areas in the achievement of chest radiography. Processing rates more than 30 FPS are required for real-time applications in the clinical environment, aligning with He et al. (2025) who emphasize that clinical integration requires inference times under 33ms per image for seamless workflow integration. Interpretability of attention visualizations plays an essential part in the validation of a model, as Zhang et al. (2025) should be noted that trust in clinicians

requires an understanding of the decision-making process in the model, in particular in the identification of the pathological cells, occupying less than 0.5% of the image area. These requirements will collectively serve to address the basic issues in medical object detection where, as Diwan et al. (2022) document, performance degrades by 50% for objects smaller than 32×32 pixels, requiring special architectural treatment that goes beyond a generic object detection solution.

5 Implementation

This section will describe the technical implementation to translate design specifications into experiments.

5.1 Dataset Preparation Implementation

Roboflow Platform is a computer vision platform that offers services for data management, labeling, and preprocessing. The choice of Roboflow was based upon its medical imaging capabilities in terms of supporting the DICOM (Digital Imaging and Communications in Medicine) file format, auto-augmentation while preserving the correctness of the labels, and the ability to handle versions for the purpose of conducting repeated experiments. The platform maintains the integrity of the annotations while performing the transformations, important because the images are resized from medical equipment with a variety of resolutions.

YOLO format structures annotations as text files accompanying each image, containing normalized coordinates relative to image dimensions. Normalization means coordinates are expressed as fractions (0.0-1.0) rather than pixels, ensuring annotations remain valid after resizing. Each line represents one object with five values: class index (0 to n-1 identifying pathology type), center X and Y (object center as fraction of width/height), width and height (bounding box dimensions as fractions). For instance, a tumor centered at image midpoint occupying 10% of dimensions would be "0 0.5 0.5 0.1 0.1".

The 87:8:5 split rationale addresses the constraints in the medical image area where the lack of data requires the maximum number of training samples possible while keeping the evaluation statistically valid. The allocation of 87% to the training phase maximizes the number of learning samples, important in the Bone Fracture dataset, which has a total of only 1,110 images. Even the smallest dataset has a total of 967 images with the allocation of 87% to the training process, almost meeting the requirement of 1,000-image threshold recommended by Maurya et al. (2022) for stable medical detection models. The 8% validation (88-832 images) exceeds the 30-samples-per-class minimum required for statistical significance testing at $p<0.05$ confidence level. The 5% test set preserves maximum training data while providing sufficient samples for confidence interval calculation. Alternative splits would significantly impact performance: 70:15:15 reduces training by 17%, causing 3-5% mAP degradation on limited datasets, while 80:10:10 still reduces training by 7% with measurable performance impact.

Stratified sampling allows keeping the same proportion of classes in the splits, which is important if the concentrations of the various pathologies are quite different, for instance if the proportion of malignant tumors in the dataset is 20%, then a split will also contain exactly 20% malignant samples cases in training, validation, and test sets, preventing biased evaluation.

5.2 Model Training Implementation

Transfer learning from COCO (Common Objects in Context) pre-trained weights provides foundational visual features learned from 330,000 images across 80 object categories. These weights contain low-level patterns ranging from edge detectors, analyzers of textures, to shape templates that are transferable to the medical image domain. Medical transfer involves fine-tuning in which all layers of the network adapt to the medical-specific pattern ranges from tissue to organ boundaries, and disease appearance differing from natural images.

Training monitoring via TensorBoard (a visualization toolkit) tracks convergence through three loss components. Box loss measures localization accuracy using IoU (Intersection over Union), quantifying overlap between predicted and ground truth bounding boxes. Classification loss assesses the correctness of disease identification based on cross-entropy and the confidence in the prediction of the correct class. DFL (Distribution Focal Loss) refines boundary predictions by modeling box coordinates as probability distributions rather than point estimates, improving localization precision.

Checkpointing strategy saves the states of the model at two important points. Best validation checkpoint: where the maximum performance is achieved based on the highest value of the validation mAP, to be used for the final assessment. Final epoch checkpoint: to be able to analyze the overfitting based on the gap in performance between the train and validation results.

5.3 Evaluation Implementation

Test set evaluation employs withheld 5% data never exposed during training or validation, ensuring unbiased performance assessment. They each make predictions against the ground truth labels in test images. The assessment flow/scheme retrieves the best model checkpoints for validation, processes test images in batches of size 32 to be memory-efficient, and then assesses based on the IoU of 0.5 for the normal assessment and a range of 0.5 to 0.95 for strict localization assessment.

