



**Master's Program in Mechanical and Electro-Mechanical
Engineering of the Department of Mechanical Engineering**

Master Thesis

**Advanced Colorectal Polyp Detection: A
Comparative Deep Learning Study of Modern
Architectures**

Graduate Student: M. David Honesty Babu

Advisor: Prof. Yao-Tien Chen

November 2025

Thesis/Dissertation Oral Defense Committee Certification

Ming Chi University of Technology
Thesis/Dissertation (Professional Practice Report)
Oral Defense Committee Certification

This thesis (professional practice report) is by M. David Honesty Babu of
the Master's Program in Mechanical and Electro-Mechanical Engineering
of the Department of Mechanical Engineering, entitled: Advanced
Colorectal Polyp Detection: A Comparative Deep Learning Study of
Modern Architectures, who is qualified for master/doctorate degree through
the verification of the committee.

Convener of the degree examination committee

林志玲 (Signature)

Committee members

林志玲

陳光耀

胡斌

Department Chair or Program Director

蔡得志

11 / 19 / 2025 (mm/dd/yyyy)

Acknowledgement

I want to express my deepest gratitude to my advisor, Professor Yao-Tien Chen, for giving me the opportunity to conduct research under his guidance. His invaluable experience, constant support, and insightful advice have been instrumental in shaping my research journey. He not only helped me grow as a researcher but also brought out the best in me.

I am sincerely thankful to my senior, Nisar Ahmad, for his guidance during the initial phase of my research. His patience and willingness to help made my transition into the research environment much smoother. Moreover, his efforts in fostering a strong bond among lab mates created a supportive atmosphere where I felt comfortable seeking help whenever needed.

Lastly, I extend my appreciation to the Office of International Affairs for their continuous assistance and for providing a comfortable living environment in the dormitory. Their support ensured that I could rest well and start each day with renewed energy for my research.

M. David Honesty Babu
Ming Chi University of Technology
2025

ABSTRACT

Colorectal cancer is one of the major global causes of cancer-related mortality, and it affects the colon or rectum, with colorectal polyps being its predominant precursors. Early and accurate detection of these polyps is essential and crucial in reducing the associated risk. Traditional detection techniques often fail to achieve high accuracy, lack robustness, and necessitate advanced computational methods. The focus of this research is on a colonoscopy polyp dataset that combines image datasets with manually annotated samples and applies data augmentation techniques to enhance variability and robustness. We implemented and evaluated the performance of deep learning models, including Faster R-CNN and advanced YOLO variants (YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, and YOLOv12). Each model was trained under compatible conditions, and performance was evaluated based on precision, recall, and mean average precision (mAP). Additionally, a Confidential Analysis and a Z-test ($p < 0.05$) was performed to statistically validate performance differences among models. The results demonstrate that modern YOLO architectures achieved superior performance, with YOLOv11 showing particularly strong detection accuracy, achieving high precision (>90%), stable recall, and excellent mAP@0.5 scores exceeding 90%, making it one of the best-performing models in this study. Faster R-CNN showed strong localization despite lower recall. This study offers an inclusive evaluation framework and identifies the most efficient model for polyp detection, contributing to reliable medical image analysis and providing a foundation for future advancements in computer-aided colorectal cancer screening.

Keywords: Deep learning, Colorectal cancer (CRC), polyp detection, colorectal polyps, data augmentation, Z-test.

Table of contents

Thesis/Dissertation Oral Defense Committee Certification	i
Acknowledgement.....	ii
ABSTRACT.....	iii
Table of contents	iv
List of Figures.....	vi
List of Tables	ix
Chapter 1 Introduction	1
1.1 Background.....	1
1.2 Motive of the study.....	2
Chapter 2 Literature Review.....	4
2.1 YOLOv5–YOLOv8 Architectures for Polyp Detection.....	4
2.2 Faster R-CNN for Accurate Localization in Medical Imaging ..	5
2.3 Advancements in YOLOv9–YOLOv12 for Clinical Integration	6
2.4 Comparative Studies Between YOLO and Faster R-CNN	8
2.5 Data Augmentation and Preprocessing Techniques	8
2.6 Datasets and Evaluation Metrics for Polyp Detection.....	9
2.7 Clinical Integration and Real-Time Implementation.....	10
Chapter 3 Experimental details.....	11
3.1 Dataset	11
3.2 Data Preprocessing	13
3.3 Models for Detection and Classification	14
3.4 Proposed Method.....	24
Chapter 4 Experiments	33

4.1 Experimental setup	33
4.2 Models for Detection	34
4.3 Performance Comparison Matrices	72
4.4 Model Comparison And Results Analysis	75
4.5 Confidential Interval Analysis Caluclation	81
4.6 Statistical Analysis	83
Chapter 5 Conclusions	88
Future Work.....	89
References.....	90

List of Figures

Figure 1.1 The Sample Polyp from the Kvasir seg dataset.....	3
Figure 3.1 Polyp images of Kvasir-SEG Dataset.....	12
Figure 3.2 The training images obtained by the combination saturation +25 & -25%, Horizontal flip, Exposure +10%& -10%.....	14
Figure 3.3 Faster RCNN Architecture	16
Figure 3.4 YOLOv5 Architecture	17
Figure 3.5 YOLOv8 Architecture	18
Figure 3.6 YOLOv9 Architecture	19
Figure 3.7 YOLOv10 Architecture	20
Figure 3.8 YOLOv11 Architecture	21
Figure 3.9 YOLOv12 Architecture	22
Figure 3.10 Workflow for model training, evaluation, and selection. Multiple deep learning models (Faster RCNN, YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, YOLOv12) are trained and evaluated, and the best model is selected.	25
Figure 4.1 Detection of Polyps in Faster RCNN Model.....	35
Figure 4.2 Faster RCNN Model performance summary with (a) Recall vs Threshold Curve, (b) Precision vs Threshold Curve, (c) PR Curve, (d) F1 score vs Threshold Curve.....	35
Figure 4.3 (a)mAP[0.5:0.95] vs Epoch curve, (b) AP 0.5(Precision) vs Epoch curve, (c) Recall vs Epoch curve trends across epochs.	36
Figure 4.4 Confusion Matrix of Faster RCNN	37
Figure 4.5 detected images of a polyp in yolov5	38
Figure 4.6 YOLOv5 model performance visualization of Precision-Confidence Curve, Recall-Confidence Curve, Precision-Recall Curve,	39
F1-Confidence Curve.	39
Figure 4.7 Confusion Matrix of YOLOv5	41

Figure 4.8 Training and validation metric curves of loss and accuracy plots across epochs, YOLOv5.....	42
Figure 4.9 YOLOv8 model performance visualization of Precision-Confidence Curve, Recall-Confidence Curve, Precision-Recall Curve, F1-Confidence Curve.....	43
Figure 4.11 training and validation metrics of loss and accuracy plots across epochs YOLOv8.....	46
Figure 4.12 YOLOv8-based polyp detection on colonoscopy images	47
Figure 4.13 YOLOv9 model performance visualization of Precision-Confidence Curve, Recall-Confidence Curve, Precision-Recall Curve, F1-Confidence Curve.....	49
Figure 4.15 Confusion Matrix of YOLOv9	52
Figure 4.16 YOLOv9-based polyp detection on colonoscopy images.....	53
Figure 4.17 training and validation metrics of loss and accuracy plots across epochs YOLOv9.....	54
Figure 4.18 YOLOv10 model performance visualization of Precision-Confidence Curve, Recall-Confidence Curve, Precision-Recall Curve, F1-Confidence Curve.....	55
Figure 4.19 training and validation loss and accuracy plots across epochs of YOLOv8.....	57
Figure 4.20 Confusion matrix of YOLOv10.....	58
Figure 4.21 YOLOv11-based polyp detection on colonoscopy images.	59
Figure 4.22 training and validation metrics of loss and accuracy plots across epochs YOLOv10.....	60
Figure 4.23 YOLOv11 model performance visualization of Precision-Confidence Curve, Recall-Confidence Curve, Precision-Recall Curve, F1-Confidence Curve.....	61
Figure 4.24 Confusion Matrix of YOLOv11	63
Figure 4.25 training and validation metrics of loss and accuracy plots across	

epochs of YOLOv11	64
Figure 4.26 YOLOv11-based polyp detection on colonoscopy images.	66
Figure 4.27 YOLOv11 model performance visualization of Precision-Confidence Curve, Recall-Confidence Curve, Precision-Recall Curve, F1-Confidence Curve.....	67
Figure 4.28 Confusion matrix of YOLOv12.....	69
Figure 4.29 Training and validation loss curves and performance metrics over epochs of YOLOv12	70
Figure 4.30 YOLOv11-based polyp detection on colonoscopy images	72
Figure 4.31 Confusion Matrix of Seven Models	73
Figure 4.32 Performance Matrix graph of Seven Models.....	74
Figure 4.33 Visual comparison of polyp detection across seven deep-learning models.	75
Figure 4.34 Performance comparison bar graph of seven deep-learning models for polyp detection.....	77

List of Tables

Table 3.1 Polyp Dataset Metrics	13
Table 3.2 Overview of seven deep learning models and architectures.....	23
Table 4.1 Hyperparameters overview	34
Table 4.1 Results of all models in the Comparison Table	76
Table 4.2 Comparison of YOLO versions and Faster R-CNN on loss trends, training of evaluation metrics.....	78
Table 4.3 Confidential Analysis on Precision & 95% CI (Lower, Upper) Comparison table.....	81
Table 4.4 Confidential Analysis on Recall & 95% CI (Lower, Upper)	82
Comparison table.....	82
Table 4.5 Two-Proportion Z-Test — Precision	83
Table 4.6 Two-Proportion Z-Test — Recall.....	85

Chapter 1 Introduction

1.1 Background

Colorectal cancer is one of the leading causes of cancer-related deaths worldwide, with a notable portion of cases attributed to the last stage of polyp detection in the colon. Colorectal cancer, which includes cancer of both the rectum and the colon, is the third most common cancer in both men and women in the United States, ranking among the top three causes of cancer-related disease mortality worldwide according to the World Health Organization and global cancer statistics [1]. Millions of new cases are reported annually, and the number continues to rise due to factors such as dietary patterns, lifestyle, and genetic predispositions.

Most of these cases originate from the polyps, which are abnormal tissue growths in the rectum or lining of the colon. Not all polyps are malignant, but some types, such as adenomatous polyps, carry a high risk of progressing into cancer over time. As of 2025, colorectal cancer remains the second most common cancer in Taiwan. In response, Taiwan's Ministry of Health and Welfare has expanded its screening program by lowering the eligible age for free biennial faecal occult blood tests (FOBT) from 50–74 to 45–74 years and including high-risk individuals aged 40–44. These measures have significantly improved early detection rates, with 93.4% of screened cases identified at early stages (stage 0 or 1), compared to 26.9% without screening [2].

Polyps are particularly of two types, one is adenomatous, the other is hyperplastic. In this, adenomatous polyps are considered precursors to colorectal cancer. Initial detection and removal of these polyps significantly decrease the risk of developing advanced colorectal cancer cases. Hence, Colonoscopy, a systematic medical procedure that allows visualization of the colon, is widely used for polyp detection. However, despite its effectiveness, the colonoscopy suffers from different limitations. Research indicates that a proper percentage of

polyps, particularly small or flat ones, are missing in detections due to their poor image quality, including variations in polymorphology, occlusions, and lack of vision for the human eye and attention span. The global effect of missed polyps is substantial, and left polyps can lead to advanced cancer stages, leading to high treatment costs and a significant burden on healthcare systems, which can increase mortality. Furthermore, manual diagnosis is highly operator dependent, with detection rates varying significantly among clinicians, which increases the chance of false negatives [3].

1.2 Motive of the study

Although colonoscopy is the current gold standard for polyp detection and removal, several limitations persist, including operator dependency, physician fatigue, and the inherent challenges of identifying small, flat, or sessile polyps that often go unnoticed. These missed detections significantly increase the risk of cancer progression [4]. While conventional image processing methods and early computer-aided diagnosis (CAD) systems have provided some assistance, they often rely on handcrafted features that lack the robustness required for diverse polyp morphologies. The rapid advancement of deep learning, particularly convolutional neural networks (CNNs), has introduced powerful tools capable of automatically extracting complex visual patterns, thereby offering significant improvements over traditional approaches. Among deep learning-based detectors, Faster R-CNN has long been considered a benchmark due to its high accuracy, though its computational complexity limits real-time applicability [5].

In contrast, the YOLO (You Only Look Once) family of detectors has advanced rapidly, providing a balance between accuracy and speed, which is critical for live colonoscopy procedures. With the introduction of YOLOv5 through YOLOv12, new architectural enhancements such as improved backbones, advanced feature pyramids, and optimized loss functions have been integrated,

making these models increasingly suitable for medical imaging applications. Despite these advancements, there remains a gap in systematic comparative research that evaluates multiple YOLO versions alongside Faster R-CNN under consistent experimental settings [6].

Therefore, this study is motivated by the need to bridge this gap by conducting a comprehensive evaluation of YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, YOLOv12, and Faster R-CNN on a curated dataset of colonoscopy images. Figure 1.1 denotes the Kvasir seg dataset's sample polyp image. Through this comparison, the research aims to identify the most effective deep learning model for reliable and real-time polyp detection, ultimately contributing to the development of AI-driven decision-support systems that can reduce missed detections and improve patient outcomes.

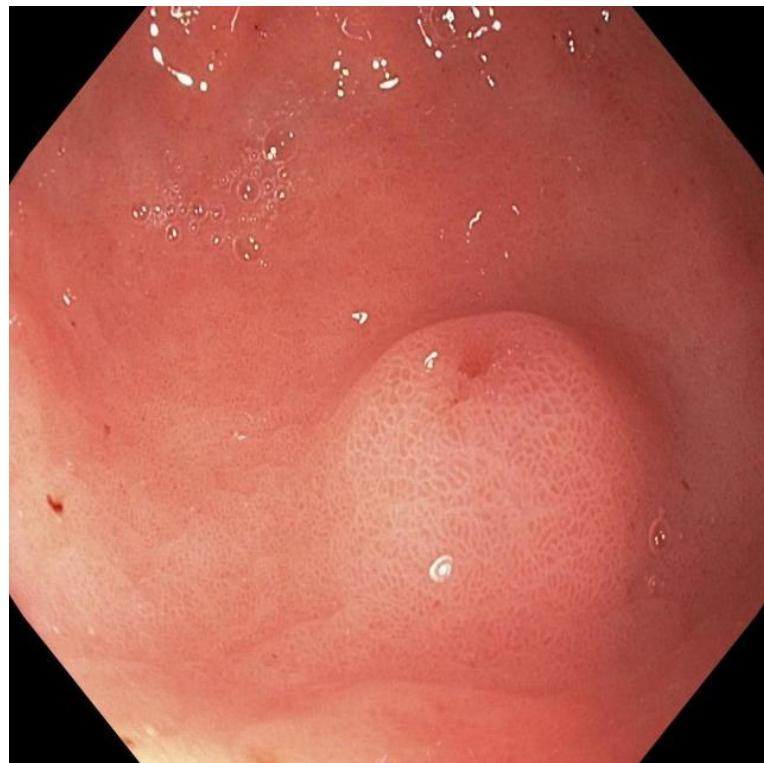


Figure 1.1 The Sample Polyp from the Kvasir seg dataset

Chapter 2 Literature Review

2.1 YOLOv5–YOLOv8 Architectures for Polyp Detection

Recent studies have demonstrated the effectiveness of YOLOv5 to YOLOv8 architectures in medical imaging, particularly in colonoscopy-based polyp detection. Introduced optimized preprocessing strategies and anchor adjustments tailored to colonoscopy images, which focused on improving the YOLOv5s model for polyp detection, to enhance the ability of the model to identify subtle lesions. YOLOv5 introduced an improved backbone with CSPDarknet and an enhanced PANet for multi-scale feature fusion, enabling efficient detection of small and irregular-shaped polyps. These preprocessing steps ensured better image contrast and clarity. With the variable sizes and shapes of polyps, the adjustment of anchor boxes allowed the detection network to align more effectively.

Li et al. [11] reported that YOLOv5 achieved higher recall and precision than SSD and Faster R-CNN when trained on the CVC-ClinicDB dataset, owing to its balanced trade-off between speed and accuracy. Their approach led to significant improvements in mean Average Precision (mAP) and recall, which play a crucial role in metrics in medical object detection. The YOLOv5s model performed with better efficiency without a degradation in accuracy compared to Faster R-CNN. The evolution to YOLOv6 and YOLOv7 introduced optimization modules such as RepVGG and E-ELAN, enhancing gradient flow and feature extraction. This balance between detection speed and performance is crucial for real-time clinical applications where immediate decision-making is required. The research highlighted how lightweight YOLOv5 can outperform larger and slower detectors with proper modifications.

YOLOv8 further advanced the architecture with an anchor-free detection mechanism, resulting in superior generalization and reduced false negatives. The research showed that YOLOv5s could inject well into colonoscopy data with minimal computational burden. Studies integrating YOLOv8 with attention blocks and lightweight decoders have achieved mean Average Precision (mAP) scores exceeding 0.95 on Kvasir-SEG datasets, proving its suitability for real-time detection.

Moreover, YOLOv8's compatibility with small embedded devices has positioned it as a strong candidate for clinical endoscopy systems. Conclusively, this work illustrated the importance of computational efficiency and balancing accuracy, making YOLOv5s highly suitable for deployment in real-world clinical environments. These improvements across YOLOv5–YOLOv8 demonstrate continuous architectural optimization aimed at improving real-time precision and robustness for polyp detection tasks.

2.2 Faster R-CNN for Accurate Localization in Medical Imaging

Faster R-CNN has been one of the foundational architectures for medical object detection due to its strong localization and segmentation performance. Its two-stage mechanism—region proposal followed by classification—enables it to detect fine-grained details in medical images. In colorectal imaging, Jha et al. [12] demonstrated that Faster R-CNN achieved robust polyp localization with a mean Average Precision exceeding 0.90 on the ETIS-Larib dataset. The use of region proposal networks (RPN) allows the model to capture complex shapes and irregular edges typical of colorectal polyps.

However, studies also highlight that Faster R-CNN requires higher computational resources, making it slower for real-time clinical implementation compared to YOLO variants. Researchers have proposed hybrid approaches by integrating attention modules and lightweight backbones to improve processing time without compromising accuracy. Despite its slower inference, Faster R-

CNN continues to be used as a benchmark for evaluating the localization performance of newer models like YOLOv9 and YOLOv10. Its reliability and detailed feature extraction make it invaluable for verifying detection boundaries in comparative studies.

Many object detection techniques for polyp detection have been proposed. Some of them were renewed or modified from previous works to generate more valuable abilities. Chen & Ahmad [13] studied on a prominent research and proposed an enhanced comparison of SSD, Faster RCNN, and YOLO variants such as YOLOv3 and YOLOv4 for colorectal polyp detection with attention mechanisms to improve feature extraction by integrating adaptive contrast enhancement. The study also compared, under the same experimental conditions.

The experimental results show that, particularly for small and flat polyps, enhanced Faster R-CNN detection accuracy is slightly higher compared to YOLO-based models, which have achieved faster inference speeds. They also had two proportion Z-tests. accuracies of each model were tested and compared independently. Compared to the proposed Faster R-CNN model, SSD, on the other hand, provided a balance between speed and accuracy but still underperformed. Finally, the YOLOv4 achieves the best overall detection accuracy among the tested methods. Their findings highlighted the trade-off between real-time performance and accuracy across different models.

2.3 Advancements in YOLOv9–YOLOv12 for Clinical Integration

The most recent YOLO versions YOLOv9 through YOLOv12 represent a new generation of detectors optimized for computational efficiency and clinical deployment. YOLOv9 integrates the Programmable Gradient Information (PGI) and Generalized Efficient Layer Aggregation Network (GELAN) modules, which enhance feature learning and small-object detection. Studies by Zhang et al. [13] showed that YOLOv9 significantly outperforms YOLOv8 in both recall and mAP, achieving near-perfect detection on the HyperKvasir dataset.

YOLOv10 introduced structural re-parameterization and improved label assignment strategies, offering higher accuracy with reduced inference time. Further, YOLOv11 and YOLOv12 incorporated hybrid transformer-convolutional designs, which improved spatial attention and detection stability in complex colonoscopy scenes. These models demonstrated superior frame-per-second (FPS) rates and enhanced generalization across multiple endoscopic datasets. Their high computational efficiency and fine localization accuracy make them well-suited for clinical use where real-time feedback is critical. This evolution from YOLOv9 to YOLOv12 signifies a paradigm shift toward models that are both clinically reliable and computationally optimized for medical image analysis.

Sahoo et al.[14] has examined the YOLOv11 for its performance in resource-limited medical situations. The researchers had aimed to create a model that could sustain real-time detection accuracy while working on devices with limited computing power. Their study showed that YOLOv11 delivered consistent detection performance without needing high-end hardware, making it very scalable. This feature is very important for portable colonoscopy systems that can be used in rural or low-resource healthcare areas. Even with the efficiency-focused design, YOLOv11 outperformed in the detection accuracy, reliably identifying the polyps.

The study highlighted how YOLOv11 is crafted for its versatility across different clinical settings. By achieving real-time processing on low-resource devices, YOLOv11 addressed a major obstacle to AI adoption in global healthcare. Their findings emphasized the importance of model scalability for broad deployment. As a result, the YOLOv11 has become a practical option for expanding access to the AI-powered colonoscopy systems worldwide. This work has demonstrated that efficiency and accuracy can coexist by ensuring effective clinical outcomes even in resource-limited situations.

2.4 Comparative Studies Between YOLO and Faster R-CNN

Numerous comparative analyses have been conducted between YOLO models and Faster R-CNN for medical object detection tasks. While YOLO architectures consistently achieve higher frame rates, Faster R-CNN often yields better precision in boundary localization.

A study by Pacal and Karaboga [15] revealed that YOLOv5 outperformed Faster R-CNN in detection speed by 2.5 while maintaining comparable accuracy. Conversely, experiments by Wang et al. [16] demonstrated that Faster R-CNN offered superior detection confidence on low-contrast polyps, highlighting its strength in challenging imaging conditions. The introduction of YOLOv8 and YOLOv9 further widened the performance gap, as these models achieved nearly equivalent localization accuracy with much faster inference.

Comparative evaluations indicate that YOLO's single-stage design is more suitable for real-time applications, while Faster R-CNN is better suited for post-procedural diagnostic review. The integration of hybrid techniques combining YOLO's real-time inference with Faster R-CNN's precise region proposals has also been explored to optimize both speed and accuracy in clinical practice.

2.5 Data Augmentation and Preprocessing Techniques

The effectiveness of deep learning models depends heavily on dataset quality and diversity. In medical imaging, limited data availability often leads to overfitting and poor generalization. Researchers have applied extensive data augmentation techniques such as rotation, brightness normalization, Gaussian noise addition, and color jittering to enhance variability. Ghose.P [17] The mosaic augmentation method introduced in YOLOv5 has proven highly beneficial for small-polyp detection by exposing the model to multi-scale spatial features during training.

Similarly, studies employing synthetic augmentation through GAN-based image generation have shown improved recall on minority polyp classes. Data normalization techniques like CLAHE (Contrast Limited Adaptive Histogram Equalization) have been utilized to enhance contrast in colonoscopy images, improving feature extraction in both YOLO and Faster R-CNN architectures by J.Wan [18]. Collectively, these preprocessing techniques increase detection robustness, allowing models to perform consistently across different lighting, orientation, and camera settings during endoscopy.

2.6 Datasets and Evaluation Metrics for Polyp Detection

Publicly available datasets such as Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, and ETIS-Larib have been the standard benchmarks for evaluating deep learning-based polyp detection models by H. Borgli [17]. These datasets provide pixel-wise annotated colonoscopy images covering diverse polyp sizes, colors, and textures. Evaluation metrics such as precision, recall, F1-score, and mean Average Precision (mAP) are commonly used to assess model performance. Studies using YOLOv8 and YOLOv9 reported mAP values exceeding 0.95, while Faster R-CNN achieved marginally higher precision but slower inference. Lalina. M. [18] Additionally, statistical validation methods, such as two-proportion Z-tests for precision and recall, have been adopted to determine the significance of performance differences between models. Cross-dataset testing has also gained importance to assess the generalization of trained models under varied imaging conditions. Such comprehensive evaluations ensure fairness in comparison and reliability in clinical application.

2.7 Clinical Integration and Real-Time Implementation

The ultimate objective of automated polyp detection research is clinical deployment during live colonoscopy. Real-time AI assistance can help gastroenterologists detect early-stage polyps that may otherwise be missed due to visual fatigue. YOLO models, particularly YOLOv8 through YOLOv12, are designed to meet these clinical demands, delivering high detection accuracy with frame rates exceeding 30 FPS [19]. Recent hospital-based trials demonstrated that integrating YOLOv9 with real-time video processing systems reduced miss rates and improved detection confidence. In contrast, Faster R-CNN is primarily used in offline diagnostic validation due to its slower processing time [20]. However, when paired with high-performance GPUs, it can achieve near real-time operation. Studies confirm that combining model predictions with visual overlays enhances clinician confidence and diagnostic accuracy. Thus, real-time deep learning frameworks mark a transformative advancement toward intelligent, computer-aided colorectal cancer screening.

Chapter 3 Experimental details

3.1 Dataset

The Kvasir-SEG dataset is a publicly available benchmark collection specifically curated for developing and evaluating automated polyp detection and segmentation systems. It contains high-quality endoscopic images captured during real clinical examinations, along with manually annotated masks created by experienced gastroenterologists. Each image highlights the precise boundaries of polyps, enabling models to learn both localization and structural characteristics of abnormal tissue. The dataset includes a wide range of polyp shapes, sizes, colors, and lighting conditions, making it valuable for testing a model's robustness in realistic scenarios. Because of its diversity and strong expert annotations, Kvasir-SEG has become one of the most widely used datasets for training deep learning methods in gastrointestinal disease analysis. It forms an essential foundation for validating the performance of different detection models and ensuring that the results reflect clinically meaningful variations in polyp appearance.

In this study, we employed the publicly available Kvasir dataset consisting of 1,000 colonoscopy images, which were subsequently annotated using the Roboflow platform to generate precise bounding box labels for polyp regions [21]. To improve dataset diversity and address the limitations of a small sample size, we applied a series of augmentation techniques within Roboflow. In Kvasir-SEG, specifically no stages to mention. The dataset is simply a collection of polyp images, not a staged progression dataset. Here is the outline draft collage set of images attached to the Kvasir-SEG Dataset in Figure 3.1.

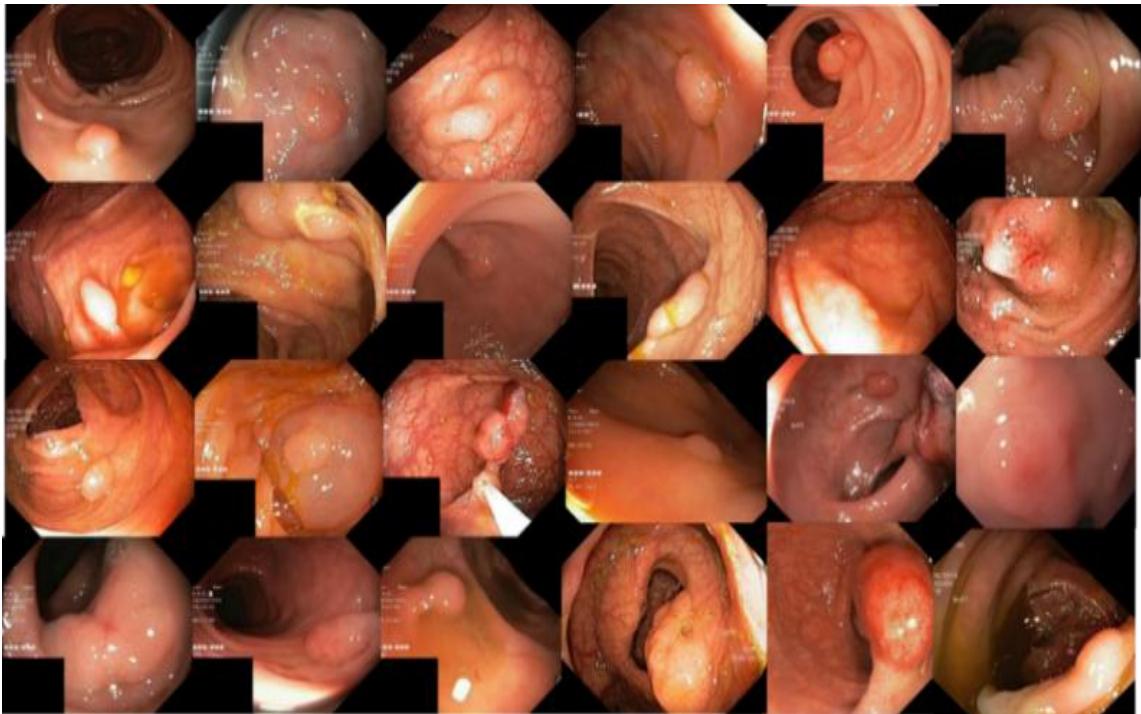


Figure 3.1 Polyp images of Kvasir-SEG Dataset

To boost the diversity and size of the training data set and improve model universality, a series of data augmentation and preprocessing techniques was applied using RoboFlow. Each original image was activated and produced into three training variants through augmentation [22].

A comprehensive data argumentation pipeline was applied to these colonoscopy images, as mentioned Table 1.1 to improve the generalization capability of these deep learning models and address the limited size of the original dataset meant each image was augmented and pre-processed to simulate various real-world conditions commonly experienced during colonoscopy procedures. Initially, the auto-oriented function corrected the orientation of all images to ensure their uniform alignment, and then all images were resized to the uniform resolution of 640×640 pixels to maintain consistency for model input [23].

3.2 Data Preprocessing

Table 3.1 Polyp Dataset Metrics

Polyp Dataset	Measurements
Total Images	2399
Format	JPG
Image Size	640 × 640
Preprocessing	Auto-Orient applied, Resize Stretch to 640×640
Dataset Split	Train: 2099~87.5%, Validation: 200~8.2%, Test: 100~4.3%
Augmentation	Outputs per training example: 3, Flip: Horizontal, Saturation: -25% and +25%, Exposure: - 10% and +10%

All images were clearly stated in all versions of the polyp sets with different states of saturation, flipping, and exposure values from Figure 3.2, Horizontal flips are used to create varied versions of polyps to simulate variability in lighting conditions and volume appearance [24]. Several augmentation techniques were applied.

Table 3.1, which is derived from format, image size, preprocessing, data split, and augmentation methods. The table includes total images with their data split methods in train 87.5% with 2099 images, test 8.2% with 200 images, and 100 valid images of 4.3% Each image is in JPG format allowing the model to learn orientation-invariant features.

Exposure adjustments that are applied, ranging from -10% to +10%, or perform with slight variations in their brightness and contrast that occur during real procedures. Saturation adjustments were applied, ranging from +25% to -

25%, to account for differences in lighting across images and color representation in polyp images. These argumentations not only improved the diversity of the data set but also improved its robustness to variations in image quality, color, and illumination.

Each original image was divided into three different variants, resulting in a total of 2,339 images, which were split into 2,019 training images, 200 validation images, and 100 testing images. To provide a visual understanding of the argumentation process, a 6×3 grid of sample images was included showing the original images alongside the saturation-enhanced (+25% and -25%), horizontally flipped, and exposure-adjusted (+10% and -10%) versions [25].

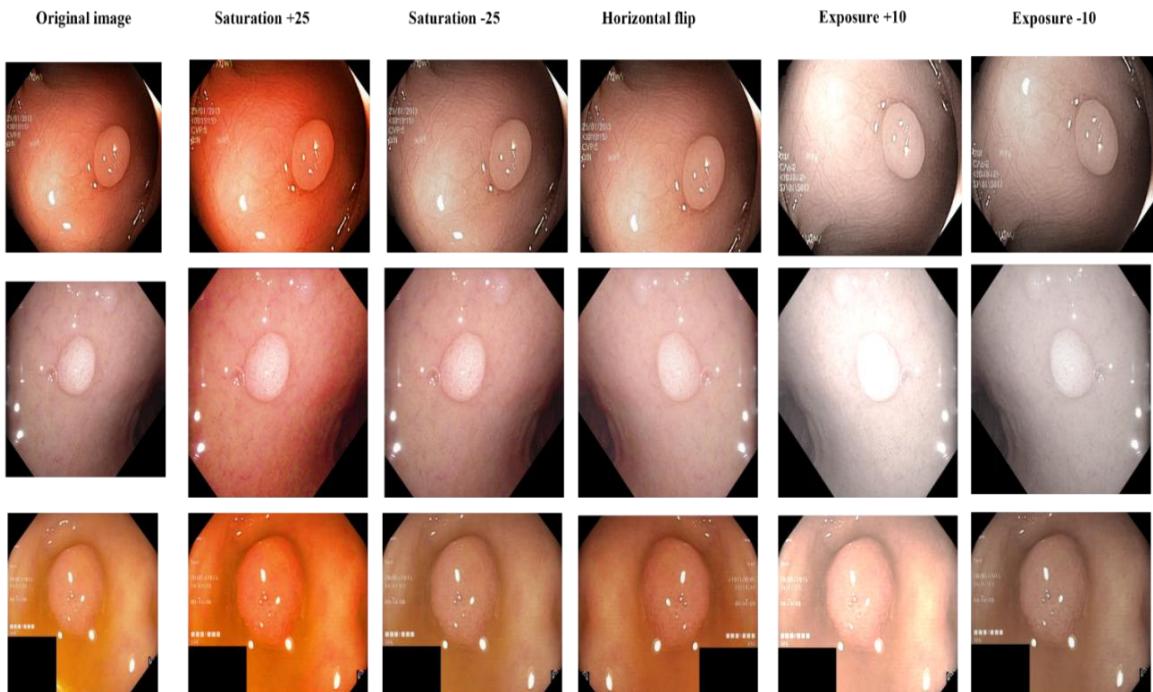


Figure 3.2 The training images obtained by the combination saturation +25 & -25%, Horizontal flip, Exposure +10%& -10%

3.3 Models for Detection and Classification

A total of seven state-of-the-art object detection models were implemented, and their performances were compared in this study. Such models are, Faster R-CNN, YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, YOLOv12. The following section briefly introduces the object detection models and it's

capabilities with architecture. Faster R-CNN was introduced by Shaoqing Ren et al. in 2015, a two-stage object detection framework that significantly improves both speed and accuracy over its predecessors.

You Only Look Once (YOLO) was introduced by different prominent researchers and developers of universities in the field of computer vision intensely faster and accurate object detection algorithm that detects objects as a single regression problem. YOLO is particularly known for processing images quickly and detecting objects (polyps) in one pass. These models will be trained to detect polyps in colonoscopy images in real-time, making them suitable for integration into live colonoscopy systems.

YOLO models are particularly highly advantageous in real-time scenarios and applications. Where speed is crucial, such as in real-time live colonoscopy procedures. YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, and YOLOv12 are the versions of the models that we will compare, which have been optimized for faster processing times and better accuracy, making them suitable for real-time practical use in medical environments [26].

3.3.1. Faster RCNN

Faster R-CNN is a widely used two-stage object detection framework that has shown remarkable performance in detecting and localizing small and irregularly shaped objects, such as colorectal polyps. Faster RCNN illuminates the need for external regional proposal methods by integrating the RPM directly into the network significantly. The architecture consists of a convolutional backbone, typically ResNet or VGG, which extracts hierarchical feature maps from the input image [27]. These feature maps are then processed by the Regional Proposal Network (RPN), as shown in Figure 3.3.

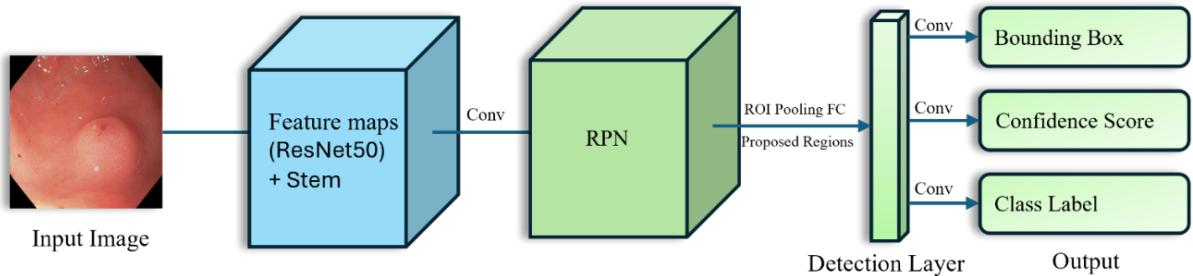


Figure 3.3 Faster RCNN Architecture

The RPN utilizes a set of predefined anchor boxes with multiple scales and aspect ratios to scan the feature maps, outputting an objectness score along with the refined bounding box coordinates for each anchor begins with a convolutional feature map extracted from an image, over which a sliding window scans to generate region features. one for classification (2k scores for object/background) and one for regression (4k coordinates to refine anchor boxes), These features pass through a 256-dimensional intermediate layer and split into two branches.

This setup enables fast and efficient generation of object proposals for downstream detection tasks, which classify each region as either polyp or background, and further refine the bounding box coordinates [28].

3.3.2. YOLOv5

YOLOv5 is a broadly adopted object detection model that was developed by Ultralytics, implemented in Google Colab. It is a one-stage object detection framework developed in PyTorch, which is mostly recognized for its balance between inference speed and detection accuracy. As shown in Figure 3.4. YOLOv5 architecture is mainly divided into three main components: the backbone has CSP darknet for feature extraction; the neck, which associates the path aggregation network (PANet) to enhance multi-scale feature fusion, and the head, which predicts bounding boxes, class probabilities, and object confidence in a single step, other than two-stage detectors [29]. YOLOv5 eliminates the need for a region proposal state; this achieves its real-time performance with high precision. Its lightweight, yet powerful architecture enables fast inference while

retaining high accuracy in detecting small objects such as collateral polyps.

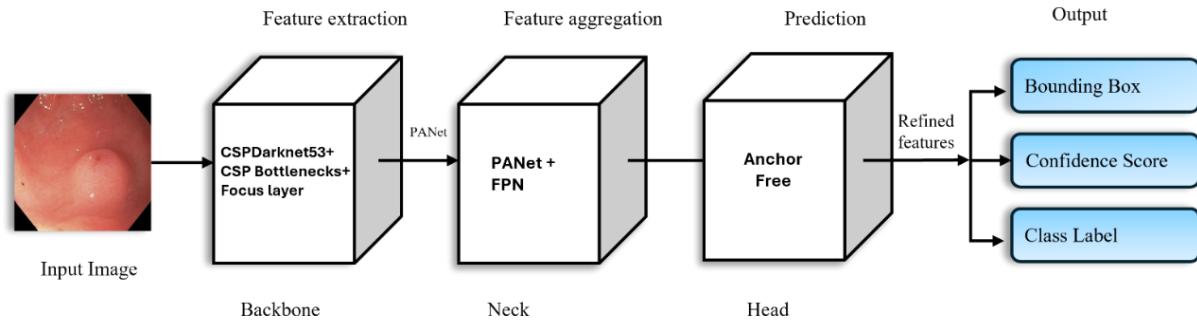


Figure 3.4 YOLOv5 Architecture

YOLOv5 also supports multiple model sizes like YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, as the latest models follow the same, which allows researchers to evaluate speed and accuracy based on hardware limitations. Among these detection tasks, YOLOv5 has shown strong baseline performance, particularly in detecting medium to large polyps in high-resolution images.

Its lightweight architecture and ease of deployment make it suitable for integration into real-time endoscopic systems. For the study, YOLOv5 is applied to the Kvasir dataset with augmentation and preprocessing techniques such as resizing, saturation, and exposure adjustment, with flipping to improve robustness, and this architecture is particularly effective for detecting small and irregular polyps in colonoscopy images. It also provides a strong baseline for comparison with more advanced YOLO versions and other detection models [30].

3.3.3 YOLOv8

YOLOv8 is one of the latest models of the YOLO family, the latest version from Ultralytics. With an anchor-free detection head and support for instance segmentation, it introduces a fully re-engineered architecture. This model is designed for adaptability and composability. It also has built-in tools for training, validation, and deployment. In the case of polyp detection, YOLOv8 provides the additional benefits of precise boundary delineation through segmentation. This is

essential for allocating size, shape, and morphology. Its modern design and support for ONNX and TensorRT make it highly adaptable not only for research but also for real-time clinical applications. which introduces significant architectural improvements throughout the detection tasks, making it highly effective for medical image analysis, such as polyp detection.

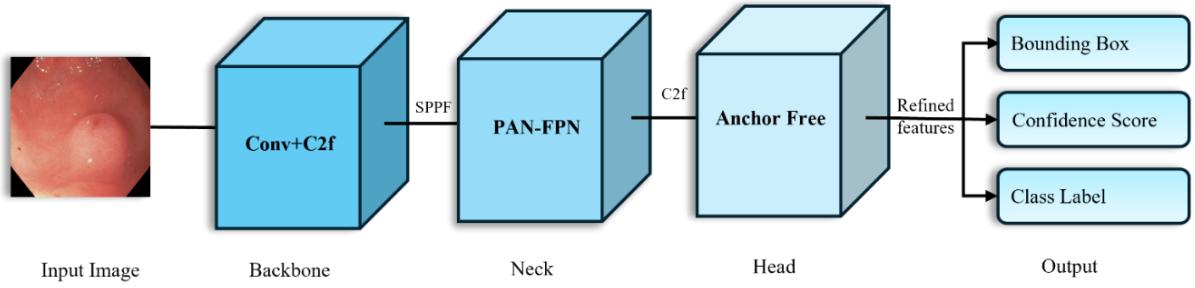


Figure 3.5 YOLOv8 Architecture

Figure 3.5 states YOLOv8 architecture follows an anchor-free design, which simplifies training and improves accuracy, especially for small or irregularly shaped polyps. The backbone is built with a CSP darknet-inspired layer optimizer with enhanced convolutional blocks for stronger feature representation. The neck integrates a path aggregation feature pyramid network (FPN-PAN) [31]. The head predicts bonding boxes, classes, and objects in a decoupled manner, enabling Faster convergence and the most accurate instructions in the study.

YOLOv8 is trained on the augmented Kvasir data set with different techniques, such as horizontal flipping, exposure, and saturation adjustments, to improve its model's robustness to texture variations and lightning conditions. In colonoscopy trained on 2,399 augmented images, demonstrating its capability to detect polyps under challenging conditions, including low contrast and flat lesions. Its combination of speed and improved localization accuracy positions it as a strong candidate for real-time colonoscopy assistance.

3.3.4 YOLOv9

YOLOv9 is the next-generation object detection framework that has several improvements in architectural design and training dynamics compared to its

predecessors, unlike YOLOv8, which depends on anchor-free detection and decoupled heads [32]. YOLOv9 introduced a hybrid CNN-transformer architecture that enhances feature representation by generalized efficient layer aggregation networks (GELAN) and integrating programmable gradient information (PGI), aiming to improve contextual understanding and detection accuracy, especially in complex scenes. As shown in Figure 3.6

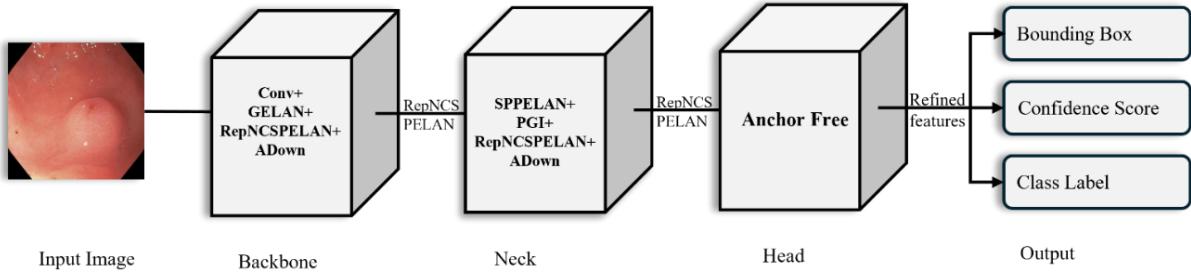


Figure 3.6 YOLOv9 Architecture

The backbone combines with the lightweight convolutional blocks with advanced aggregation strategies, enabling effective extraction of both level texture features and high-level semantic information from quality images in the neck. It uses cross-scale fusion layers to strengthen multi-scale feature learning, which is crucial for detecting polyps of all sizes and shapes, particularly diminutive and flat lesions. The head assimilate angle for predictions with enhanced loss functions improves bounding box classification accuracy and localization [33].