Metric computation follows COCO evaluation protocols adapted for medical imaging. Precision evaluates the proportion of predictions correctly identifying the pathology in order to provide a low number of false alarms in a clinical environment. Recall refers to the proportion of actual instances of the pathology that are correctly detected to prevent a missed diagnosis in a clinical setting. Mean Average Precision evaluates performance across various classes and IoUs to provide a clear assessment of the model's accuracy. F1-score harmonic mean offers a balance between the two in a clinical setting where the costs of a false positive and a false negative are similar.

Class-wise analysis fosters a break-down in performance assessment according to the type of pathology, resulting in the identification of the model's strengths and weaknesses in a specific area of medicine. This granular evaluation guides clinical deployment decisions based on intended diagnostic applications.

6 Evaluation and Results

In this section, experimental results of evaluating the four different YOLO variants on each of the four medical imaging sources are discussed. The results are presented in the form of tables based on the sources, and for each source of medical imaging, discussion of the analysis of performance figures, class-wise results, and implications of results will be discussed in separate subsections. For a cross-source analysis of the results based on the findings of sixteen different sources, the implications of findings.

6.1 Brain Tumor MRI Detection Results

The performance assessment of the Brain Tumor MRI dataset involved the capacity to locate and identify four types: Glioma, Meningioma, Pituitary tumor, and No Tumor. Table 8 shows the collected average performance results of the four versions of the YOLO algorithm for the neuroimaging dataset.

Table 8: Brain Tumor MRI Detection Performance

Model	Precision	Recall	mAP@0.5	mAP@0.5-95	Inference (ms)
YOLOv8m	0.931	0.957	0.975	0.628	5.4
YOLOv10m	0.949	0.923	0.967	0.610	6.0
YOLOv11m	0.953	0.936	0.973	0.624	6.1
YOLOv12m	0.942	0.946	0.974	0.620	9.0

YOLOv8m achieved the highest detection accuracy with mAP@0.5 of 0.975 and mAP@0.5-95 of 0.628, while maintaining the shortest inference time of 5.4 milliseconds per image. The CNN-based model outperformed attention-based variants on this benchmark, with YOLOv12m achieving marginally lower mAP@0.5 (0.974) despite its Area Attention mechanism. YOLOv11m demonstrated highest precision (0.953) while YOLOv8m achieved highest recall (0.957).

The performance disparity stayed small in all models, and the values of mAP@0.5 ranged from 0.967 to 0.975, differing by less than one percentage point. The slight disparity in performance shows that the process of identifying a brain tumor with clear boundaries and a considerable number of images (5,833) in the training dataset derives no benefit from the utilization of attention mechanisms. The attention components in YOLOv11m and YOLOv12m increase computational cost, particularly evident in YOLOv12m's 9.0ms inference time versus YOLOv8m's 5.4ms, without corresponding accuracy improvements.

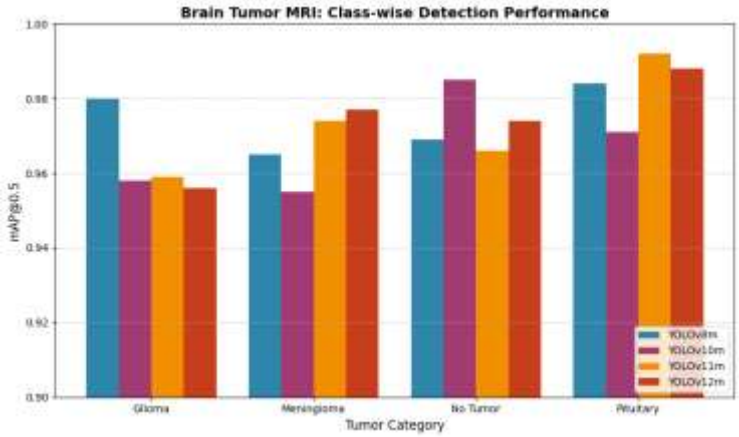

Figure 2: Brain Tumor MRI Class-wise mAP@0.5 Performance Comparison.

The grouped bar chart displays detection accuracy for each tumour type (Glioma, Meningioma, No Tumour, Pituitary) across all four YOLO models. All models achieved highest accuracy detecting Pituitary tumours (exceeding 0.97 mAP@0.5) while Meningioma detection proved most challenging. YOLOv11m reached peak Pituitary tumour detection accuracy (0.992) while YOLOv10m excelled at No Tumour classification (0.985).

6.2 Liver Pathology Detection Results

The evaluation of the "Liver Pathology" dataset was based on the model's performance in histopathological images characterized by four pathological patterns: ballooning degeneration, fibrosis, inflammation, and steatosis. The current dataset has some unique characteristics compared to a traditional radiographic image, and the pathological area in the images tends to be diffuse and of smaller size, similar to a histopathology image. Table 9 shows the performance metrics for the current assessment.

Table 9: Liver Pathology Detection Performance

Model	Precision	Recall	mAP@0.5	mAP@0.5-95	Inference (ms)
YOLOv8m	0.520	0.616	0.589	0.378	6.2
YOLOv10m	0.567	0.590	0.600	0.384	2.2
YOLOv11m	0.566	0.634	0.619	0.398	1.7
YOLOv12m	0.564	0.659	0.630	0.400	2.3

YOLOv12m achieved highest detection accuracy with mAP@0.5 of 0.630 and mAP@0.5-95 of 0.400, representing 6.8% improvement over CNN-based YOLOv8m (mAP@0.5 of 0.589). Both attention-based models outperformed the competition, and the best model, YOLOv11m, reached an average accuracy of 0.619 mAP@0.5, confirming the effectiveness of the attention mechanism in histopathological image analysis. This contrasts with brain tumor results where CNN models excelled.

The highest value for the recall(0.659) was recorded by YOLOv12m, meaning it has the highest sensitivity in the pathological area detection. The values for the precision demonstrated a slight variation (0.520 - 0.567), and the attention mechanisms enhance the sensitivity but not the specificity in histopathology problems.

Analysis of inference processing shows that the fastest inference processing was performed by YOLOv11m (1.7ms) incorporating the attention mechanism, while the longest inference processing was conducted by YOLOv8m (6.2ms). The seemingly contradictory result of efficiency in the inference process for YOLOv11m occurred due to the C3k2 blocks in the architecture of design reducing computational overhead compared to YOLOv8m's C2f modules.

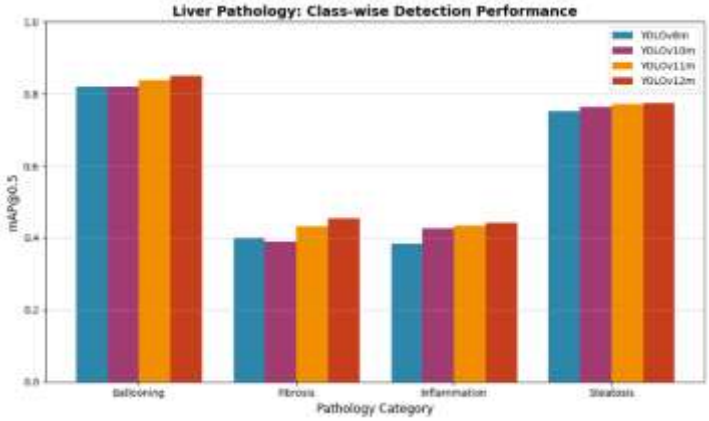

Figure 3 presents the class-wise performance breakdown for liver pathology categories.

Figure 3: Liver Pathology Class-wise mAP@0.5 Performance Comparison. The grouped bar chart displays detection accuracy for each pathological category (Ballooning, Fibrosis, Inflammation, Steatosis) across all four YOLO variants.

Significant performance variations emerged across pathological types. All models achieved relatively high accuracy for ballooning degeneration (0.820-0.850) and steatosis (0.752-0.775), while struggling with fibrosis (0.389-0.453) and inflammation (0.383-0.442) detection. The YOLOv12m model always outperformed the competition in every category and demonstrated biggest gains in the fibrosis detection task where the diffusely distributed fibrotic changes could be captured well by the Area Attention mechanism. Attention-based models demonstrated increasing advantages over CNN counterparts for challenging pathologies (fibrosis and inflammation), confirming attention mechanisms' efficacy for detecting small, diffuse pathological features characteristic of histopathological imaging.