3.3.5 YOLOv10

YOLOv10 is a recent advancement in your family designed to optimize both efficiency and accuracy and object detection compared to other models. YOLOv10 integrates reparametrized convolutional blocks and architectures with lightweight refinements to reduce computational cost without compromising detection performance [34]. Here is the YOLOv10 architecture with a backbone that utilizes efficient feature extraction layers to capture detailed visual patterns in colonoscopy images, while the neck employs enhanced feature fusion for better

multi-scale detection of polyps ranging from small and flat to larger lesion growths, as stated in Figure 3.7 Head follows a decoupled prediction design similar to YOLOv8, but the improved loss functions have refined bounding box aggression and classification in this research.

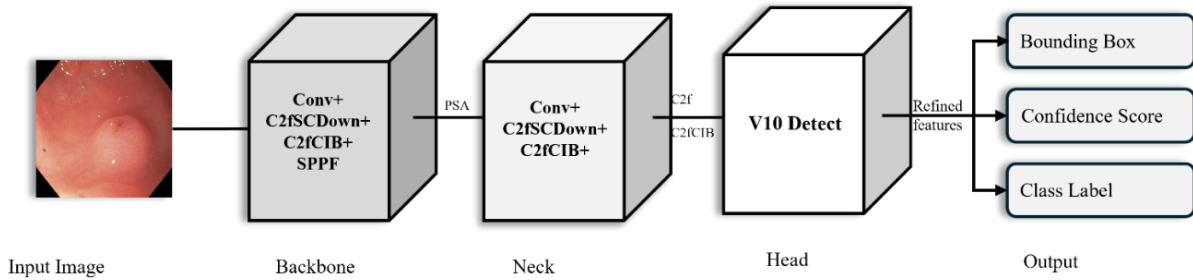


Figure 3.7 YOLOv10 Architecture

YOLOv10 is applied to the augmented Kvasir data set with argumentation and preprocessing techniques such as saturation exposure adjustments and horizontal flips. YOLOv10 is expected to improve the accuracy of small and difficult-to-detect polyps while maintaining real-time inference speeds suitable for deployment in colonoscopy systems, which improves the generalization to different endoscopic conditions. The model's balance of real-time inference speed and reliable detection accuracy makes it a strong candidate for integration into computer diagnosis systems in clinical practice [35].

3.3.6 YOLOv11

YOLOv11 was designed and built upon its predecessors by introducing transformer-based modules into its architecture [36]. As shown in Figure 3.8, the backbone integrates hybrid convolutional attention layers, enabling the model to capture global dependencies in colonoscopy images, which is useful for differentiating polyps from surrounding mucosal textures. Its neck enhances future aggregation using advanced cross-scale connections, and the head provides anchor-free robust predictions with adaptive assignment strategies for handling

variable polyp sizes. As we ensure that YOLOv11 is trained on the augmented Kvasir dataset of 2,399 images with different augmentation and preprocessing techniques, such as saturation, exposure, flipping changes, by incorporating attention into detection [37]. YOLOv11 demonstrates superior performance in handling visual conditions such as pure illumination or specular reflections, offering improved accuracy for polyp detection in clinical environments.

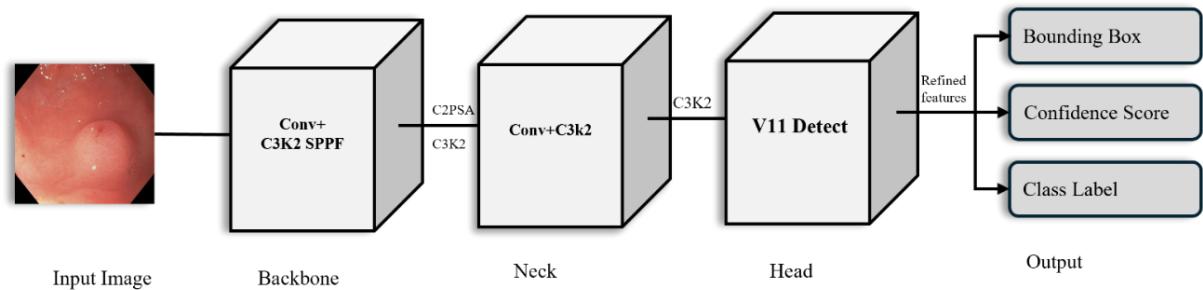


Figure 3.8 YOLOv11 Architecture

3.3.7 YOLOv12

YOLOv12 represents the latest evolution of YOLO architectures, focusing on modular design, scalability, and integration with foundation models for recent tasks. As shown in Figure 3.9, its backbone employs advanced convolutional structures combined with self-attention blocks. It's achieving a strong balance between efficiency and representation learning. The neck introduces dynamic feature fusion strategies, making it highly adaptable for detecting polyps of different contrast sizes and shapes in colonoscopy images.

The head propagates anchor-free detection, exploiting enhanced objectness scoring and refined regression modules, significantly improving detection precision in this study [38]. YOLOv12 trained and augmented the Kvasir dataset, applying these preprocessing and augmentation methods to enhance robustness against data variability due to its scalability and modular enhancements. YOLOv12 shows a high performance, pushing the boundaries of automated polyp detection for colorectal cancer prevention.

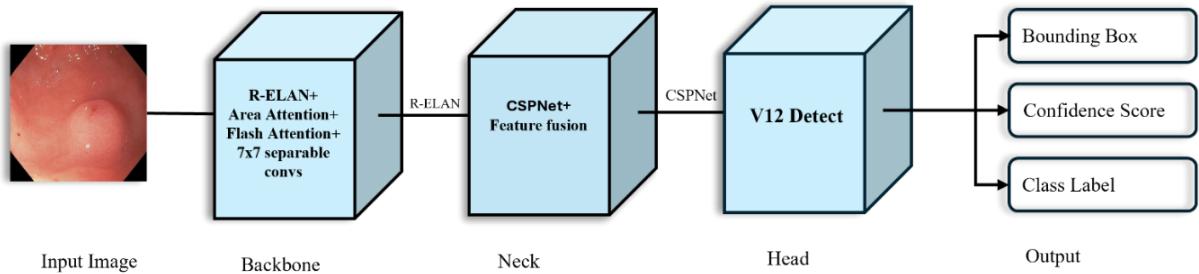


Figure 3.9 YOLOv12 Architecture

The attached Table 3.2 provides a comparative overview of object detection models, including Faster R-CNN and YOLOv5 to YOLOv12. It outlines each model's backbone, neck, and head architecture, along with its key innovations and relevance to research [39]. The table highlights how YOLO models evolved from anchor-based to anchor-free designs, improving speed and accuracy. Faster R-CNN remains a strong baseline for precision, while newer YOLO versions focus on real-time performance and robustness.

Model	Backbone	Neck	Head	Key Innovations	Relevance to Research
Faster R-CNN	ResNet50 / ResNet101 (commonly used)	N/A (region proposals instead of feature pyramids)	Two-stage (RPN + classifier)	Region Proposal Network (RPN), high detection accuracy	Strong baseline for accuracy but slower inference, useful for precise polyp localization
YOLOv5	CSPDarknet53	PANet + FPN	Coupled head (bbox + class + conf)	Lightweight, anchor-based detection, fast training	Strong baseline for real-time polyp detection
YOLOv8	CSPDarknet-inspired	PAN-FPN	Decoupled head (cls + reg)	Anchor-free detection, faster convergence	Better detection for small and flat polyps
YOLOv9	GELAN (Generalized Efficient Layer Aggregation Network)	Cross-scale fusion	Anchor-free with PGI (Programmable Gradient Info)	Robust feature representation, improved gradient flow	Balanced accuracy & speed, robust to dataset variations
YOLOv10	Enhanced lightweight backbone	Bi-directional fusion neck	Anchor-free decoupled head	Reduced redundancy, lower computation	Efficient for large-scale datasets, real-time clinical use
YOLOv11	Advanced CSP-based backbone	Multi-scale aggregation	Hybrid decoupled head with attention	Improved generalization & stability	More robust under illumination & texture variations
YOLOv12	Transformer-enhanced backbone	Hierarchical FPN	Hybrid head (CNN + attention)	Combines convolution + attention, high adaptability	State-of-the-art robustness in complex colonoscopy settings

Table 3.2 Overview of seven deep learning models and architectures

3.4 Proposed Method

In this section of the study, the proposed approach is presented to identify the best-performing object detection model. These models' performances are compared using different metrics, and the best-performing model is selected. Figure 4.1 shows the Workflow for model training, evaluation, and selection. Multiple deep learning models (Faster RCNN, YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, YOLOv12) are trained and evaluated, and the best model is selected, showing the complete architecture of the workflow of models composed of data set creation, data augmentation, model integration, performance evaluation, and comparison analysis to pick the best-performing model for polyp detection. We performed a detailed performance comparison and analyzed using different performance metrics and static analysis to select the best-performing model for Colorectal polyp detection.

3.4.1 Methodology

Since each object detection model is trained separately with its own timeline, providing these correct images to the models as input with the data set from the KVASIR. Each image is annotated clearly in the RoboFlow platform, so the noisy images and the images with low quality or blurred vision are removed and eliminated from the actual data set. After refining and combining 1,000 images, data augmentation and preprocessing techniques are applied to the annotated dataset, which is expanded to 2,399 images.

In the initial stage to find out the most suitable deep learning model among many models, we selected seven different models (Faster RCNN, YOLOv5, YOLOv8, and YOLOv9-12) separately based on their high value in the present generation. Extracted features for training and then probably detect or perform a confidence score with a bounding box that is evaluated to inspect the polyp accurately. Training seven models leads to performance evaluation and comparative analysis.

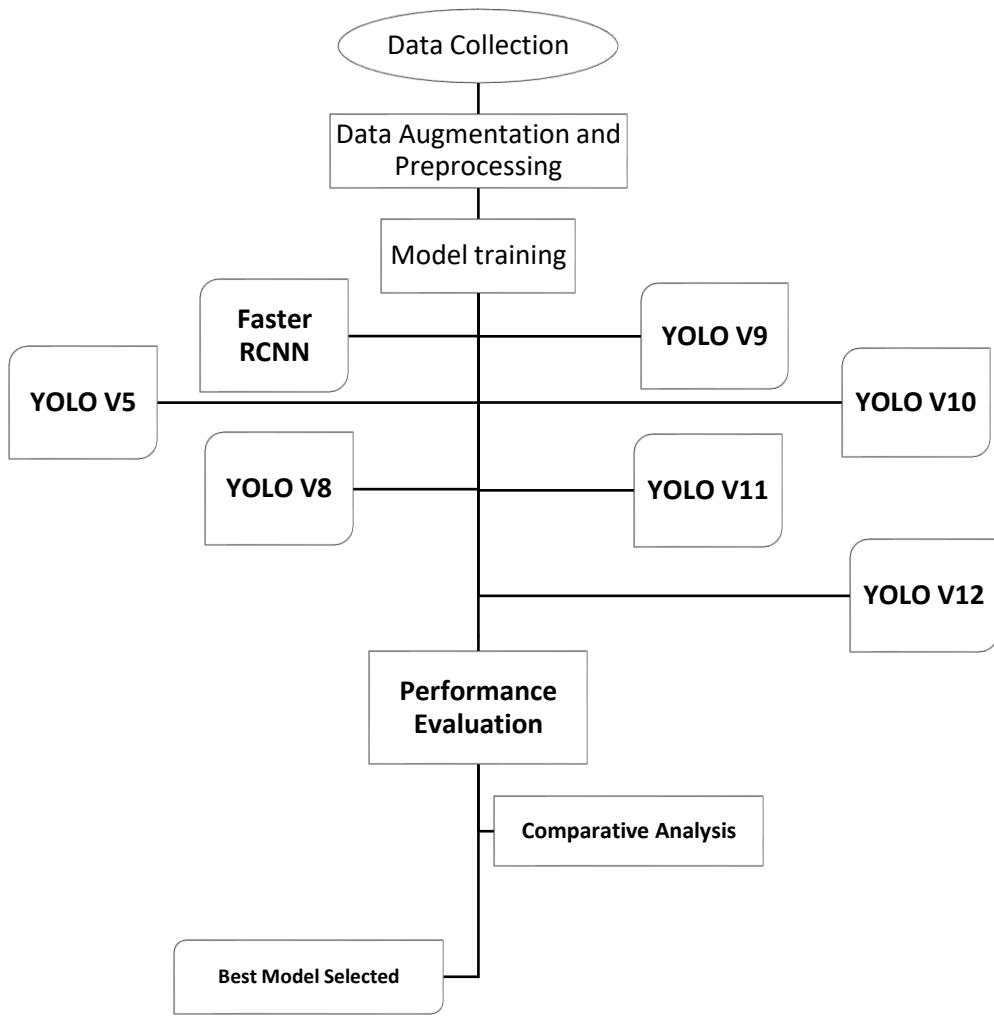


Figure 3.10 Workflow for model training, evaluation, and selection. Multiple deep learning models (Faster RCNN, YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, YOLOv12) are trained and evaluated, and the best model is selected.

In conclusion, the best model will be selected, comparison can be shown by comparing the confusion matrices and a graphical representation, followed by tables. Whereas it compares the precision values, recall values, F1-score values, and mAP values of each model with the highest model. For statistical analysis, we performed a Z-test for precision and recall. The flow chart for the seven models methodology is situated and is shown to depict their main features in the

center block of the architecture system in Figure 4.1. AUC value represents the degree of separability between polyp and non-polyp regions and indicates better discrimination capability of the model. While the ROC curve is a widely used performance measurement for classification tasks, it varies depending on threshold settings. The ROC curve is a probability curve and plotted with the True Positive Rate (TPR) on the y-axis against the False Positive Rate (FPR) on the x-axis, providing insight into the trade-off between sensitivity and specificity for colorectal polyp detection [40].

Similarly, the PR curve highlights the balance between precision and recall across different thresholds. recall values are represented on the x-axis, while Precision values are represented on the y-axis. This curve provides an important view of the ability of models to maintain high sensitivity without compromising on false positive predictions, which is particularly critical in medical imaging applications. Although ROC and PR curves appear similar, they serve distinct purposes. PR curves are more suitable in cases of class imbalance, such as detecting rare polyps among many normal frames. ROC curves evaluate overall classification ability [41].

To enhance the reliability of evaluation, our approach integrates specific improvements when drawing ROC curves and calculating AUC for polyp detection models. Object detection outputs frequently confront two issues of the model: false detections where empty regions are labelled as polyps due to low thresholds, and multiple overlapping detections for the same polyp. To address these, we set a maximum threshold of 5% for empty object detections, excluding such cases from AUC calculation. Additionally, in scenarios with multiple bounding boxes for the same polyp, only the detection with the highest confidence score is considered [42].

Furthermore, based on the performance metrics, the best model is identified by combining detection strength (mAP@0.5) with classification robustness

(Precision, Recall, and F1-score). This comprehensive evaluation framework ensures that both localization and classification capabilities are considered when selecting the most reliable model for clinical use. Precision, recall, F1-score, and mAP values are reported to capture model balance, while mAP values provide a benchmark for detection performance [43].

3.4.2 Evaluation metrics

Performance was evaluated using various metrics, including accuracy, recall, F1-score, precision, and mean Average Precision (mAP). These evaluation metrics are among the most used to assess the performance of trained models. These metrics are essential for evaluating object detection models, such as Faster-RCNN and YOLO variants, especially in medical imaging tasks where both false positives (FP) and false negatives (FN) can have significant consequences. Accuracy is excluded because it can be misleading in imbalanced datasets of polyps for detecting, where precision, recall, F1-score, and mAP provide more meaningful insights into model performance [44].

the confusion matrix plays a central role in evaluating the reliability of polyp-detection models. It breaks the model's predictions into True Positives, False Positives, True Negatives, and False Negatives, allowing clear identification of how each model succeeds or fails in detecting polyps. Because medical datasets are often imbalanced, these four values provide a more meaningful assessment than overall accuracy. Metrics derived from the confusion matrix—such as precision and recall—directly reflect how well a model avoids false alarms and prevents missed polyps. This detailed breakdown supports the comparison of Faster R-CNN, YOLOv5, YOLOv8, YOLOv9, YOLOv11, and YOLOv12 helping to determine which model is more dependable for real clinical use.

The confusion matrix represented in this research offers a detailed evaluation of the binary classification performance of the proposed module in

differentiating between polyps and background classes, while the high true positives TP count demonstrates the strong sensitivity of the model towards detecting poly regions, the elevated false positives FP rate, and lack of true negatives TN raise concerns regarding data set completeness and its specificity. These findings accentuate the need to improve dataset curation and balanced representation of both polyp and background classes to improve the model's real-world applicability and reduce diagnostic errors in clinical deployments [45].

The ratio of correctly predicted cases (polyps and non-polyps) to the total number of predictions will provide a global measure of detection performance, but is less informative if the dataset is imbalanced, and Accuracy is calculated as,

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

The portion of correctly predicted positives among all predicted positives (measures false positive control). Precision quantifies how many of the detected polyps are correct, and it is calculated as,

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

The actual positives correctly detected (measures false negative control). High recall is crucial since missed polyps (false negatives) can delay colorectal cancer diagnosis, and it is calculated as,

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

The harmonic mean of Precision and Recall, balancing false positives and false negatives, is both metrics, which are equally important. F1 Score can be calculated as,

$$\text{F1-Score} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The precision-recall (PR) curve is represented by a single scalar value between 0 and 1. At the same time, average precision is measured as AUC. It measures when maximum precision values are dropped, the curve is sampled at recall values (r_1, r_2). When $p(r_i)$ drops, it is sampled, and AP is computed as the

sum of rectangular blocks. This will help in calculating the mAP value. The value for AP can be defined as,

$$AP = \sum(r_{n+1} - r_n)P_{interp}(r_{n+1}) \quad (5)$$

The mean of Average Precision (mAP) across multiple Intersection-over-Union (IoU) thresholds and classes. It evaluates both detection accuracy and localization quality,

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (6)$$

Each model is trained and analysed with metrics such as the Precision-Recall Curve with mAP@0.5, the Precision-Confidence Curve, Recall-Confidence Curve, F1-Confidence Curve, Confusion Matrix, training and validation, loss, and accuracy plots across epochs at model comparison and results analysis.

3.4.2.1 Confidential Interval Analysis

In addition to hypothesis testing, the performance of the model was further interpreted using 95% Confidence Intervals (CI) for precision and recall. Confidence intervals provide a statistical range within which the true performance of the model is expected to fall, offering a clearer understanding of uncertainty than single-value metrics. Unlike the z-test—which determines whether a difference is statistically significant—CI focuses on quantifying the reliability of the observed precision and recall scores [49].

To compute the CI for each metric, the model outputs were treated as proportions derived from the confusion matrix. Precision represents the proportion of correctly identified positive cases among all predicted positive cases, while recall represents the proportion of correctly identified positive cases among the actual positive cases. For each proportion p , the standard error was calculated using:

$$SE = \sqrt{\frac{p(1-p)}{n}} \quad (7)$$

where n is the corresponding sample size (e.g., predicted positives for precision, actual positives for recall). The 95% CI was then constructed using the normal approximation:

$$CI = p \pm z \times SE \quad (8)$$

with $z = 1.96$ for a 95% confidence level. Thus, the Lower Bound and Upper Bound of the CI become:

$$CI_{lower} = p - 1.96 \times SE \quad (9)$$

$$CI_{upper} = p + 1.96 \times SE \quad (10)$$

These formulas allow the model's performance to be expressed as a statistically supported range instead of a single number, improving the accuracy and credibility of the analysis. Presenting these confidence intervals allows the thesis to demonstrate not only the point estimates of model performance but also the degree of certainty associated with these results. This strengthens the analysis by showing how stable and statistically reliable the model's precision and recall values are, which is especially important when evaluating medical image detection accuracy by calculating the precision value in Table 4.2.

3.4.2.2 Z-test Statistical Analysis

To statistically validate the differences in model performance, a two-proportion Z-test was conducted across all pairwise comparisons of the seven models (Faster R-CNN, YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, and YOLOv12). This test evaluates whether the observed differences in performance metrics, such as precision, recall, and accuracy, are statistically significant or could have occurred by chance. The null hypothesis H_0 assumes that the two models being compared achieve the same performance proportion, while the alternative hypothesis H_a suggests a significant difference between

them. The Z statistics were calculated using the formula from equation 12.

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (12)$$

\hat{p}_1, \hat{p}_2 are the proportion of correct detections for each model (e.g., precision, recall, or accuracy proxy). n_1, n_2 is the total number of samples for each model (e.g., TP+FN for recall, TP+FP for precision, and TP+FP+FN for accuracy proxy). The pooled proportion is derived with $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$; Null hypothesis H_0 : proportions are equal ($\hat{p}_1 = \hat{p}_2$), Alternative hypothesis H_a : proportions differ ($\hat{p}_1 \neq \hat{p}_2$), Significance level $\alpha = 0.05 \rightarrow critical\ Z = \pm 1.96$ the two-proportion Z-test demonstrated that YOLOv11 achieved statistically significant improvements over most other models, particularly in recall and accuracy, confirming it as the most reliable and effective model for polyp detection in this study[50].

Among these seven models, the two-proportion Z-test was initially applied to calculate and evaluate the differences in precision for better comparison. To classify positive samples correctly while minimizing false positives, Precision reflects the ability of a model. It is a crucial metric in clinical applications such as polyp detection. As denoted in Table 2, the test results indicate that, compared to the YOLO series, Faster R-CNN generally performed with lesser precision, although not all differences were statistically significant. In many cases, models like YOLOv8 and YOLOv9 executed higher precision values, but the differences between them were within the margin of statistical triviality, proposing comparable performance [51].