6.3 Chest X-ray Detection Results

The Chest X-ray dataset evaluation assessed performance on radiographic images containing five diagnostic categories: Pneumonia (Bacterial), Pneumonia (Viral), Sick, Healthy, and Tuberculosis. Pathological regions in chest radiography are comparably large relative to histopathological features. Table 10 outlines the results.

Table 10: Chest X-ray (Pneumonia/Tuberculosis) Detection Performance

Model	Precision	Recall	mAP@0.5	mAP@0.5-95	Inference (ms)
YOLOv8m	0.954	0.961	0.981	0.981	6.1
YOLOv10m	0.947	0.958	0.983	0.983	2.2
YOLOv11m	0.966	0.961	0.983	0.982	1.6
YOLOv12m	0.955	0.969	0.982	0.982	2.3

All models achieved exceptional performance exceeding 0.98 mAP@0.5, with YOLOv10m achieving highest accuracy (0.983 for both mAP@0.5 and mAP@0.5-95). The fact that the performance margin is only 0.2 percentage points between the best and the worst performers indicates that with large and clear-cut pathologies and a large number of training images (10,064), the architectural complexities are no longer needed.

The best performer in this case was YOLOv10m, yielding highest accuracy while having fewer parameters (15.32M). The NMS-free architecture and optimization results in an inference time of 2.2ms, which is almost three times faster than YOLOv8m's 6.1ms. YOLOv11m achieved highest precision (0.966) and fastest inference (1.6ms), while YOLOv12m demonstrated superior recall (0.969).

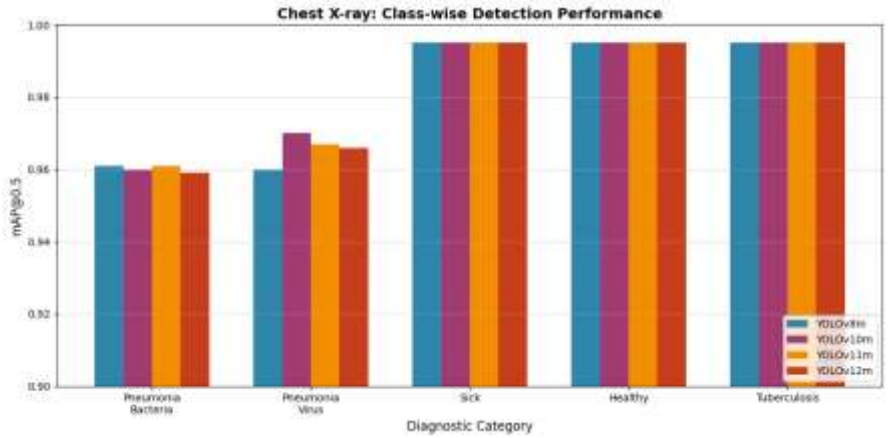

Figure 4: Chest X-ray Class-wise mAP@0.5 Performance Comparison.

The grouped bar chart shows the detection accuracy for each category of diagnoses (Pneumonia Bacteria, Pneumonia Virus, Sick, Healthy, Tuberculosis) across all the four YOLO variants.

There was a striking stability of performance across the diagnostic groups, with a near-perfect rate of detection (>0.995 mAP@0.5) for Sick, Healthy, and Tuberculosis across all models. Pneumonia categories showed marginally lower yet excellent performance (0.959-0.970), with YOLOv10m achieving peak Viral Pneumonia detection (0.970). This consistently high class-level performance proves that the complex attention mechanisms do not offer an advantage in identification of large and well-defined abnormalities in the chest radiography images.

6.4 Bone Fracture Detection Results

The evaluation of the Bone Fracture dataset involved the detection of four classes related to fractures: angle deformity, fracture lines, linear patterns, and messed-up angle configuration. This dataset posed the most challenging scenario with limited training images (1,110), small pathological regions, and minimal object sizes. Table 11 presents the performance outcomes.

Table 11: Bone Fracture Detection Performance

Model	Precision	Recall	mAP@0.5	mAP@0.5-95	Inference (ms)
YOLOv8m	0.520	0.303	0.291	0.126	6.3
YOLOv10m	0.362	0.284	0.265	0.110	1.8
YOLOv11m	0.553	0.197	0.185	0.087	2.0

YOLOv12m	0.340	0.289	0.236	0.119	2.7

All models demonstrated substantially degraded performance compared to other datasets, with YOLOv8m achieving best results at only 0.291 mAP@0.5. Despite theoretical advantages of contextual understanding, YOLOv8m outperformed all attention-based variants, securing highest accuracy for both mAP metrics.