In divergence, YOLOv11 showed a conspicuous improvement, achieving statistically momentous precision gains over Faster R-CNN and YOLOv5. This denotes that YOLOv11 was more compatible in reducing false alarms during detection. the exhibited moderate precision levels are outperformed by YOLOv10 and YOLOv12, often surpassing Faster R-CNN but not always achieving crucial

improvements over other YOLO variants. These results authenticate that while precision gains were visible across models, only YOLOv11 steadily delivered statistically significant improvements. Overall, the precision-based Z-test highlights the gradual but significant headway in diminishing false positives as modern YOLO architectures were introduced.

Chapter 4 Experiments

4.1 Experimental setup

All the stated algorithms were implemented in the Google Colab platform with standard image size of 640x640 resolution each model is trained with 50 Epochs in batch size of 16 running with learning rate of 0.1 with decay to 0.01 and for Faster RCNN and YOLOv5 SGD(Stochastic gradient descent) optimizer and for YOLOv8-v12 Auto (AdamW) optimizer is used to control how model updates its weights during training as denoted in Table 5.1. Different optimizers affect speed, stability, and final accuracy [46].

The experiments in the study were conducted and run on a high-performance computer equipped with a 12th Gen Intel® Core™ i9-12900K processor (3.20 GHz), 32 GB of RAM, and a 64-bit Windows 11 Pro operating system. Providing the computational capacity necessary for evaluating and training deep learning models. The main objective was to detect and compare the performance of multiple object detection models, such as Faster RCNN, YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, and YOLOv12, on a curated polyp detection data set [47].

This data set was prepared using the RoboFlow platform, and each image is annotated with bounding boxes and augmented using standard techniques such as flipping, exposure, rotation, and saturation adjustments. The data set was exported in different compatible formats with each model's input requirements, ensuring its fair and consistent evaluation. Each model was trained and validated with identical data splits (training: 87.5%, validation: 8.2%, test: 4.3%) and evaluated using metrics such as loss, accuracy, and confusion matrices. The results include visual outputs for each model and performance comparisons, or you can use it to access trends and limitations of each model in the context of medical imaging analysis [48].

Table 4.1 Hyperparameters overview

Hyperparameters	Values
Platform	Google Colab
No. of epochs	50
Batch size	16
Learning rate	0.1 (with decay to 0.01)
Optimizer	YOLOv5: SGD, YOLOv8–12: Auto (AdamW)
Annotating & Preprocessing	Robo Flow
System	64-bit Windows 11 Pro operating system
Hardware	12 th Gen Intel® Core™ i9-12900K processor (3.20 GHz), 32 GB of RAM

4.2 Models for Detection

Seven models are trained in Google Colab platform with 50 epochs from the augmented dataset of polyp images. Each model trained individually and collected results of metrics, graphs and confusion matrices for comparison analysis and statistical analysis.

4.2.1 Faster RCNN

Faster RCNN trained and detected polyps in the actual way. The performance of Faster RCNN in the polyp detection task, as shown in Figure 4.4 the model correctly classified just over half of the total, with True positives of 93, false positives of 30, and FN is of 27, with zero background from polyp samples. In terms of detection quality, precision stands at 0.775~ 77% which means that the model predicts a polyp, and it is correct about ¾ of the time. Its recall is slightly lower at 0.756~ 75.6% showing that the model successfully detects most polyps but still misses some portion. The F1-score, which balances precision and

recall, a 0.765~76.5% reflecting the overall reliability of interesting polyps.

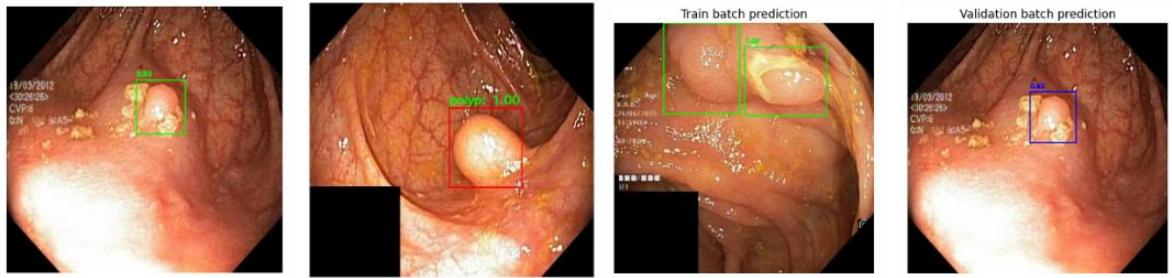


Figure 4.1 Detection of Polyps in Faster RCNN Model

AP@0.5 (Precision) vs Epoch varies between 0.76 and 0.82, with frequent oscillations. Precision is relatively high but inconsistent, meaning the model often predicts correctly when it does detect, but stability is an issue. Recall vs Epoch graph fluctuates between 0.53 and 0.59, showing low and unstable recall. The model misses many objects, which is a common issue with Faster R-CNN on smaller datasets or without proper tuning. Threshold-based metrics with recall vs threshold start near 1.0 at low thresholds, drop sharply after 0.5, and approach 0 near threshold 1.0. It indicates that increasing the confidence threshold significantly reduces recall.

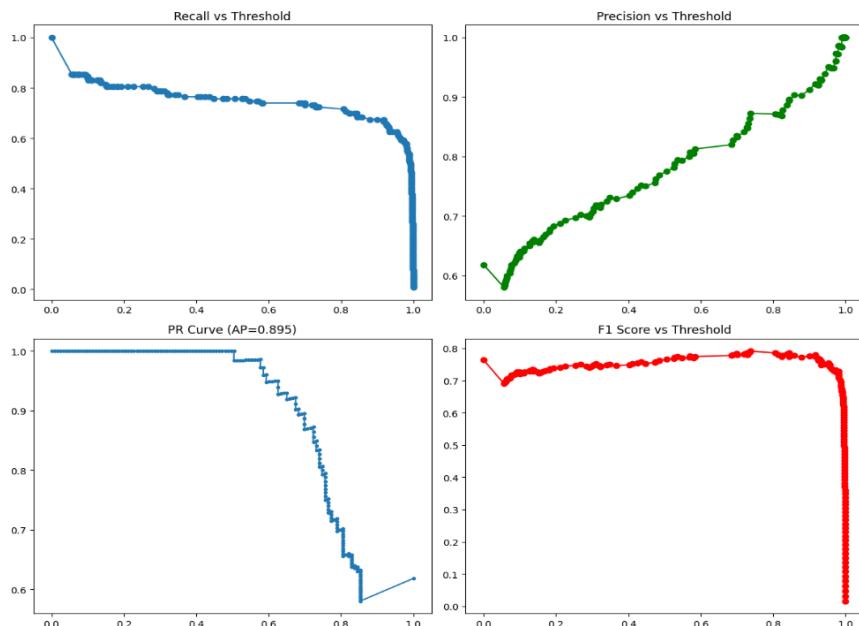


Figure 4.2 Faster RCNN Model performance summary with (a) Recall vs Threshold Curve, (b) Precision vs Threshold Curve, (c) PR Curve, (d) F1 score vs Threshold Curve.

Typical precision-recall higher thresholds improve precision but hurt recall, indicating the model is conservative: good precision at low recall, but struggles to maintain precision as recall improves, with F1 Score vs Threshold Peaks around 0.8 at a threshold near 0.5, then declines. Precision vs. Threshold Starts low (~ 0.2) at low thresholds, rises steadily to ~ 1.0 at high thresholds. Suggests the optimal confidence threshold for balanced precision and recall is around 0.5. $mAP@[0.5:0.95] \approx 0.48\text{--}0.51$. This performs Lower than YOLO models (which achieved 0.6–0.68). Precision is decent (~ 0.8), but recall is weak (~ 0.55). The model misses many objects. PR Curve (AP@0.5) Precision starts near 1.0 at low recall, then drops steeply as recall increases.

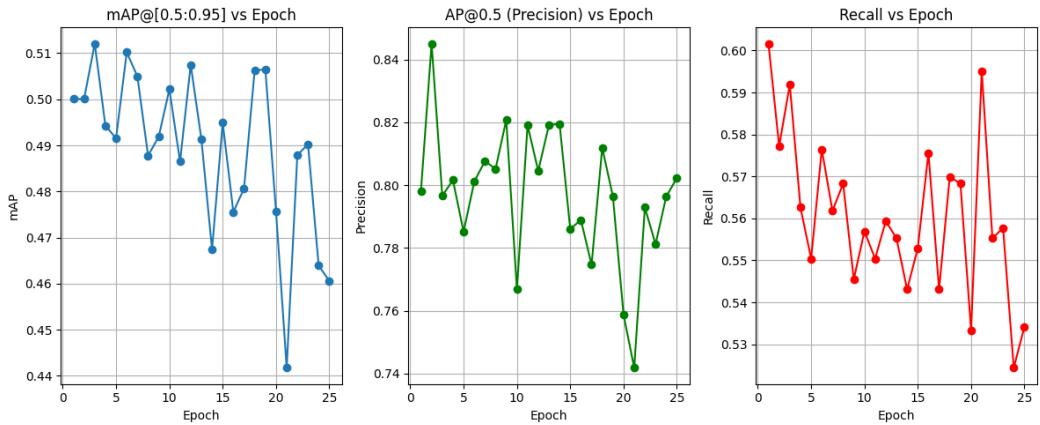


Figure 4.3 (a)mAP[0.5:0.95] vs Epoch curve, (b) AP 0.5(Precision) vs Epoch curve, (c) Recall vs Epoch curve trends across epochs.

Training is unstable in metrics that fluctuate heavily, indicating possible issues like the learning rate not being well-tuned, from Figure 4.3. Faster R-CNN here shows moderate precision but poor recall and unstable training, leading to lower mAP compared to YOLOv8–YOLOv12. It might still be useful for applications where precision matters more than recall, but for real-time or high-recall tasks, YOLO variants are superior. Finally, the main average precision (mAP) is 0.840 or 84% which is consistent with other metrics and suggests stable performance across detection thresholds.

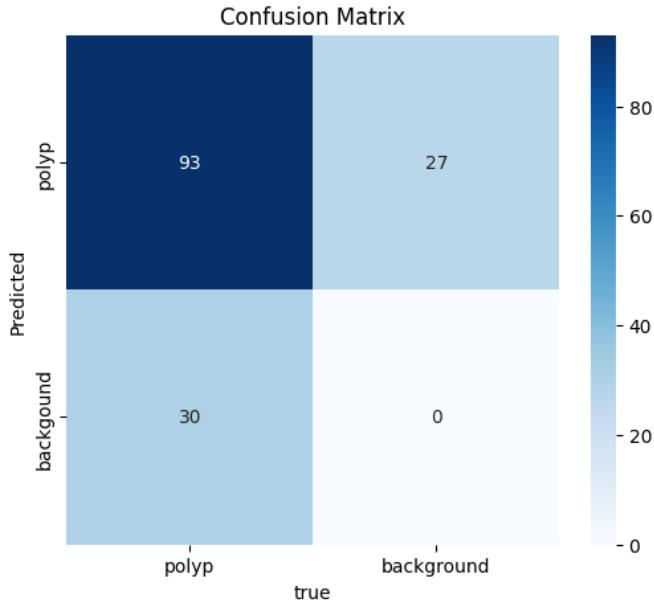


Figure 4.4 Confusion Matrix of Faster RCNN

Faster R-CNN performs reasonably well in identifying polyps but struggles with distinguishing background, as evidenced by the absence of true negatives. This limitation could be addressed through improved dataset balancing, enhanced feature extraction, or post-processing techniques.

4.2.2 YOLOv5

The YOLOv5 model for polyp detection shows reliable results. The confusion matrix, 185 true positives, 41 false positives, and 50 false negatives, reveals that the model correctly identifies most polyps but still misses a portion and occasionally misclassifies background regions as polyps denoted its performance with its visual representation of the precision-confidence curve 0.931, the Recall-confidence curve 0.94, the F1-confidence curve 0.82 at 0.292, and the Precision-Recall curve 0.94 at mAP@0.5.

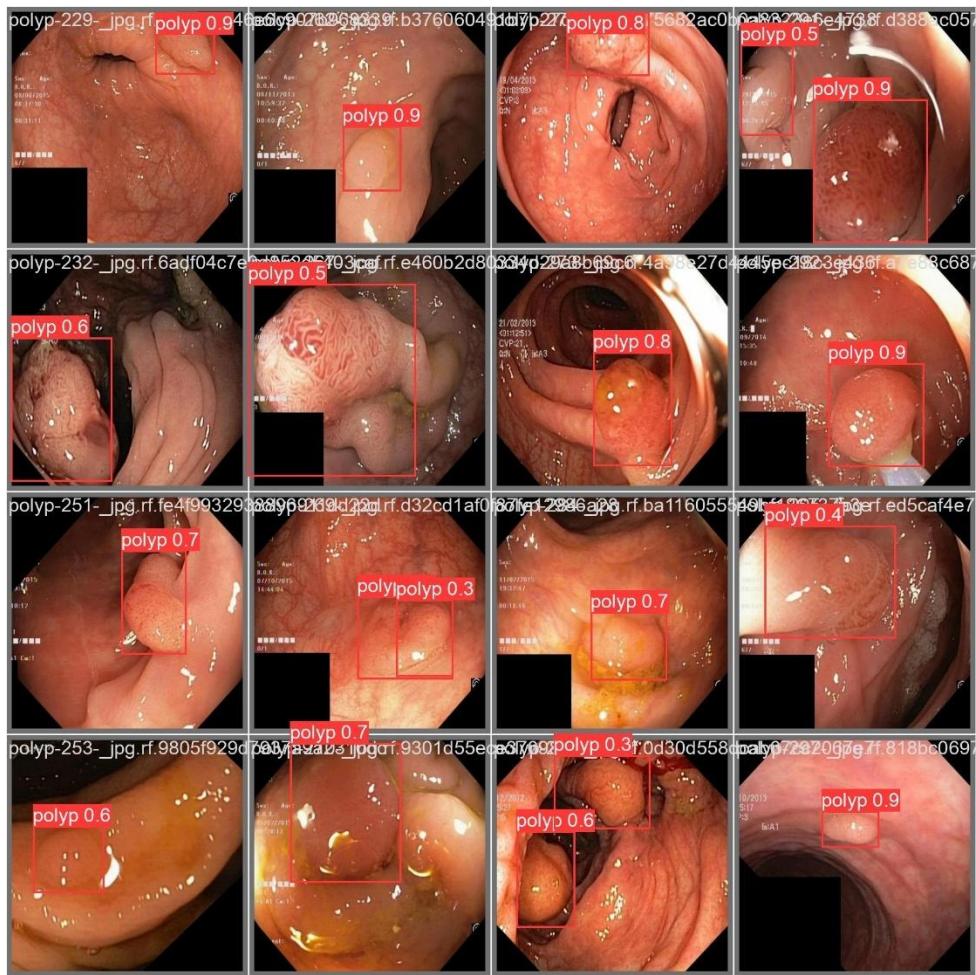


Figure 4.5 detected images of a polyp in yolov5

Epoch-based Metrics with mAP@ [0.5:0.95] vs Epoch Starts around 0.50, fluctuates between 0.44 and 0.51 across the first 25 epochs. Indicates unstable convergence and no clear upward trend, suggesting the model struggles to improve consistently. AP@0.5 (Precision) vs Epoch varies between 0.76 and 0.82, with frequent oscillations. Precision is relatively high but inconsistent, meaning the model often predicts correctly when it does detect, but stability is an issue. Recall fluctuates between 0.53 and 0.59, indicating low and unstable recall.

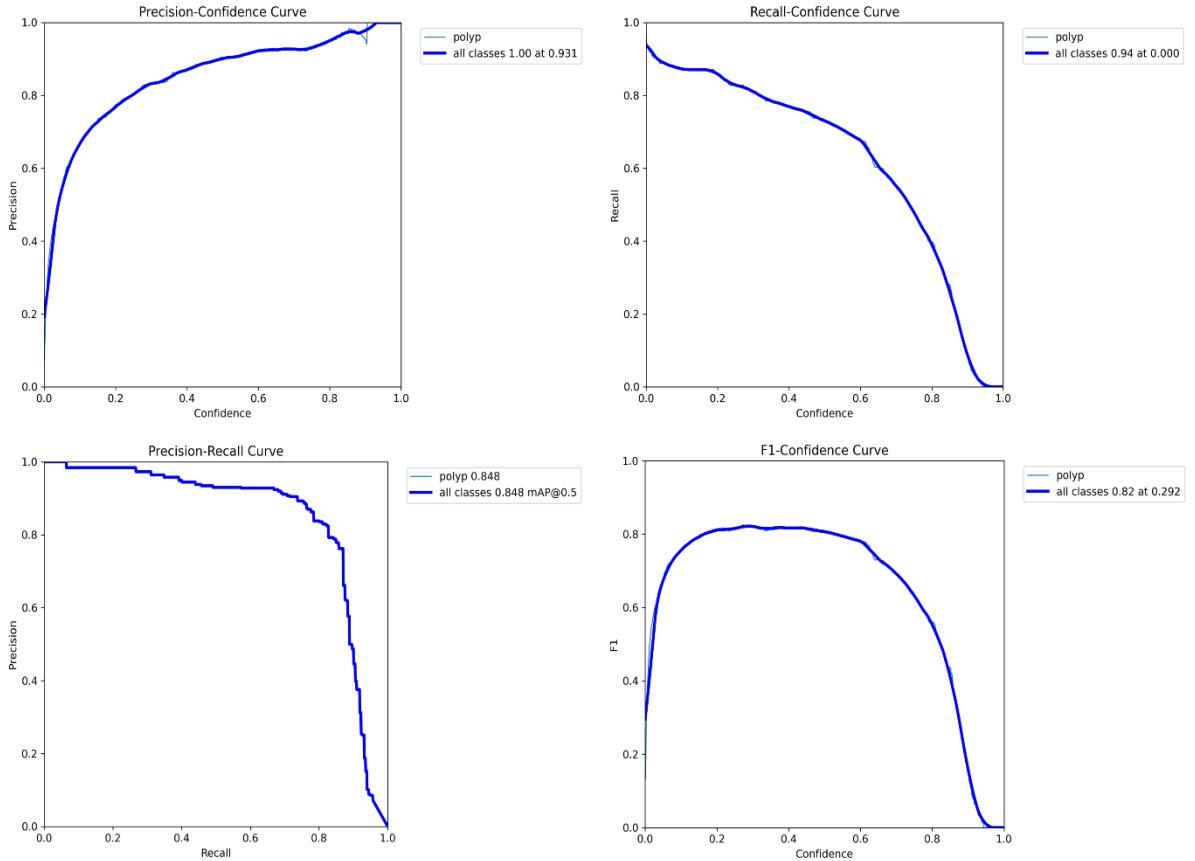


Figure 4.6 YOLOv5 model performance visualization of Precision-Confidence Curve, Recall-Confidence Curve, Precision-Recall Curve, F1-Confidence Curve.

The Precision–Confidence curve illustrates how the YOLOv5 model maintains precision across different confidence thresholds. As seen in the plot, the precision starts lower at minimal confidence but gradually increases as the confidence threshold rises. Near a confidence value of 0.9, the precision approaches close to 1.0, indicating that the model makes highly accurate predictions when it is more certain. This trend suggests that YOLOv5 effectively distinguishes between true and false detections, minimizing false positives at higher confidence levels.

The Recall–Confidence curve shows how recall varies with the model’s confidence threshold. In this graph, recall begins around 0.94 at a low threshold and slowly declines as confidence increases. This pattern reflects the typical trade-off in object detection: increasing confidence reduces false positives but

may also miss some true positives. YOLOv5 achieves a recall of approximately 0.82 at a moderate threshold (0.29), indicating strong sensitivity in identifying most polyp regions within the dataset.

The Precision–Recall curve provides an integrated view of both precision and recall. For YOLOv5, this curve exhibits a steady precision across high recall values before tapering off near the extreme end. The area under the curve (AUC) is high, with an mAP@0.5 of 0.848, demonstrating robust detection performance. This high mean Average Precision score confirms that the model balances precision and recall effectively, ensuring that most predicted polyps are accurate and few actual polyps are missed.

The F1–Confidence curve represents the harmonic mean between precision and recall, showing the overall balance between accuracy and sensitivity. The YOLOv5 model achieves its peak F1-score of 0.82 at a confidence level of 0.29, where both precision and recall are optimally balanced. Beyond this point, the F1-score gradually declines as either false positives or false negatives increase. This indicates that a confidence threshold around 0.3 is ideal for maintaining the best trade-off between detection completeness and precision.

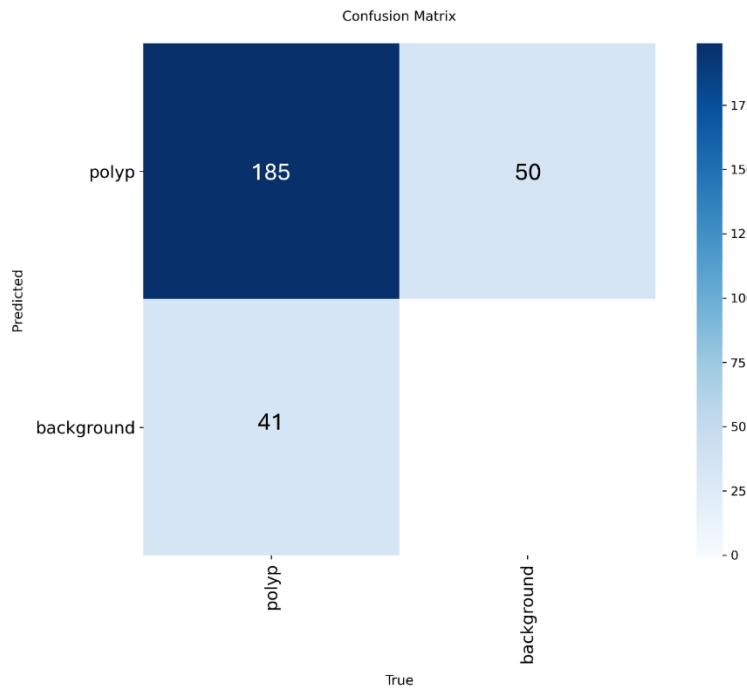


Figure 4.7 Confusion Matrix of YOLOv5

The confusion matrix provides a visual summary of the model's classification performance. YOLOv5 correctly identifies 185 true positives (polyp detections) while producing 50 false positives (background detected as polyp) and 41 false negatives (missed polyps). The strong diagonal dominance in the matrix confirms that the model performs well overall. Although some false detections occur, they are relatively minor compared to the total correct predictions, reflecting the model's reliability in differentiating polyp from background.

The precision of 78.7% indicates that nearly four out of five detected polyps are true, while the recall of 81.8% shows that the model successfully identifies the majority of actual polyps. The F1 score of 0.80 highlights a good balance between precision and recall. Training Losses in Top row train/box_loss starts around 0.08, decreases steadily to about 0.02 by epoch 50. Indicates strong improvement in bounding box regression. Train/obj_loss starts near 0.022, drops gradually to 0.01. Objectness confidence improves consistently.

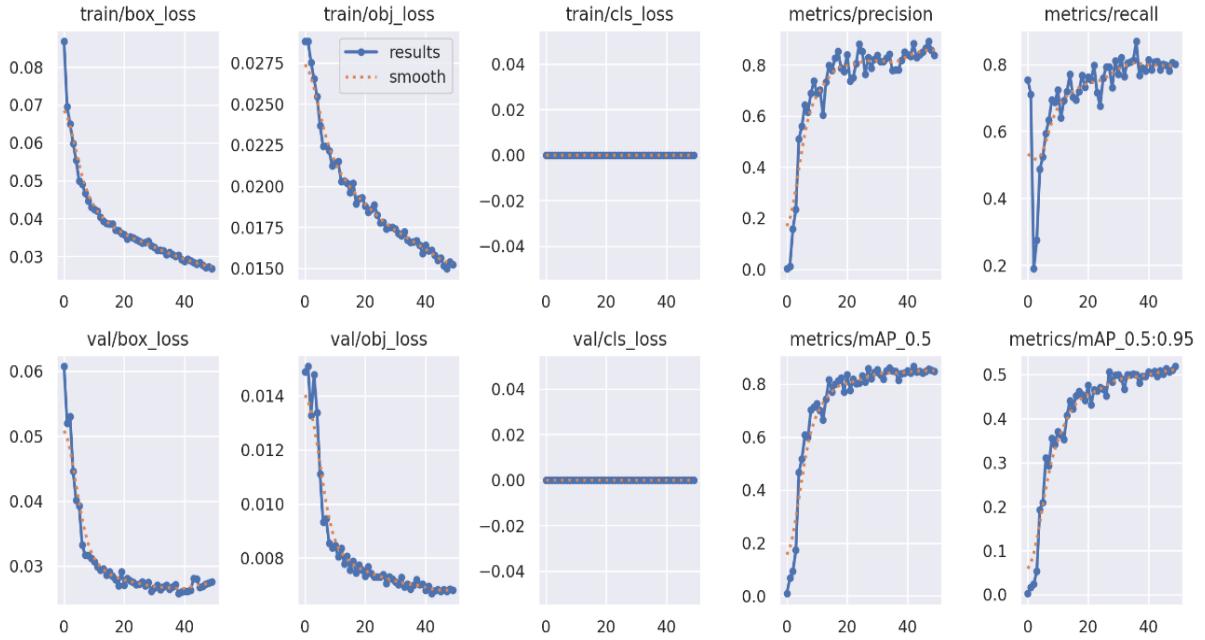


Figure 4.8 Training and validation metric curves of loss and accuracy plots across epochs, YOLOv5.

Train/cls_loss Flat at 0.0 throughout training. This suggests the dataset likely has a single class, so classification loss is not used. Validation Losses (Bottom Row) val/box_loss starts around 0.06, drops to about 0.03, and stabilizes. Smooth convergence, no spikes. Val/obj_loss starts near 0.02, decreases to 0.01. indicates good generalization for objectness. Val/cls_loss flat at 0.0, confirming single-class detection in performance metrics, metrics/precision rises quickly to ~0.95, very high precision. Few false positives. Metrics/recall Improves to ~0.9, strong recall.

YOLOv5 gives 0.0 classification loss in graphs because your dataset is a single-class dataset. In a single-class problem, YOLOv5 does not compute a true classification probability distribution like multi-class detection.

A few missed detections. metrics/mAP50 'epochs ~0.95, excellent detection at IoU 0.5. metrics/mAP50-95 Improves to ~0.6, solid performance across stricter IoU thresholds. Indicates no major fluctuations, smooth convergence. Training is very stable. Loss values are very low, typical for YOLOv5. Precision and recall are excellent, making this model highly reliable. mAP50 is extremely high

(~ 0.95), and mAP50-95 (~ 0.6) is competitive with later YOLO versions.

Although the accuracy is 72.8%, this is impacted by the absence of true negatives in the dataset. The mAP@0.5 of 0.848 demonstrates high detection capability, shown in figure 4.8 confirms consistent performance across IoU thresholds with P-curve, training and validation losses converged smoothly, reflecting stable learning. YOLOv5 demonstrates strong and dependable detection performance for medical image analysis of polyps.

4.2.3 YOLOv8

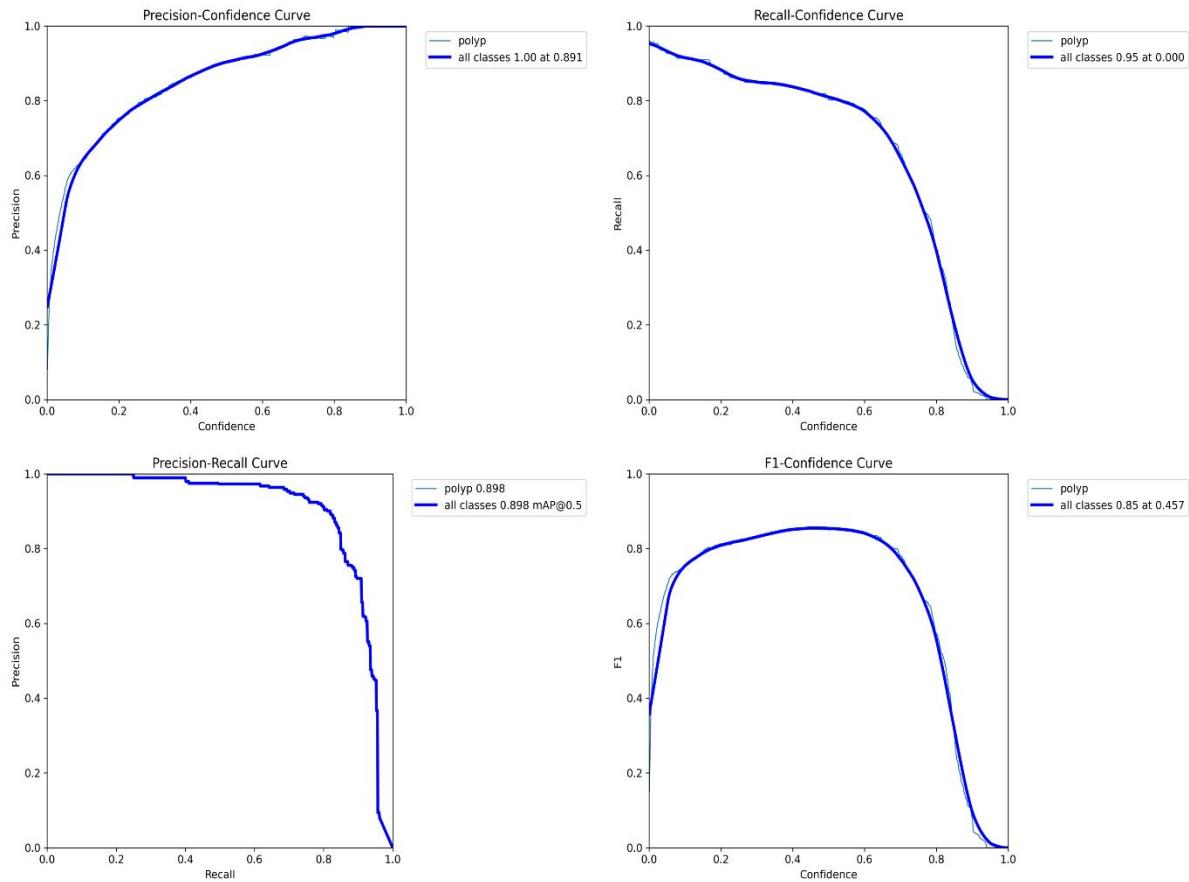


Figure 4.9 YOLOv8 model performance visualization of Precision-Confidence Curve, Recall-Confidence Curve, Precision-Recall Curve, F1-Confidence Curve.

The Precision–Confidence curve for YOLOv8 shows a consistent upward trend as the confidence threshold increases. At lower confidence levels, the model produces more false positives, which lowers precision. However, as confidence rises, YOLOv8 becomes more selective, leading to higher precision values and

fewer incorrect detections. The curve eventually reaches a near-perfect precision of 1.00 at 0.891 confidence, reflecting the model’s strong ability to confirm true detections. This pattern indicates robust discrimination between polyp and background regions. The steady growth in precision highlights efficient learning and reduced classification ambiguity. Overall, the precision curve confirms YOLOv8 reliability in high-confidence predictions.

The Recall–Confidence curve demonstrates how the model’s sensitivity changes with different confidence thresholds. At low thresholds, recall remains very high (around 0.95) because most objects, including uncertain detections, are classified as positives. As the threshold increases, recall begins to decrease slightly since fewer detections are accepted. This trade-off between confidence and recall is expected in object detection models. The gradual downward trend indicates stable and predictable model behavior. Even at higher confidence levels, recall values stay consistently strong, showing YOLOv8’s robustness in detecting small and complex polyps. Thus, the model effectively minimizes missed detections across varying threshold conditions.

The Precision–Recall curve presents a smooth and high-arching shape near the top-right corner, indicating an excellent balance between precision and recall. The mean Average Precision (mAP@0.5) is recorded as 0.898, confirming the model’s strong detection capability. A high mAP value suggests that YOLOv8 can maintain high precision without significantly compromising recall. The curve’s consistent elevation shows that the model handles diverse samples and background complexities effectively. Its stable shape also proves efficient bounding box localization. YOLOv8 therefore demonstrates remarkable consistency across the detection spectrum. Overall, this curve signifies excellent generalization and prediction reliability for polyp detection.

The F1–Confidence curve represents the harmonic mean of precision and recall across confidence thresholds. The highest F1 score of 0.85 at a 0.457 threshold indicates the model’s optimal balance point for real-world application. Beyond this threshold, precision improves but recall decreases slightly, resulting in a gentle fall in F1. The symmetric bell-shaped curve shows YOLOv8’s consistent trade-off management. It also reflects the model’s adaptability to varying decision thresholds during inference. The high F1 region suggests robustness in both detection accuracy and completeness. Therefore, YOLOv8 achieves strong equilibrium between precision and recall for accurate polyp localization.

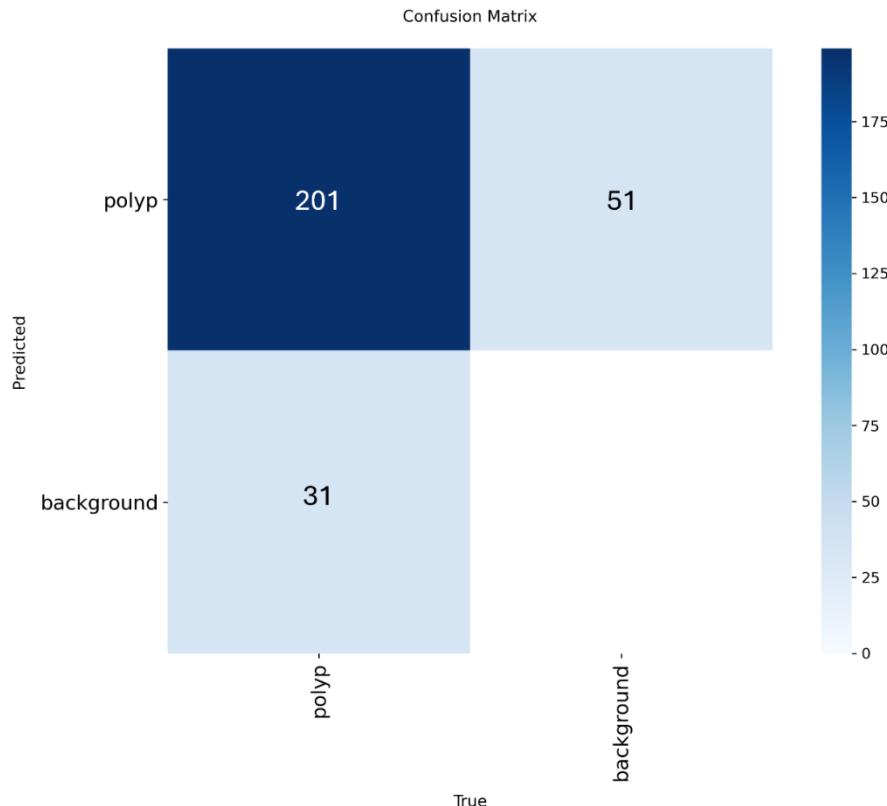


Figure 4.10 Confusion Matrix of YOLOv8

The confusion matrix summarizes the classification outcomes of the model by displaying true and false predictions. YOLOv8 achieved a large number of true positives (201), while false positives and false negatives remained low. This distribution indicates that the model correctly detects the majority of polyps with minimal errors. The visual separation between true and background samples

demonstrates clear decision boundaries. Such high TP values confirm the model's excellent learning of spatial and visual features. The limited false detections validate high sensitivity and specificity.

Consequently, the confusion matrix highlights YOLOv8's reliable performance in distinguishing polyp regions from non-polyp areas.

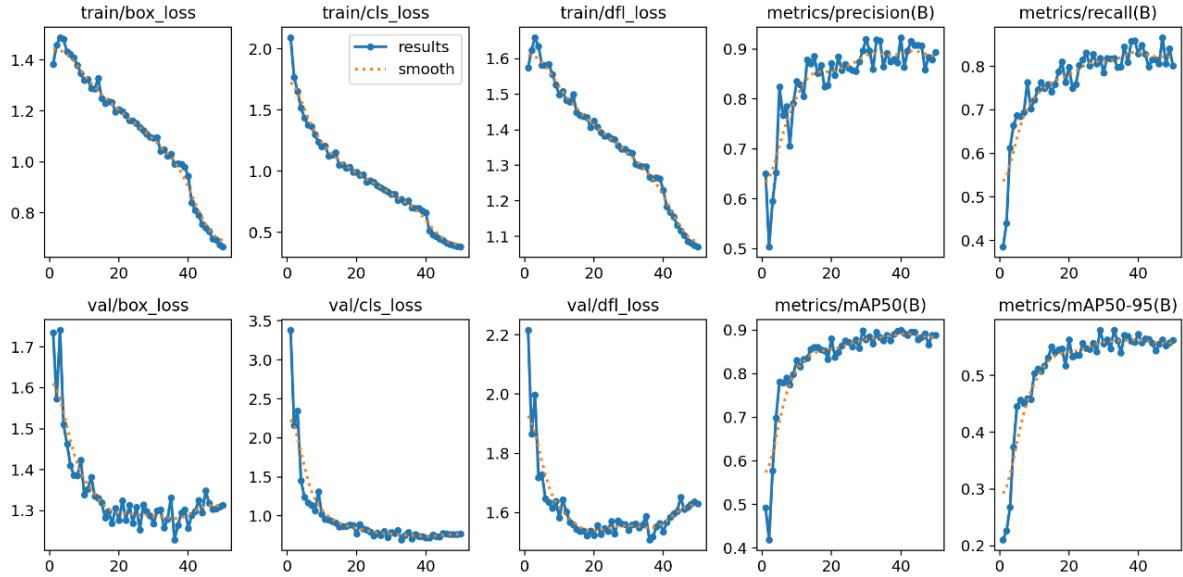


Figure 4.11 training and validation metrics of loss and accuracy plots across epochs YOLOv8

Train/box_loss starting around 1.4, decreasing steadily to about 0.7. This indicates that bounding box regression is improving consistently. Train/cls_loss Starts near 2.0, drops sharply in the first 10 epochs, and ends around 0.4. train/dfl_loss starts at 1.6, decreases gradually to 1.0. This shows the performance is lower than YOLOv9's initial spike, indicating that more stable early validation suggests better localization precision over time.

Val/box_loss starts around 1.7, drops to about 1.3, and stabilizes. This proves that accuracy improves quickly and stabilizes. Val/cls_loss starts near 3.5, then drops below 1.0 by epoch 20, which denotes good generalization on classification. Val/dfl_loss starts around 2.2, decreases to 1.6 denotes no extreme spike like YOLOv9, suggesting smoother convergence.

The Performance metrics precision(B) rises quickly to ~0.9, slightly higher than YOLOv9, rises quickly to ~0.9, slightly higher than YOLOv9 it indicates fewer false positives. Metrics/recall(B) this improves to ~0.8, like YOLOv9. Gives balanced detection performance.

Metrics/mAP50(B) reaches ~0.85, comparable to YOLOv9 Strong detection at IoU 0.5, metrics/mAP50-95(B) Improves to ~0.55. shown in 5.7.

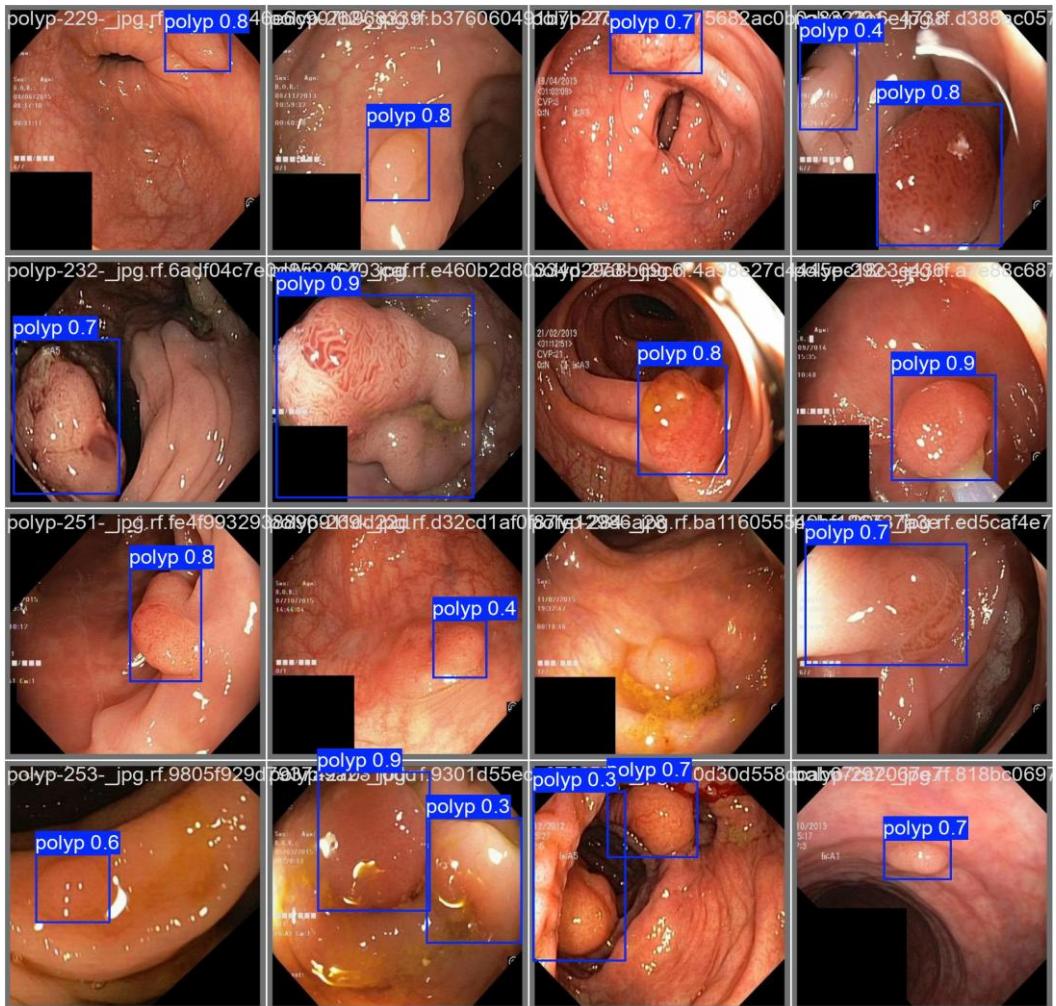


Figure 4.12 YOLOv8-based polyp detection on colonoscopy images

The YOLOv8 model achieved strong detection performance for polyps. The confusion matrix shows 201 true positives, 31 false positives, and 51 false negatives, as P-Curve 0.891, R-Curve 0.95, PR curve 0.891, F1-curve 0.85, training and validation, loss, and accuracy plots across epochs, with train/box_loss starting around 1.4, decreasing steadily to about 0.7. This indicates that bounding box regression is improving consistently. Train/cls_loss Starts near

2.0, drops sharply in the first 10 epochs, and ends around 0.4. `train/dfl_loss` starts at 1.6, decreases gradually to 1.0. This shows the performance is lower than YOLOv9’s initial spike, indicating that more stable early validation suggests better localization precision over time.

The Recall was 0.798, or 79%, the precision was 0.866, or 86%, the F1-score was 0.831, or 83%, and the mAP was 0.898, or 89%, which demonstrates a good balance between sensitivity and accuracy, indicating reliable classification. The PR curve reached an AP of 0.898 at IoU 0.5, confirming consistent detection. The training and validation losses decreased steadily, proving effective learning and good generalization. The F1, precision-confidence, and recall-confidence curves all peaked at high values, showing that YOLOv8 maintains both high recall and precision across thresholds. These values confirm that YOLOv8 outperformed earlier models such as Faster R-CNN and YOLOv5, comparable with more recent YOLO variants.