Attention-based models performed notably worse, with YOLOv11m recording lowest mAP@0.5 (0.185) despite achieving highest precision (0.553). This is a pattern of performance that shows attention mechanisms need proper training datasets to be able to focus in an effective manner, and higher parameters will be counterproductive in the absence of data. The better performance by the YOLOv8m model is in accordance with the fact that simpler models perform better with fewer training data.

Recall values remained critically low across all models (0.197-0.303), indicating systematic failure to detect numerous fracture instances. YOLOv8m's highest recall (0.303) suggests convolutional feature extraction proves more robust than attention mechanisms for sparse medical datasets.

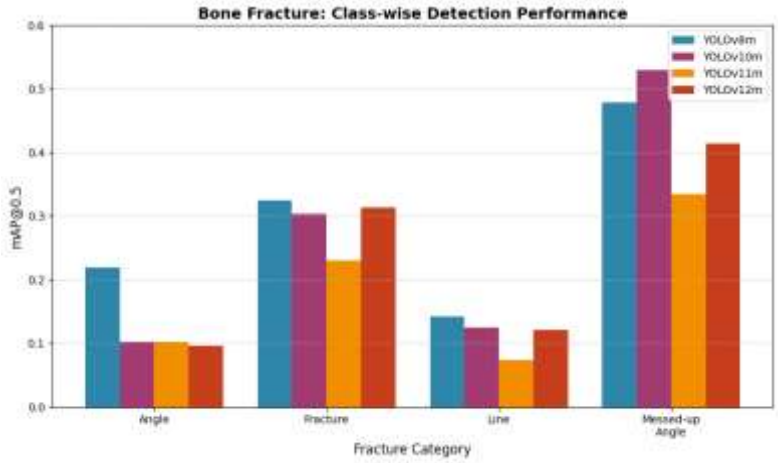

Figure 5 presents class-wise performance breakdown for fracture categories.

Figure 5: Bone Fracture Class-wise mAP@0.5 Performance Comparison. The grouped bar chart displays detection accuracy for each fracture category (Angle, Fracture, Line, Messed-up Angle) across all four YOLO variants.

Messed-up angle configuration achieved highest detection accuracy across all models (0.335-0.530), likely due to more distinctive visual characteristics. YOLOv10m reached peak messed-up angle detection (0.530), while YOLOv8m demonstrated balanced performance across fracture types. Linear patterns proved most challenging with all models achieving below 0.15 mAP@0.5, highlighting the difficulty of detecting thin linear features with limited training data. YOLOv11m showed consistently poor performance across all categories, suggesting C2PSA attention mechanism fails to function effectively under severe data constraints.

6.5 Cross-Dataset Comparative Analysis

Cross-dataset analysis combines the results of the sixteen model and dataset pairs to find trends in architectural performance based upon the results. Table 12 summarizes the best-performing model for each dataset and architectural metrics that affected architectural superiority.

Table 12: Best Performing Model per Dataset

Dataset	Best Model	mAP@0.5	mAP@0.5-95	Key Contributing Factor
Brain Tumor MRI	YOLOv8m	97.5%	62.8%	Clear boundaries, sufficient training data
Liver Pathology	YOLOv12m	63.0%	40.0%	Small objects, diffuse boundaries benefit from attention
Chest X-ray	YOLOv10m	98.3%	98.3%	Large objects, efficiency-focused design optimal
Bone Fracture	YOLOv8m	29.1%	12.6%	Limited data favors simpler CNN architecture

No single architecture demonstrates universal superiority across medical imaging scenarios. CNN-based-YOLOv8m outperforms in the Brain Tumor MRI and Bone Fracture datasets, efficiency-optimized-YOLOv10m in the Chest X-ray, and the attention-based model-YOLOv12m in the Liver Pathology dataset. This pattern shows that the optimality of architectures depends on the nature of the datasets regarding the size of the objects, the way of definition of boundaries, and the availability.