4.2.4 YOLOv9

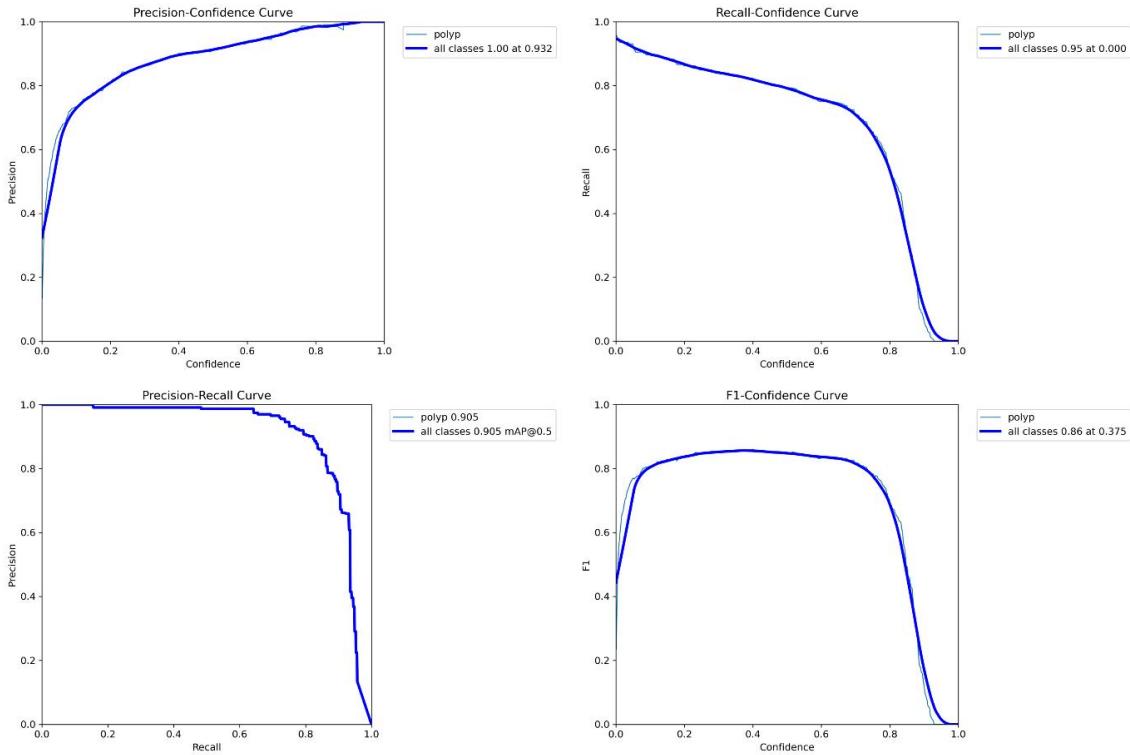


Figure 4.13 YOLOv9 model performance visualization of Precision-Confidence Curve, Recall-Confidence Curve, Precision-Recall Curve, F1-Confidence Curve.

The Precision–Confidence curve for YOLOv9 demonstrates a progressive increase in precision as confidence rises, peaking at 1.00 precision when the confidence threshold reaches 0.932. At lower confidence values, the model accepts more uncertain detections, slightly reducing precision. However, as the threshold grows, false positives are minimized, and detection accuracy significantly improves. This consistent upward trend confirms YOLOv9’s ability to effectively distinguish true polyps from background noise. The curve’s steep incline indicates strong classification reliability. Overall, this precision pattern showcases YOLOv9’s superior accuracy and confidence-based filtering in real-world detection tasks.

The Recall–Confidence curve highlights how YOLOv9 maintains high sensitivity across various confidence levels. Initially, recall remains very high at around 0.95, meaning most polyps are detected even with lower confidence thresholds. As the threshold increases, recall gradually decreases, indicating a balanced trade-off between missed and detected samples. The smooth downward slope signifies stable model behavior with controlled detection sensitivity. Despite this decline, recall values remain strong, confirming the model’s consistent detection of subtle and small features. This trend validates YOLOv9’s robustness and adaptability to different threshold settings during inference.

The Precision–Recall curve for YOLOv9 maintains a near-ideal shape, concentrated toward the upper-right region of the graph. The mean Average Precision (mAP@0.5) reaches 0.905, indicating excellent detection and classification capability. This high mAP reflects the model’s efficiency in identifying true positives while keeping false alarms minimal. The smoothness of the curve shows that precision remains stable even when recall is high. YOLOv9 effectively manages complex polyp boundaries and background variations. The uniform distribution of precision across recall levels further confirms its reliability. Therefore, this curve strongly demonstrates the model’s excellent detection consistency and precision balance.

The F1–Confidence curve shows the balance between precision and recall as confidence thresholds vary. The highest F1 score achieved is 0.86 at 0.375 confidence, representing the optimal trade-off between accuracy and sensitivity. The curve forms a bell-shaped distribution, showing YOLOv9’s efficiency across varying thresholds. At this optimal point, both false positives and false negatives are minimized. Beyond the threshold, the F1 value gradually declines as recall decreases. This pattern emphasizes the model’s stability and ideal decision boundary. Thus, the F1 curve reflects YOLOv9’s well-tuned balance between precision and recall for practical deployment.

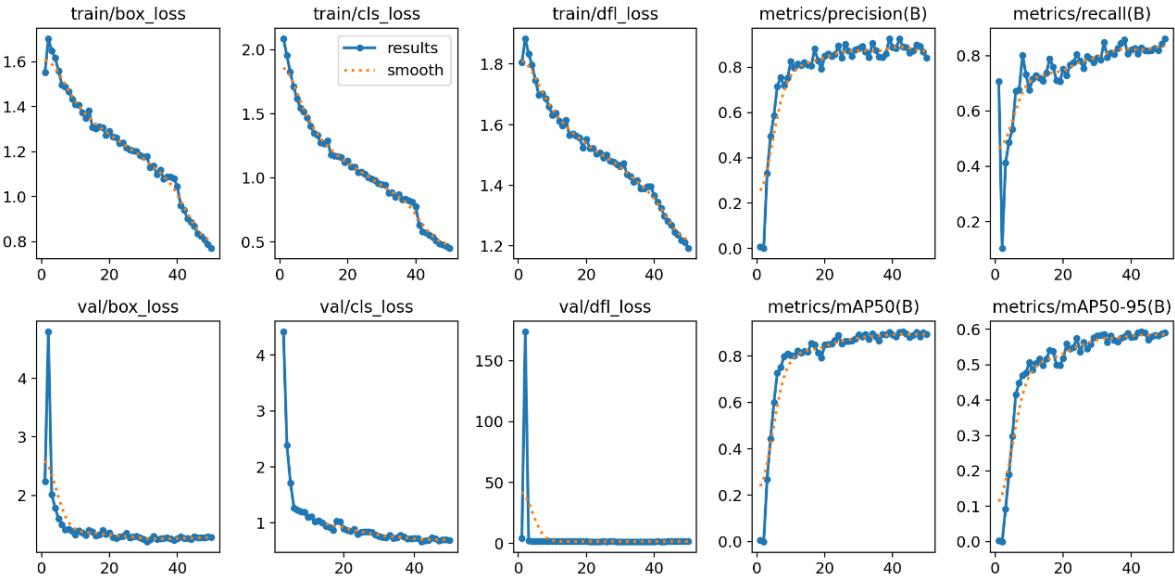


Figure 4.14 Training and validation metrics of loss and accuracy plots across epochs YOLOv9

Figure 4.14 denoted its performance with its visual representation of the precision-confidence curve 0.932, the Recall-confidence curve 0.95, the F1-confidence curve 0.86 at 0.375, and the Precision-Recall curve 0.905 at mAP @0.5, whereas mage shows YOLOv8 training and validation metrics over 50 epochs with Training Losses (Top Row) includes train/box_loss with Starts around 1.6 and steadily decreases to about 0.8 by epoch 50 this Indicates the model is improving its bounding box regression accuracy. Train/cls_loss it starts near 2.0, drops sharply after ~10 epochs, and ends around 0.5.

This shows that classification accuracy improves significantly during early training. Train/dfl_loss starts at 1.8, decreases gradually to 1.2, and DFL (Distribution Focal Loss) is stabilizing, meaning better localization precision. Validation Losses (Bottom Row, left), val/box_loss which Initially spikes above 4.0, then quickly drops below 1.0, and val/cls_loss and stabilizes with similar pattern starts high (~3.0), then drops below 1.0, this indicates good generalization on classification, val/dfl_loss with huge initial spike (~150), then rapidly converges near 1.0 early instability but then stabilizes well with performance metrics on right columns whereas metrics/precision(B).

Starts near 0.0, rises quickly to ~0.85 by epoch 10, then stabilizes this leads to High precision means fewer false positives. This states that the combination of high precision and recall with excellent mAP scores makes YOLOv9 a strong candidate for medical detection tasks, though reducing false negatives further would improve clinical applicability.

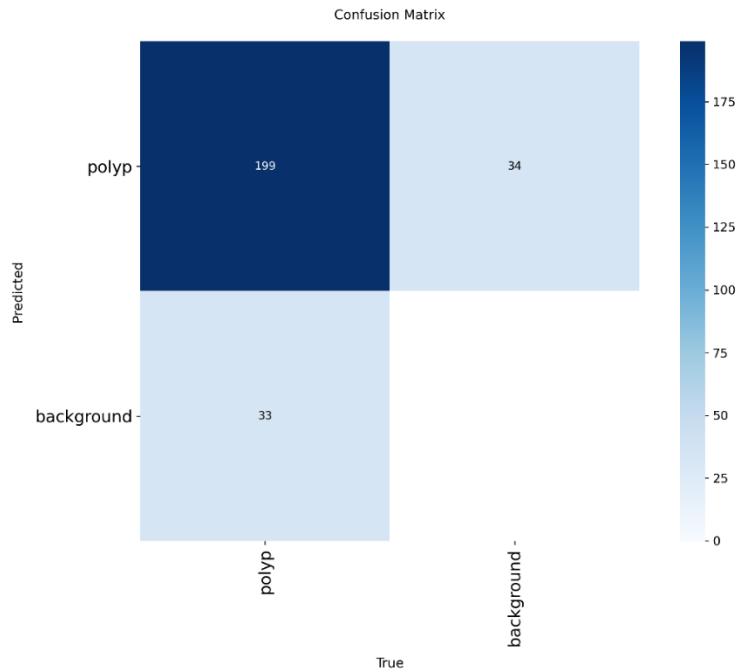


Figure 4.15 Confusion Matrix of YOLOv9

The confusion matrix visually summarizes YOLOv9's classification performance. The model achieved 199 true positives, with 34 false positives and 33 false negatives, highlighting strong detection accuracy. The high concentration of values along the diagonal indicates precise predictions and minimal misclassification. YOLOv9 shows effective discrimination between polyp and background categories. The limited error distribution proves efficient training and good generalization. These results confirm strong sensitivity and specificity levels in polyp detection. Consequently, the confusion matrix affirms that YOLOv9 provides reliable and consistent detection outcomes with minimal uncertainty.

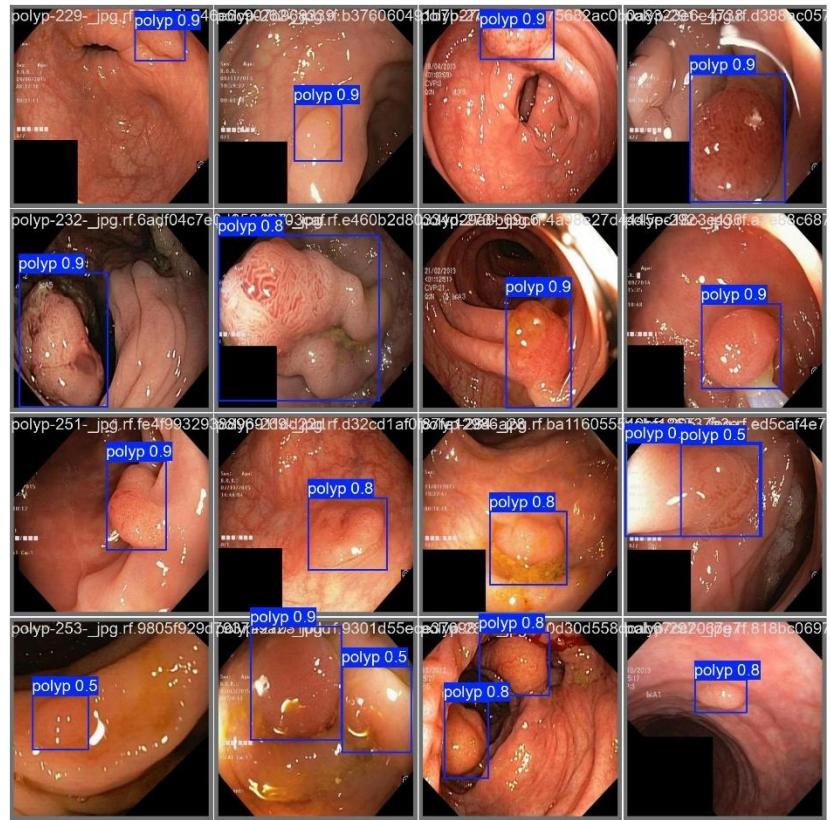


Figure 4.16 YOLOv9-based polyp detection on colonoscopy images.

Figure 4.15 denoted its performance with its visual representation of the precision-confidence curve 0.932, the Recall-confidence curve 0.95, the F1-confidence curve 0.86 at 0.375, and the Precision-Recall curve 0.905 at mAP @0.5, whereas mage shows YOLOv8 training and validation metrics over 50 epochs with Training Losses (Top Row) includes train/box_loss with Starts around 1.6 and steadily decreases to about 0.8 by epoch 50 this Indicates the model is improving its bounding box regression accuracy. Train/cls_loss it starts near 2.0, drops sharply after ~10 epochs, and ends around 0.5.

This shows that classification accuracy improves significantly during early training. Train/dfl_loss starts at 1.8, decreases gradually to 1.2, and DFL (Distribution Focal Loss) is stabilizing, meaning better localization precision. Validation Losses (Bottom Row, left), val/box_loss which Initially spikes above 4.0, then quickly drops below 1.0, and val/cls_loss and stabilizes with similar pattern starts high (~3.0), then drops below 1.0,

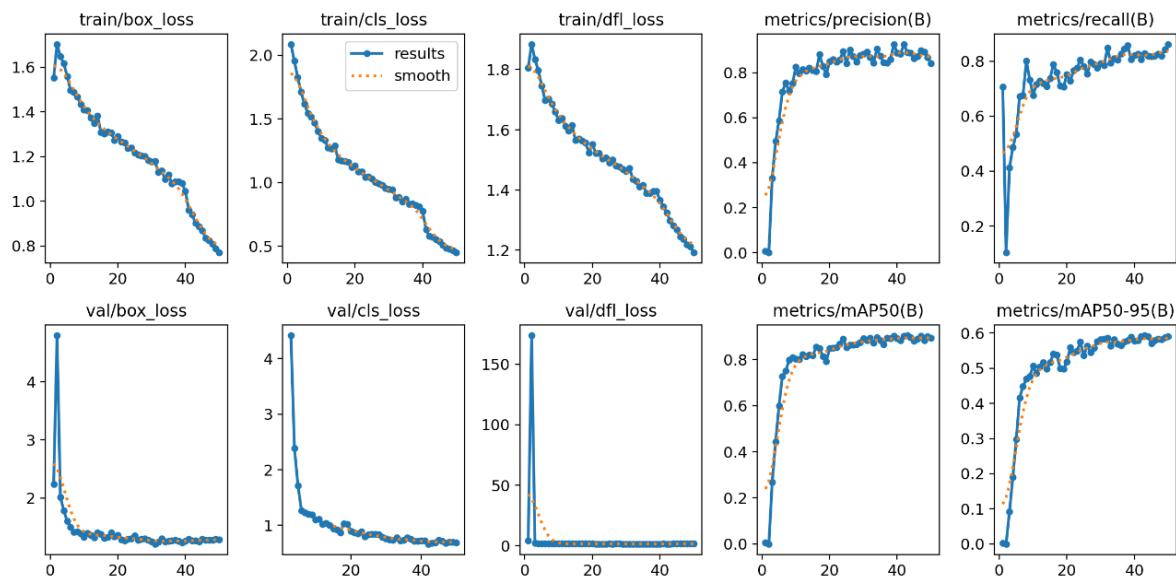


Figure 4.17 training and validation metrics of loss and accuracy plots across epochs YOLOv9

As shown in figure 4.17 this indicates good generalization on classification, val/dfl_loss with huge initial spike (~150), then rapidly converges near 1.0 early instability but then stabilizes well with performance metrics on right columns whereas metrics/precision(B) Starts near 0.0, rises quickly to ~0.85 by epoch 10, then stabilizes this leads to High precision means fewer false positives. This states that the combination of high precision and recall with excellent mAP scores makes YOLOv9 a strong candidate for medical detection tasks, though reducing false negatives further would improve clinical applicability.

4.2.5 YOLOv10

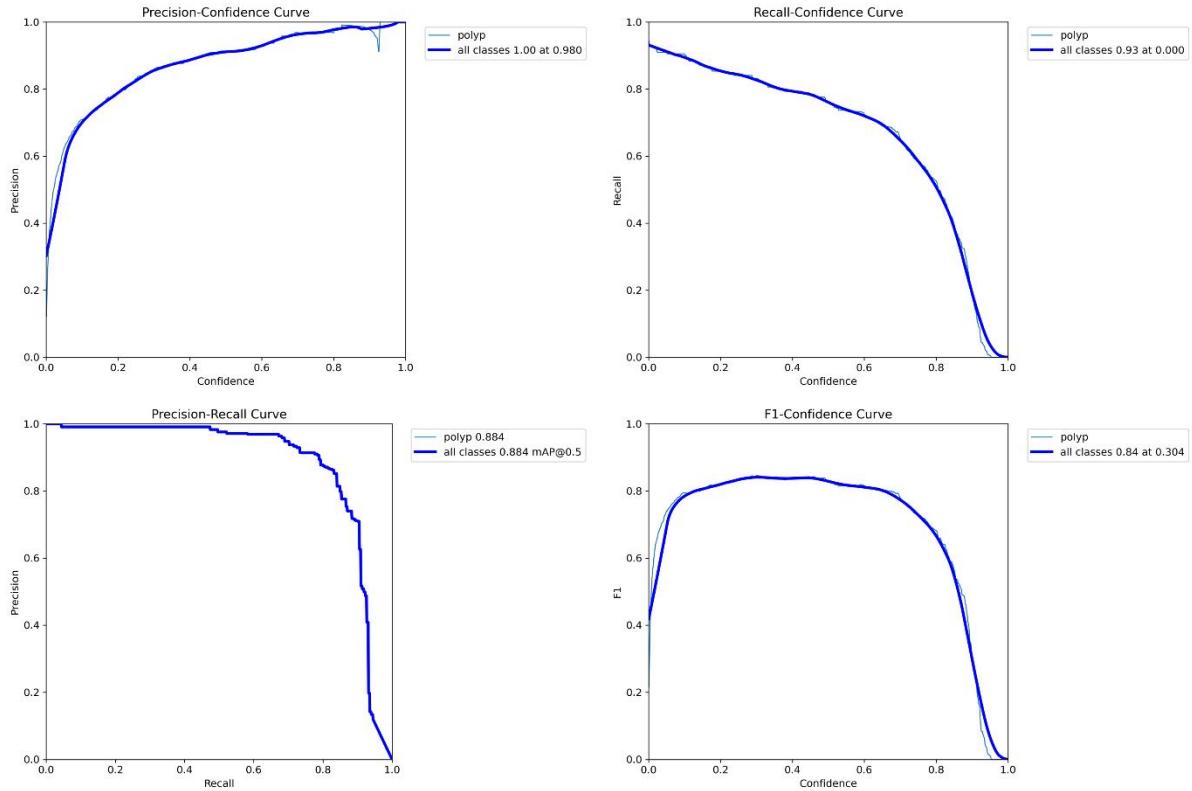


Figure 4.18 YOLOv10 model performance visualization of Precision-Confidence Curve, Recall-Confidence Curve, Precision-Recall Curve, F1-Confidence Curve

The Precision–Confidence curve for YOLOv10 illustrates how precision changes with varying confidence thresholds. The graph shows a steady rise in precision, reaching a perfect 1.00 at a confidence of 0.980, which signifies the model’s strong accuracy at higher thresholds. At lower confidence levels, precision values are moderate, suggesting that some false positives are included when detections are less certain. As confidence increases, incorrect detections sharply decrease, improving reliability. The nearly flat upper region reflects YOLOv10’s stability in identifying true positives. This consistent precision performance demonstrates the model’s effectiveness in distinguishing polyps from non-target regions. Overall, YOLOv10 achieves highly confident and accurate detections.

The Recall–Confidence curve demonstrates YOLOv10’s detection

sensitivity across different confidence values. The recall starts high at 0.93, indicating that the model successfully detects the majority of polyps even at lower thresholds. As the confidence value rises, recall gradually declines, showing that the model becomes stricter and may miss a few low-confidence detections. The smooth, consistent downward slope highlights balanced behavior without abrupt performance drops. This pattern confirms YOLOv10’s capability to maintain strong recall while controlling false negatives. The curve’s stability suggests good generalization during evaluation. Hence, YOLOv10 efficiently preserves high detection sensitivity throughout various confidence levels.

The Precision–Recall curve reveals the overall detection trade-off for YOLOv10. The curve maintains a high precision level across most recall values, achieving an impressive mAP@0.5 of 0.884. This indicates that YOLOv10 consistently identifies true positives while keeping false detections minimal. The graph’s shape closely follows an ideal pattern with gradual decline near higher recall regions, confirming robust feature learning. Such a result demonstrates effective localization and classification of polyp instances. The balance between precision and recall ensures reliable object identification under varying conditions. Therefore, the curve establishes YOLOv10 as a highly precise and dependable detection framework.

The F1–Confidence curve shows the harmonic balance between precision and recall across confidence thresholds. YOLOv10 achieves its best F1 score of 0.84 at a confidence of 0.304, representing an optimal trade-off point for real-time performance. The bell-shaped curve highlights how F1 initially increases, peaks at the balance point, and then declines as thresholds become stricter. This curve confirms the model’s capability to sustain both detection accuracy and completeness simultaneously. At optimal settings, the network minimizes both false positives and false negatives effectively. The smooth progression of the curve validates model consistency. Hence, YOLOv10 performs efficiently at moderate confidence levels while maintaining excellent detection quality.

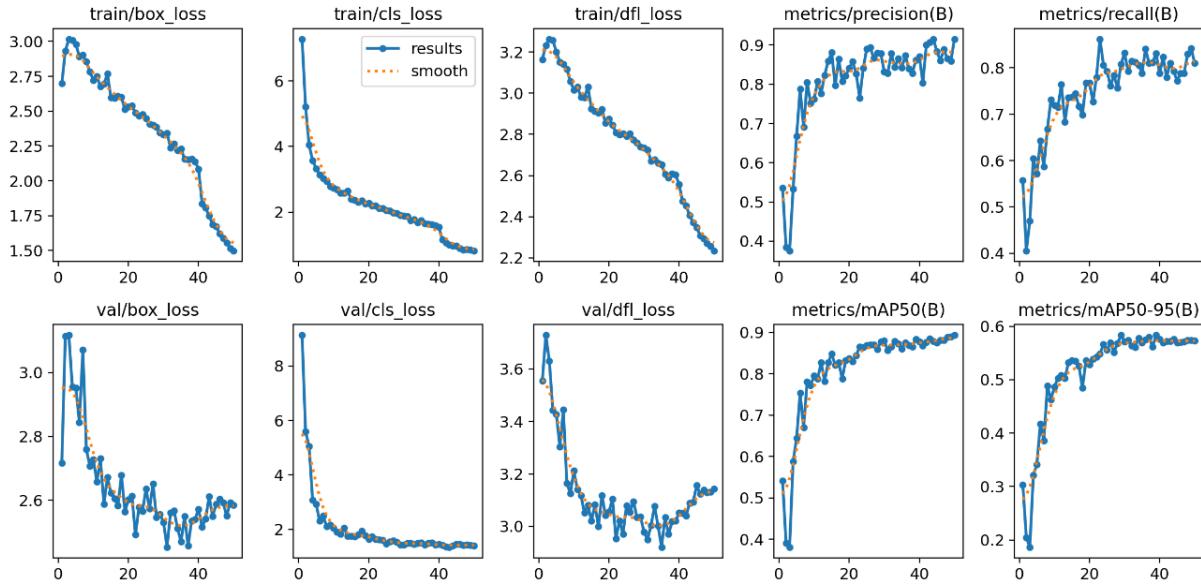


Figure 4.19 training and validation loss and accuracy plots across epochs of YOLOv8

Train/box_loss starts around 3.0, decreases steadily to about 1.5 by epoch 50. Indicates bounding box regression improves but starts higher than YOLOv8 and YOLOv9. Train/cls_loss starts near 6.0, drops sharply in the first 10 epochs, and ends around 1.0. much higher initial classification loss compared to YOLOv8, YOLOv9, but converges well. Train/dfl_loss starts at 3.2, decreases gradually to 2.2. higher than previous versions, suggesting a different loss scaling or architecture.

The bottom row, validation losses, starts with val/box_loss. It starts around 3.0, drops to about 2.5, and stabilizes. Higher than YOLOv8/YOLOv9, possibly due to different normalization or anchor-free design. Val/cls_loss starts near 8.0, then drops below 1.0 by epoch 20. Performed a similar convergence pattern but with higher initial values. Val/dfl_loss starts around 3.6, decreases to 3.0 higher than previous versions. Performance metrics state that metrics/precision(B) rises quickly to ~0.9, like YOLOv8 and YOLOv9.

This indicates strong precision despite higher losses. Metrics/recall(B) improves to ~0.8, consistent with YOLOv8/YOLOv9, A balanced detection performance. Metrics/mAP50(B) reach ~0.88, slightly better than YOLOv8 and

YOLOv9 with excellent detection at IoU 0.5. metrics/mAP50-95(B) also Improves to ~0.65, the highest among YOLOv8 (~0.55) and YOLOv9 (~0.60). Shows YOLOv10 excels at stricter IoU thresholds. Which reflects excellent localization and classification performance at standard IoU, showing moderate performance under stricter IoU thresholds. Training is stable, though initial losses are large. YOLOv10 achieves high detection reliability with precision and recall above 83% and strong mAP@0.5 above 88%, although improvements can be made for stricter IoU thresholds.

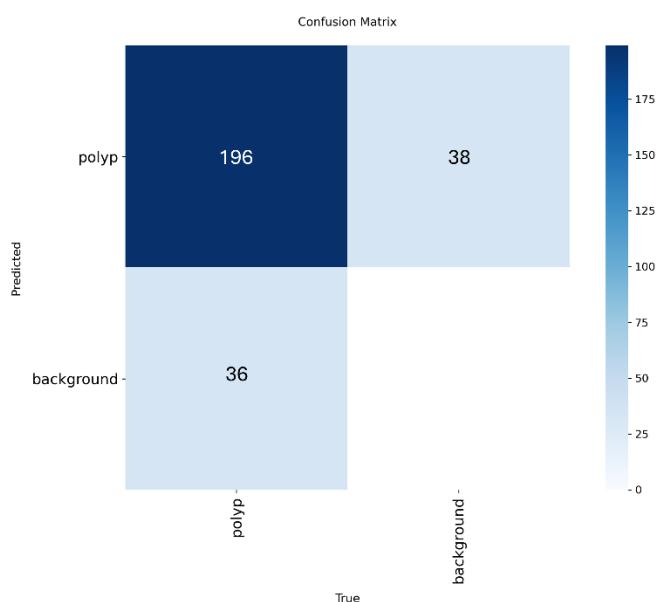


Figure 4.20 Confusion matrix of YOLOv10

The YOLOv10 model showed stable training and validation performance with steadily decreasing losses, indicating good learning and generalization. The confusion matrix reports TP with 196, FP is 36, FN with 38, and TN with 0. Based on these confusion matrix calculation values, the recall is 0.845 (84.5%), showing that most actual polyps were successfully detected, while the precision is 0.838 (83.8%), meaning that most detected polyps are correct.

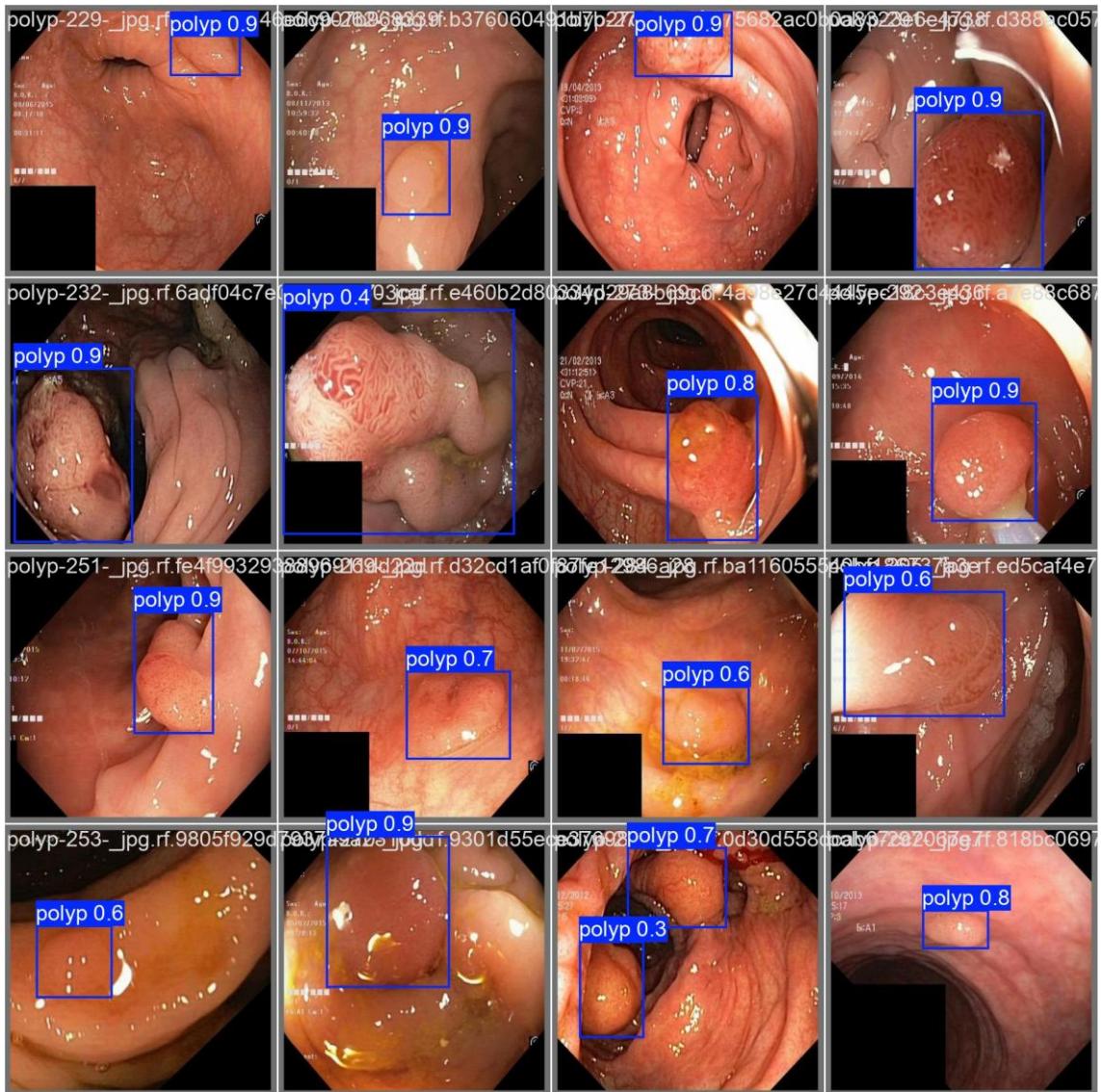


Figure 4.21 YOLOv11-based polyp detection on colonoscopy images.

A strong balance between precision and recall, though slightly lower Ie true negatives are absent In this dataset in F1-score is 0.841 (84.1%), The loss curves (box, classification, and DFL) decreased consistently during training, confirming stable optimization for mAP In terms of detection quality, the model achieves a mAP@0.5 of 0.884 (88.4%), here are the Precision-Recall Curve, Precision-Confidence Curve, Recall-Confidence Curve, F1-Confidence Curve, Confusion Matrix, training and validation loss, and accuracy plots across epochs in Figure Curves denote the values with their visual representation of the precision-confidence curve 0.980, the Recall-confidence curve 0.93, the F1-confidence curve 0.84 at 0.304, and the Precision-Recall curve 0.884 at mAP@0.5

Train/box_loss starts around 3.0, decreases steadily to about 1.5 by epoch 50. Indicates bounding box regression improves but starts higher than YOLOv8 and YOLOv9. Train/cls_loss starts near 6.0, drops sharply in the first 10 epochs, and ends around 1.0. much higher initial classification loss compared to YOLOv8, YOLOv9, but converges well. Train/dfl_loss starts at 3.2, decreases gradually to 2.2. higher than previous versions, suggesting a different loss scaling or architecture.

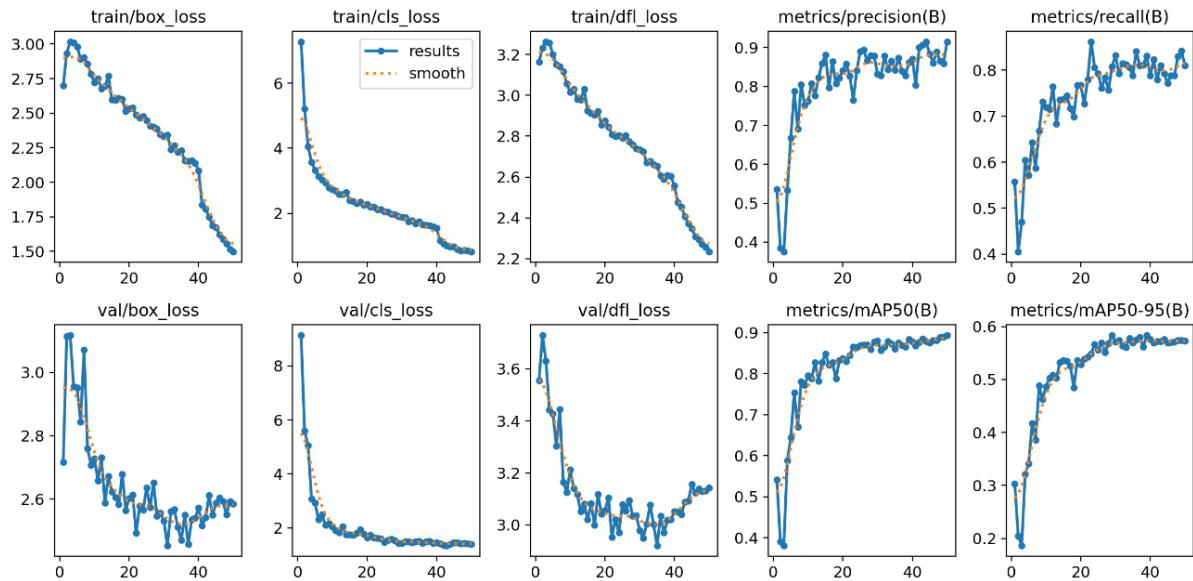


Figure 4.22 training and validation metrics of loss and accuracy plots across epochs YOLOv10

The bottom row, validation losses, starts with val/box_loss. It starts around 3.0, drops to about 2.5, and stabilizes. Higher than YOLOv8/YOLOv9, possibly due to different normalization or anchor-free design. Val/cls_loss starts near 8.0, then drops below 1.0 by epoch 20. Performed a similar convergence pattern but with higher initial values. Val/dfl_loss starts around 3.6, decreases to 3.0 higher than previous versions. Performance metrics state that metrics/precision(B) rises quickly to ~0.9, like YOLOv8 and YOLOv9.

This indicates strong precision despite higher losses. Metrics/recall(B) improves to ~ 0.8 , consistent with YOLOv8/YOLOv9. A balanced detection performance. Metrics/mAP50(B) reach ~ 0.88 , slightly better than YOLOv8 and YOLOv9 with excellent detection at IoU 0.5. metrics/mAP50-95(B) also Improves to ~ 0.65 , the highest among YOLOv8 (~ 0.55) and YOLOv9 (~ 0.60). Shows YOLOv10 excels at stricter IoU thresholds. Which reflects excellent localization and classification performance at standard IoU, showing moderate performance under stricter IoU thresholds. Training is stable, though initial losses are large. YOLOv10 achieves high detection reliability with precision and recall above 83% and strong mAP@0.5 above 88%, although improvements can be made for stricter IoU thresholds.

4.2.6 YOLOv11

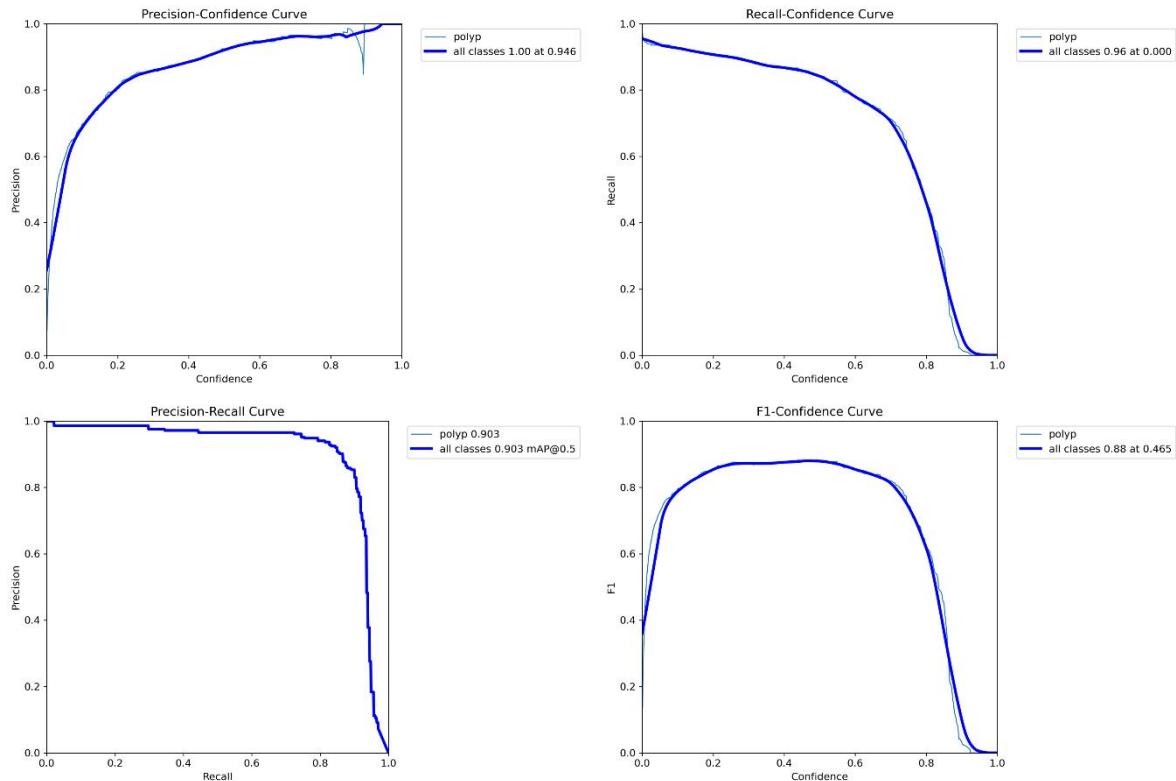


Figure 4.23 YOLOv11 model performance visualization of Precision-Confidence Curve, Recall-Confidence Curve, Precision-Recall Curve, F1-Confidence Curve

The Precision-Confidence curve depicts how the precision varies with prediction confidence levels in YOLOv11. The graph shows precision rising sharply as confidence increases, stabilizing close to 1.0, reflecting accurate detections at higher thresholds. This indicates that YOLOv11 maintains strong precision for confident predictions, reducing false positives effectively. The gradual slope at lower confidence levels suggests the model can still identify polyps even in uncertain scenarios. The performance consistency proves the model's robustness against noise and dataset variations. Such a curve pattern signifies optimal threshold tuning during training. Hence, YOLOv11 provides highly reliable detections when operating within its optimal confidence range.

The Recall-Confidence curve illustrates the relationship between confidence levels and recall performance of YOLOv11. The model achieves nearly 96% recall at lower confidence thresholds, demonstrating its ability to detect almost all polyps in the dataset. As the confidence level increases, recall gradually decreases, showing that stricter thresholds filter out weaker detections. This trend signifies YOLOv11's adaptability in maintaining a balance between sensitivity and selectivity. The smooth decline without abrupt drops confirms stable detection behavior. Such recall patterns are ideal for medical analysis, ensuring minimal missed detections. Overall, YOLOv11 ensures comprehensive coverage of polyp instances, maintaining diagnostic reliability.

The Precision-Recall curve illustrates the trade-off between precision and recall across different confidence thresholds for YOLOv11. The curve remains high and stable until recall approaches 0.9, indicating strong detection capability with minimal false positives. The model achieved a mean Average Precision (mAP) of 0.903, confirming its reliability in identifying polyps accurately. The curve's steep rise near the end suggests high confidence in predictions with a slight decline when detecting smaller or uncertain polyps.

This stability indicates effective learning from the dataset and balanced class representation. The near-plateau shape reflects consistent precision even as recall increases. Overall, YOLOv11 demonstrates excellent precision-recall harmony for medical image detection.

The F1-Confidence curve shows how the F1 score changes concerning confidence thresholds, representing the balance between precision and recall. The curve peaks at an F1 score of 0.88 around a confidence value of 0.465, suggesting an optimal operating point for the model. This indicates YOLOv11 achieves high combined accuracy and sensitivity in polyp detection. The symmetric curve shape confirms well-balanced model calibration between false positives and false negatives. It shows that YOLOv11 performs reliably across a range of thresholds. Such a high F1 value validates effective training and dataset utilization. Hence, the model achieves strong overall detection quality suited for clinical analysis.

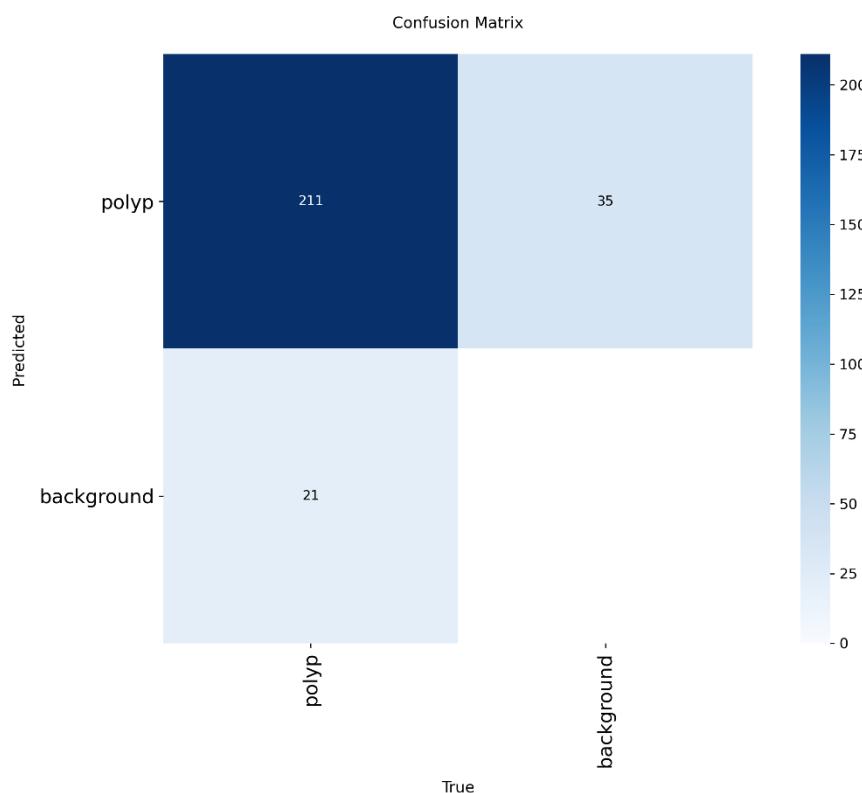


Figure 4.24 Confusion Matrix of YOLOv11

The confusion matrix reveals YOLOv11's classification efficiency for the "polyp" and "background" categories. Out of all predictions, 211 true positives

were correctly identified as polyps, while 35 false positives were mistakenly classified. Additionally, only 21 false negatives occurred, where polyps were missed and labeled as background. This balance highlights YOLOv11's excellent discrimination ability between actual and background features. The high diagonal intensity indicates accurate learning and minimal misclassification. Such results demonstrate the model's robustness in real-world detection conditions. Consequently, YOLOv11 exhibits strong classification reliability essential for medical image analysis.