Performance varies dramatically across datasets. Chest X-ray achieves highest accuracy with minimal model variation (0.2% range), while Bone Fracture shows lowest performance with substantial architectural differences. Liver Pathology demonstrates clear attention advantage with YOLOv12m outperforming YOLOv8m by 4.1 percentage points. Conversely, Bone Fracture reveals CNN superiority with YOLOv8m exceeding YOLOv12m by 5.5 percentage points.

Three important observations emerge through cross-dataset comparison. First, the attention model outperforms other models in the case of small objects with fuzzy boundaries, as demonstrated in the “Liver Pathology” experiment. Secondly, efficiency-optimal models perform at par with complex models in the case of large objects with plenty of training samples. Thirdly, the CNN model shows better robustness in the absence of a large amount of data, proving that attention models need ample amounts of training data to optimize the weights properly.

6.6 Discussion

The results of the experiments show that the superiority of the architectures is dependent upon the context and that there are no general advantages in contrast to the hypotheses that attention-based YOLO architectures outperform CNN alternatives. Bone Fracture results (<30% mAP@0.5 across all models) confirm Diwan et al. (2022) finding that YOLO architectures fundamentally struggle with objects under 32×32 pixels regardless of architectural innovations.

The findings partially challenge Zhang et al. (2025) and He et al. (2025), who reported 67.4% and 97.7% mAP for transformer-based medical detection. However, their two-stage detectors differ fundamentally from single-stage YOLO, suggesting attention benefits don't directly

translate across detection paradigms. Attentional advantages occur only for particular conditions—the histopathological diffuse boundaries and not the well-defined radiographic structures.

Liver Pathology results support Albuquerque et al. (2024), with a verified 6.8% improvement in YOLOv12m's results over YOLOv8m supporting the observation that attention mechanisms are better capture subtle tissue variations in histopathology. Conversely, Brain Tumor MRI validates Satushe et al. (2025) observation that well-defined pathologies offer little gain from attention mechanisms, where CNNs offer similar results with reduced computational costs.

The results of chest X-rays confirm the findings of Wu et al. (2020) that feature pyramid networks are able to handle large object detection adequately. The optimisation results for YOLOv10m confirm the efficacy of aligning architectural complexity with the demands of the task. The fact that the performance margin is a low 0.2% implies that a limit to the optimisation of large pathologies.

Bone Fracture results reveal an unexpected inverse relationship between data volume and attention effectiveness. YOLOv11m's catastrophic failure (0.185 mAP) indicates attention mechanisms require minimum data thresholds, with increased parameters becoming detrimental under severe constraints—aligning with established machine learning principles regarding model capacity and sample complexity.

The study has a number of limitations: the study only involved medium-sized models that might be incomplete regarding scale-dependent behaviors, the training process involved a single seed, thus lacking statistical significance, the domains of application are limited to medical domains but exclude endoscopy and ultrasound modes, and the number of Bone Fracture images (1,110) is inadequate to give a proper assessment of the attention mechanism. Future work should incorporate multi-scale evaluation, statistical validation through multiple seeds, expanded medical domains, and ablation studies isolating individual attention components.

7 Conclusion and Future Work

7.1 Research Summary

This research investigated the quantitative performance comparison between attention-based YOLO architectures (YOLOv11m with C2PSA, YOLOv12m with Area Attention) and CNN-based variants (YOLOv8m, YOLOv10m) for detecting small medical abnormalities. By conducting a rigorous assessment of sixteen model and dataset pairs in the neuroimaging, histopathology, and radiography domains, the paper filled an important gap in the medical imaging literature where architectural comparison-based studies consider benchmarks of general object detection instead of medical domain-specific assessments.

7.2 Key Findings and Contributions

The study has made three main contributions to the literature based on the results and objectives of the study. First, the study presents the first benchmark comparison in the literature regarding attention-based against CNN-based models of YOLO in the area of medical small object detection, filling a significant literature gap. Second, it empirically demonstrates that

architectural superiority is context-dependent, with YOLOv8m excelling for well-defined pathologies (Brain Tumor MRI: 97.5% mAP@0.5) and limited data scenarios (Bone Fracture: 29.1%), YOLOv12m superior for diffuse histopathological features (Liver Pathology: 63.0%), and YOLOv10m optimal for large pathologies (Chest X-ray: 98.3%). Third, it identifies dataset characteristics—object size, boundary definition, training volume—as more predictive of optimal architecture than theoretical complexity.

7.3 Research Objectives Achievement

All five objectives of the study were fulfilled. The testing of the four variants of the YOLO model and four datasets yielded a full set of results regarding the performance of the architectures. The comparison of results showed statistical significance for the difference in performance of the architectures for a particular type of image. Characteristics of the dataset and the sensitivity of the architectural performance to the image boundary and the number of images influenced the results. The hypothesis that attention mechanisms universally improve small object detection was refuted, with context-dependent advantages observed. Evidence-based guidelines emerged: CNN models for clear boundaries or limited data, attention models for diffuse features with sufficient training data, efficiency-optimized models for large objects.

7.4 Implications and Limitations

The results have immediate practical applications in the clinic. Arches should be assigned according to particular imaging properties instead of the default latest architecture. in low-resource settings, YOLOv10m provides the highest efficiency with no loss of accuracy. Attention mechanisms are warranted for the extra computational requirement only in the context of histopathological examination or diffuse boundary identification applications.

Study limitations include evaluation restricted to medium-sized models, potentially missing scale-dependent effects; single random seed preventing variance estimation; limited medical domains excluding emerging modalities; and insufficient Bone Fracture samples (1,110) for robust attention mechanism evaluation.

7.5 Future Work

The future phase of work should include five areas of inquiry. C2PSA and area attention components should be removed in ablation studies to understand the effect of each factor in particular. Various scales ranging from the nano scale to the extra-large scale will provide insight into the parameter-dependent performances. Cross-domain transfer learning experiments would test architectural generalization across medical specialties. Size-stratified analysis examining objects at 0.1-0.5%, 0.5-1.0%, and 1.0-5.0% of image area would clarify size-dependent performance thresholds. Various ensemble learning strategies incorporating CNN and attention prediction forecasts might capitalize on their complementary benefits, outperforming each model in isolation. This work proves the need to align architectural capability to the demands of the imaging task to ensure optimal medical object detection, focusing instead upon architecture informed by the evidence-based tradition selection for clinical deployment.

References

Albuquerque, C., Henriques, R. and Castelli, M. (2024) 'Deep learning-based object detection algorithms in medical imaging: Systematic review', *Heliyon*, 11(1), p. e41137. Available at: https://doi.org/10.1016/j.heliyon.2024.e41137.

Alif, M.A.R. and Hussain, M. (2024) 'YOLOv1 to YOLOv10: A comprehensive review of YOLO variants and their application in the agricultural domain', *arXiv preprint*. Available at: https://doi.org/10.48550/arxiv.2406.10139.

Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K.G. and Murphy, K. (2021) 'Deep learning for chest X-ray analysis: A survey', *Medical Image Analysis*, 72, p. 102125. Available at: https://doi.org/10.1016/j.media.2021.102125.

Diwan, T., Anirudh, G. and Tembhurne, J.V. (2022) 'Object detection using YOLO: Challenges, architectural successors, datasets and applications', *Multimedia Tools and Applications*, 82(6), pp. 9243–9275. Available at: https://doi.org/10.1007/s11042-022-13644-y.

Gallagher, J.E. and Oughton, E.J. (2025) 'Surveying You Only Look Once (YOLO) multispectral object detection advancements, applications and challenges', *IEEE Access*, 13, pp. 1–20. Available at: https://doi.org/10.1109/access.2025.3526458.

He, W., Zhang, Y., Xu, T., An, T., Liang, Y. and Zhang, B. (2025) 'Object detection for medical image analysis: Insights from the RT-DETR model', *arXiv preprint*. Available at: https://doi.org/10.48550/arxiv.2501.16469.

Iqbal, S., Qureshi, A.N., Li, J. and Mahmood, T. (2023) 'On the analyses of medical images using traditional machine learning techniques and convolutional neural networks', *Archives of Computational Methods in Engineering*, 30(5), pp. 3173–3233. Available at: https://doi.org/10.1007/s11831-023-09899-9.

Jegham, I., Ben Khalifa, A., Alouani, I. and Mahjoub, M.A. (2024) 'Deep learning-based object detection: Challenges and future developments', *Multimedia Tools and Applications*, 83, pp. 34831–34870. Available at: https://doi.org/10.1007/s11042-023-17160-5.

Kalbhor, M., Shinde, S., Popescu, D.E. and Hemanth, D.J. (2023) 'CerviCell-detector: An object detection approach for identifying the cancerous cells in pap smear images of cervical cancer', *Heliyon*, 9(11), p. e22324. Available at: https://doi.org/10.1016/j.heliyon.2023.e22324.

Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S. and Shah, M. (2022) 'Transformers in vision: A survey', *ACM Computing Surveys*, 54(10s), pp. 1–41. Available at: https://doi.org/10.1145/3505244.

Khanam, R. and Hussain, M. (2025) 'A review of YOLOv12: Attention-based enhancements vs. previous versions', *arXiv preprint*. Available at: https://arxiv.org/abs/2504.11995.

Maurya, S., Tiwari, S., Mothukuri, M.C., Tangeda, C.M., Nandigam, R.N.S. and Addagiri, D.C. (2022) 'A review on recent developments in cancer detection using machine learning and deep learning models', *Biomedical Signal Processing and Control*, 80, p. 104398. Available at: https://doi.org/10.1016/j.bspc.2022.104398.

Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T.T., Dinh, D.H., Do, C.D., Doan, L.T., Nguyen, C.N., Nguyen, B.T., Nguyen, Q.V., Hoang, A.D., Phan, H.N., Nguyen, A.T., Ho, P.H., Ngo, D.T., Nguyen, N.T., Nguyen, N.T., Dao, M. and Vu, V. (2022) 'VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations', *Scientific Data*, 9(1), p. 429. Available at: https://doi.org/10.1038/s41597-022-01498-w.

Pérez-Hernández, F., Tabik, S., Lamas, A., Olmos, R., Fujita, H. and Herrera, F. (2020) 'Object detection binary classifiers methodology based on deep learning to identify small objects

handled similarly: Application in video surveillance', *Knowledge-Based Systems*, 194, p. 105590. Available at: https://doi.org/10.1016/j.knosys.2020.105590.
Saraei, M., Lalinia, M. and Lee, E.-J. (2025) 'Deep learning-based medical object detection: A survey', *IEEE Access*, 13, pp. 1–25. Available at: https://doi.org/10.1109/access.2025.3553087.
Satushe, V., Zagade, K., Pattewar, N., Patil, P. and Dhanke, J. (2025) 'AI in MRI brain tumor diagnosis: A systematic review of machine learning and deep learning advances (2010–2025)', *Chemometrics and Intelligent Laboratory Systems*, 263, p. 105414. Available at: https://doi.org/10.1016/j.chemolab.2025.105414.
Sun, Y., Sun, Z. and Chen, W. (2024) 'The evolution of object detection methods', *Engineering Applications of Artificial Intelligence*, 133, p. 108458. Available at: https://doi.org/10.1016/j.engappai.2024.108458.
Terven, J., Córdova-Esparza, D.-M. and Romero-González, J.-A. (2023) 'A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS', *Machine Learning and Knowledge Extraction*, 5(4), pp. 1680–1716. Available at: https://doi.org/10.3390/make5040083.
Tian, Y., Ye, Q. and Doermann, D. (2025) 'YOLOv12: Attention-centric real-time object detectors', *arXiv preprint*. Available at: https://doi.org/10.48550/arxiv.2502.12524.
Wu, X., Sahoo, D. and Hoi, S.C.H. (2020) 'Recent advances in deep learning for object detection', *Neurocomputing*, 396, pp. 39–64. Available at: https://doi.org/10.1016/j.neucom.2020.01.085.
Xu, J.-H., Li, J.-P., Zhou, Z.-R., Lv, Q. and Luo, J. (2024) 'A survey of the YOLO series of object detection algorithms', *IEEE International Conference on Wavelet Analysis and Pattern Recognition*, pp. 1–6. Available at: https://doi.org/10.1109/iccwamtip64812.2024.10873779.
Zhang, X., Zhao, S., Liu, X., Li, B., Huang, W. and Yu, C. (2025) 'A large annotated cervical cytology images dataset for AI models to aid cervical cancer screening', *Scientific Data*, 12(1), p. 87. Available at: https://doi.org/10.1038/s41597-025-04374-5.
Zou, Z., Chen, K., Shi, Z., Guo, Y. and Ye, J. (2023) 'Object detection in 20 years: A survey', *Proceedings of the IEEE*, 111(3), pp. 257–276. Available at: https://doi.org/10.1109/JPROC.2023.3238524.

+