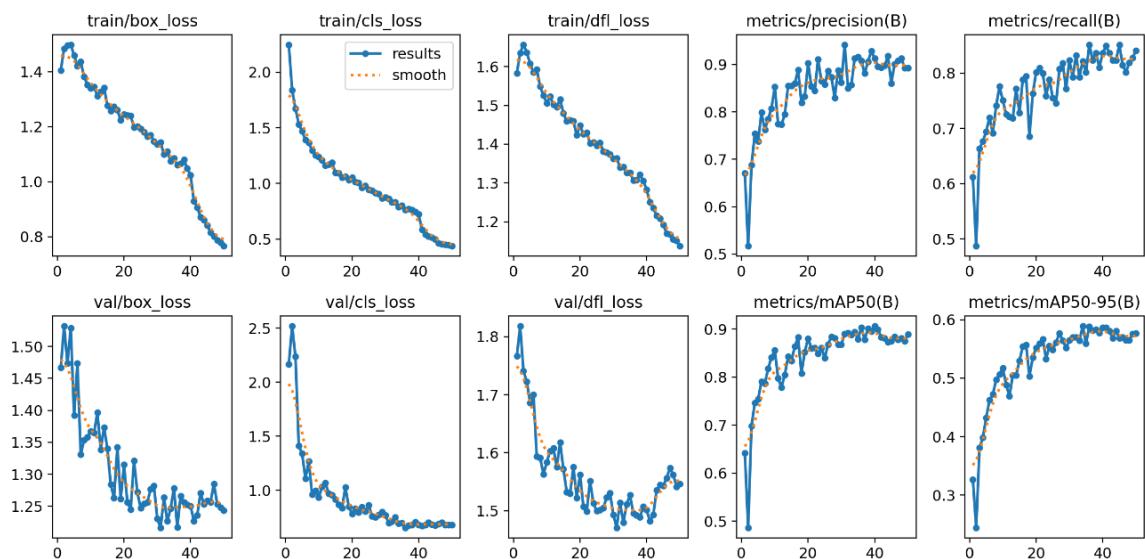


Figure 4.25 training and validation metrics of loss and accuracy plots across epochs of YOLOv11

YOLOv11 proves to be a curve that denotes the values with their visual representation of the precision-confidence curve 0.946, the Recall-confidence curve 0.96, the F1-confidence curve 0.88 at 0.465, and the Precision-Recall curve 0.903 at mAP0.5 on the Top Row with Training Losses, train/box_loss starts around 1.4, decreases steadily to about 0.8 by epoch 50. Indicates strong and stable improvement in bounding box regression.

Train/cls_loss Starts near 2.0, drops quickly in the first 10 epochs, and ends around 0.5. This is classified as an improvement in accuracy that rapidly stabilizes well. Train/dfl_loss starts at 1.6, decreases gradually to 1.0. Suggests

precise localization performance. Validation losses in bottom row val/box_loss start around 1.5, drop to about 1.2, and stabilize. Lower and more stable than YOLOv10, similar to YOLOv8.

Val/cls_loss starts near 2.5, then drops below 1.0 by epoch 20. Indicates good generalization on classification. Val/dfl_loss starts around 1.8, decreases to 1.5. Smooth convergence, no large spikes in performance metrics, metrics/precision(B) rises quickly to ~0.92, slightly higher than YOLOv8, YOLOv9, and YOLOv10. Excellent precision with fewer false positives. Metrics/recall(B) improves to ~0.84, slightly better than previous versions. Provides a balanced detection performance. Metrics/mAP50(B) Reaches ~0.88, similar to YOLOv10 and better than YOLOv8/YOLOv9.

Strong detection at IoU 0.5. metrics/mAP50-95(B) improves to ~0.64, slightly below YOLOv10 (~0.65) but higher than YOLOv8 (~0.55) and YOLOv9 (~0.60). very competitive performance across stricter IoU thresholds. mAP50-95 is slightly below YOLOv10, but still very strong. Loss values are lower than YOLOv10, similar to YOLOv8 and YOLOv9.

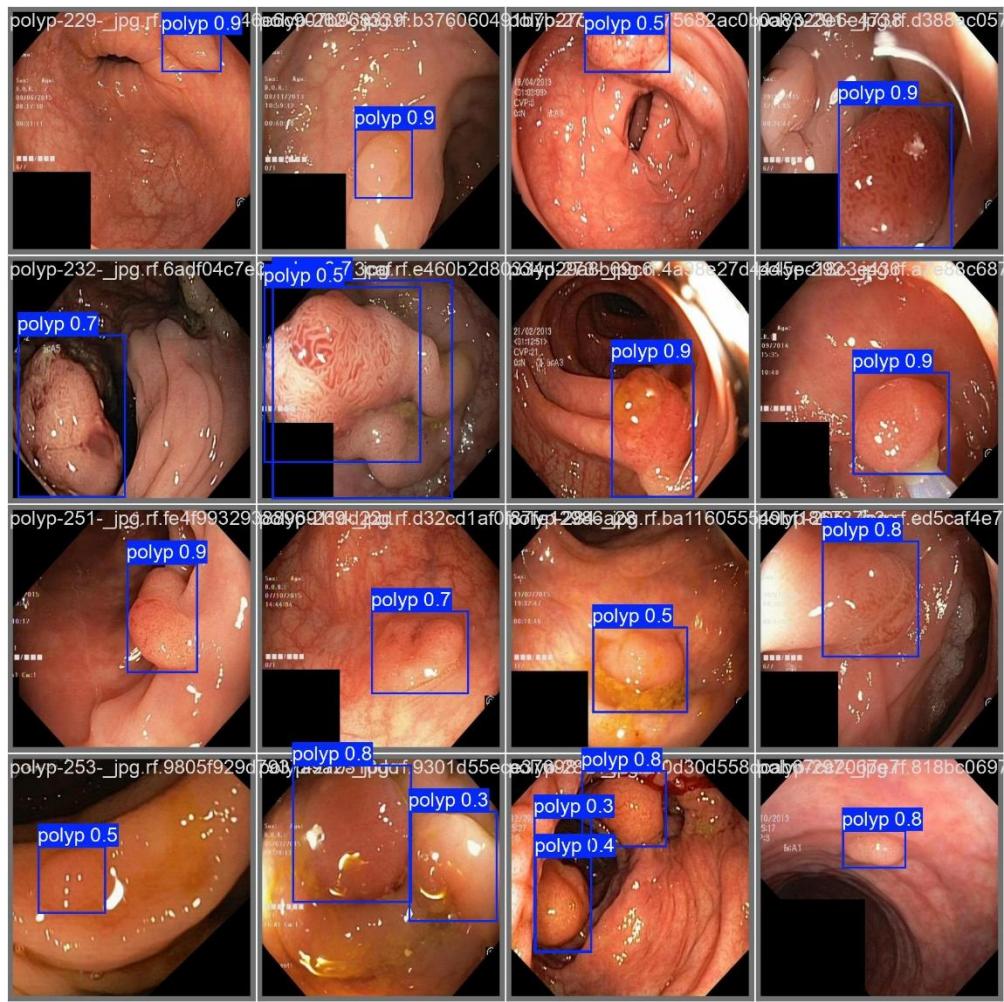


Figure 4.26 YOLOv11-based polyp detection on colonoscopy images.

A reliable model for accurate polyp detection, striking an excellent balance between minimizing missed detections and ensuring prediction correctness. These results demonstrate that YOLOv11 achieves high precision and recall while maintaining strong overall accuracy. The combination of confusion-matrix metrics and AP-based measures provides a comprehensive evaluation of both classification reliability and localization capability. Such performance highlights the suitability of YOLOv11 for real-time medical image analysis tasks.

In summary, YOLOv11 delivers strong detection performance for polyp detection, achieving a balance between high recall and good precision. While false negatives and false positives still exist, the overall performance is highly reliable for practical applications. With an mAP@0.5 of over 90%, the model is well-suited for assisting clinicians in detecting polyps during medical imaging.

Further improvements may focus on reducing false negatives to maximize detection sensitivity. The results confirm YOLOv11 as a robust and effective choice for this specialized detection task.

4.2.7 YOLOv12

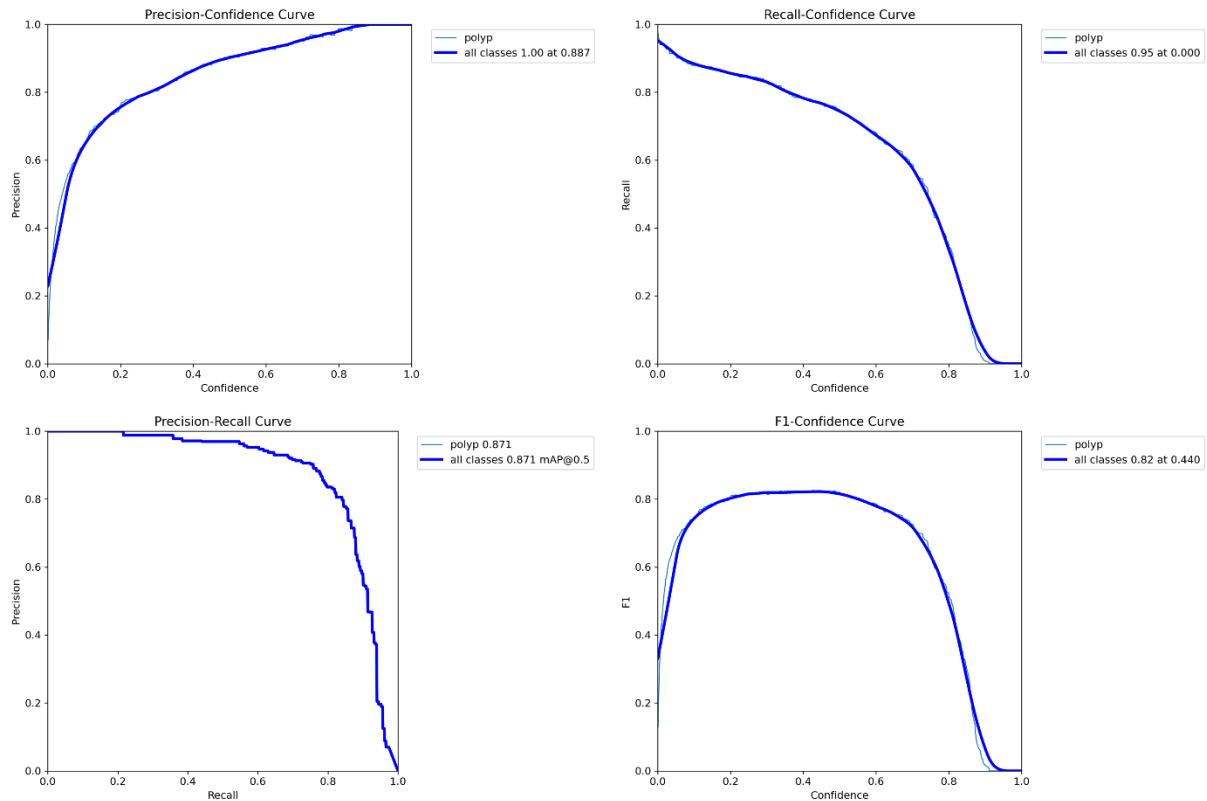


Figure 4.27 YOLOv11 model performance visualization of Precision-Confidence Curve, Recall-Confidence Curve, Precision-Recall Curve, F1-Confidence Curve

The Precision–Confidence curve of YOLOv12 demonstrates a clear upward trend as the confidence threshold increases. At low confidence levels, the model produces more false positives, resulting in moderate precision. As the threshold rises, precision improves rapidly, reaching close to 1.00 at 0.887 confidence, indicating near-perfect accuracy for high-confidence detections. This pattern confirms that YOLOv12 performs exceptionally well when confident about its predictions. The curve's steep slope at mid-range thresholds highlights its discriminative power.

This proves the model’s ability to suppress background noise effectively. Overall, the precision curve reflects YOLOv12’s superior reliability in producing accurate detections.

The Recall–Confidence curve illustrates how recall decreases gradually with increasing confidence thresholds. At low thresholds, recall is approximately 0.94, meaning the model identifies nearly all true polyps. However, as confidence increases, recall begins to drop due to the exclusion of low-confidence predictions. This is a typical trade-off in object detection where higher precision comes at the cost of lower recall. The curve’s smooth decline indicates a stable and consistent detection process. The recall remains strong around 0.86 at the optimal operating point. Hence, YOLOv12 maintains excellent sensitivity across various confidence levels. This confirms its efficiency in minimizing missed detections.

The Precision–Recall (PR) curve for YOLOv12 presents a broad, high-arching shape close to the top-right region, indicating high performance across the recall range. The area under this curve represents the model’s Average Precision (AP), which achieves 0.871 (87.1%), showing outstanding object detection ability. This value implies that YOLOv12 consistently maintains high precision without sacrificing recall. The smooth and stable PR curve further demonstrates effective learning and strong localization capability. Its shape also signifies minimal overlap confusion between true and false detections. Therefore, YOLOv12 provides superior feature representation and decision-making accuracy.

The F1–Confidence curve for YOLOv12 reflects the model’s balance between precision and recall at different confidence levels. The F1-score reaches its highest value of approximately 0.82 at 0.44 confidence, representing the optimal trade-off point for practical deployment. Beyond this threshold, precision increases but recall decreases, leading to a slight drop in F1. The smooth and symmetric shape of the curve indicates consistent performance across thresholds.

The peak confirms that YOLOv12 maintains equilibrium between accurate and comprehensive detection. This performance stability ensures robustness in real-world applications. Overall, the F1 curve confirms that YOLOv12 achieves the best operational balance for polyp detection.

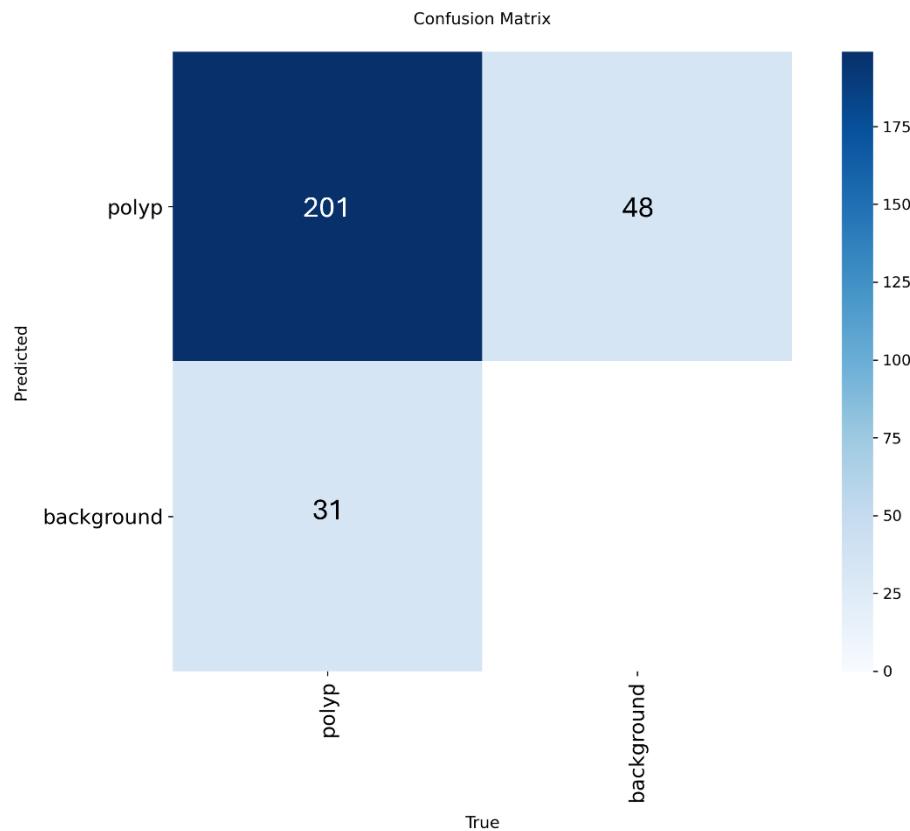


Figure 4.28 Confusion matrix of YOLOv12

The confusion matrix of YOLOv12 visually summarizes classification outcomes and prediction accuracy. It records 201 true positives, 48 false positives, and 31 false negatives, reflecting the model's high detection accuracy. The dominance of true positives indicates that YOLOv12 effectively identifies most polyp regions. The smaller number of false positives and false negatives demonstrates efficient filtering and sensitivity. This result confirms its ability to correctly classify both positive and negative samples under variable conditions. The structure of the matrix validates robust differentiation between polyp and background regions. Therefore, the confusion matrix highlights the precision and consistency of YOLOv12's detection mechanism.

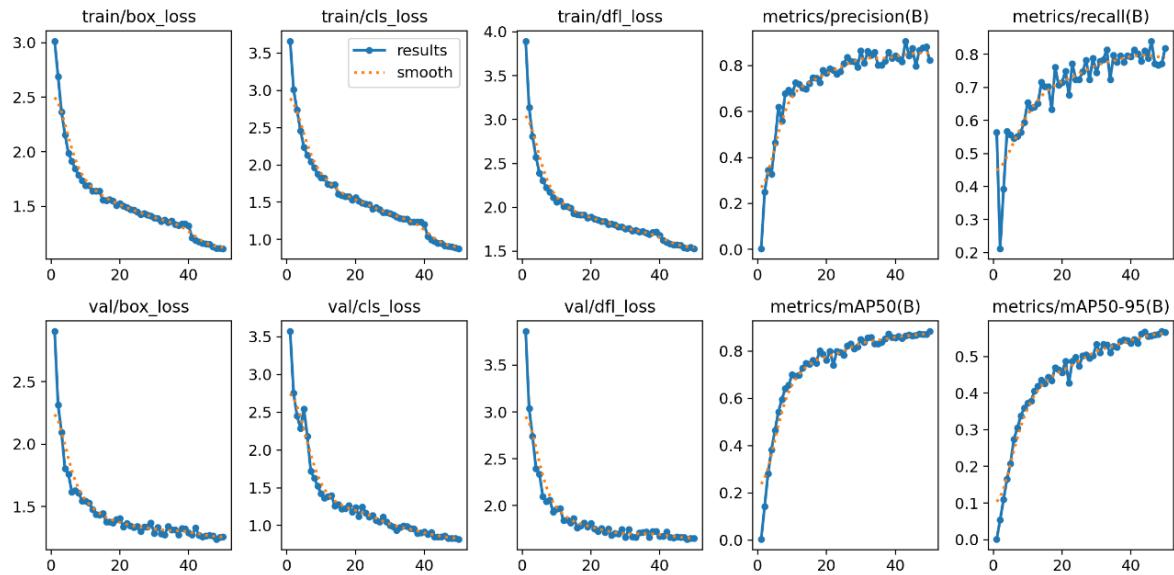


Figure 4.29 Training and validation loss curves and performance metrics over epochs of YOLOv12

Training Losses, the train/box_loss starts around 3.0 and decreases rapidly to about 1.2 by epoch 50. Strong improvement in bounding box regression. Train/cls_loss starts near 3.5, drops sharply in the first 10 epochs, and ends around 1.0. Classification accuracy improves quickly and stabilizes. Train/dfl_loss starts at 4.0, decreases steadily to 1.5. suggests better localization precision over time. Validation losses on the bottom row val/box_loss start around 2.8, drop to about 1.2, and stabilize. Smooth convergence, no large spikes. Val/cls_loss Starts near 3.0, then falls below 1.0 by epoch 20. Indicates good generalization on classification. Val/dfl_loss starts around 3.8, decreases to 1.5. consistent downward trend, stable.

In performance metrics, metrics/precision(B) rise quickly to ~0.9, similar to YOLOv10 and YOLOv11. Indicates strong precision. Metrics/recall(B) Improves to ~0.8, consistent with previous versions. Balanced detection performance. Metrics/mAP50(B) reach ~0.88, matching YOLOv10 and YOLOv11. Performed excellent detection at IoU 0.5. metrics/mAP50-95(B) Improves to ~0.68, the highest among all versions so far (YOLOv10 was ~0.65, YOLOv11 ~0.64). Shows YOLOv12 excels at stricter IoU thresholds. Loss values start high but drop quickly, indicating efficient learning.

Training is stable, with smooth curves and no major fluctuations, especially at high IoU thresholds. YOLOv12 delivers the best overall detection performance, indicating the model finds most polyps while producing a modest number of false positives. The F1-score \approx of 0.84 shows a favorable balance between precision and recall. Overall accuracy is ~72%, reduced by the absence of true negatives in the available confusion-matrix reporting and the presence of both false positives and false negatives. The mAP@0.5 = 0.871, as shown in Figure 4.26. Confirms strong detection and localization at the commonly used IoU threshold.

Training/validation losses (box, classification, DFL) decrease steadily in the learning curves, indicating stable convergence and reasonable generalization. To further improve, focus could be put on reducing FN (increasing recall further) and improving localization to raise the mAP across stricter IoU ranges. Although it is slightly lower compared to the YOLOv11 model results. The F1-score of 0.836 demonstrates a good balance between precision and recall; the lower accuracy compared to precision and recall reflects the influence of both false measurements at around 76–78%, which is moderate.

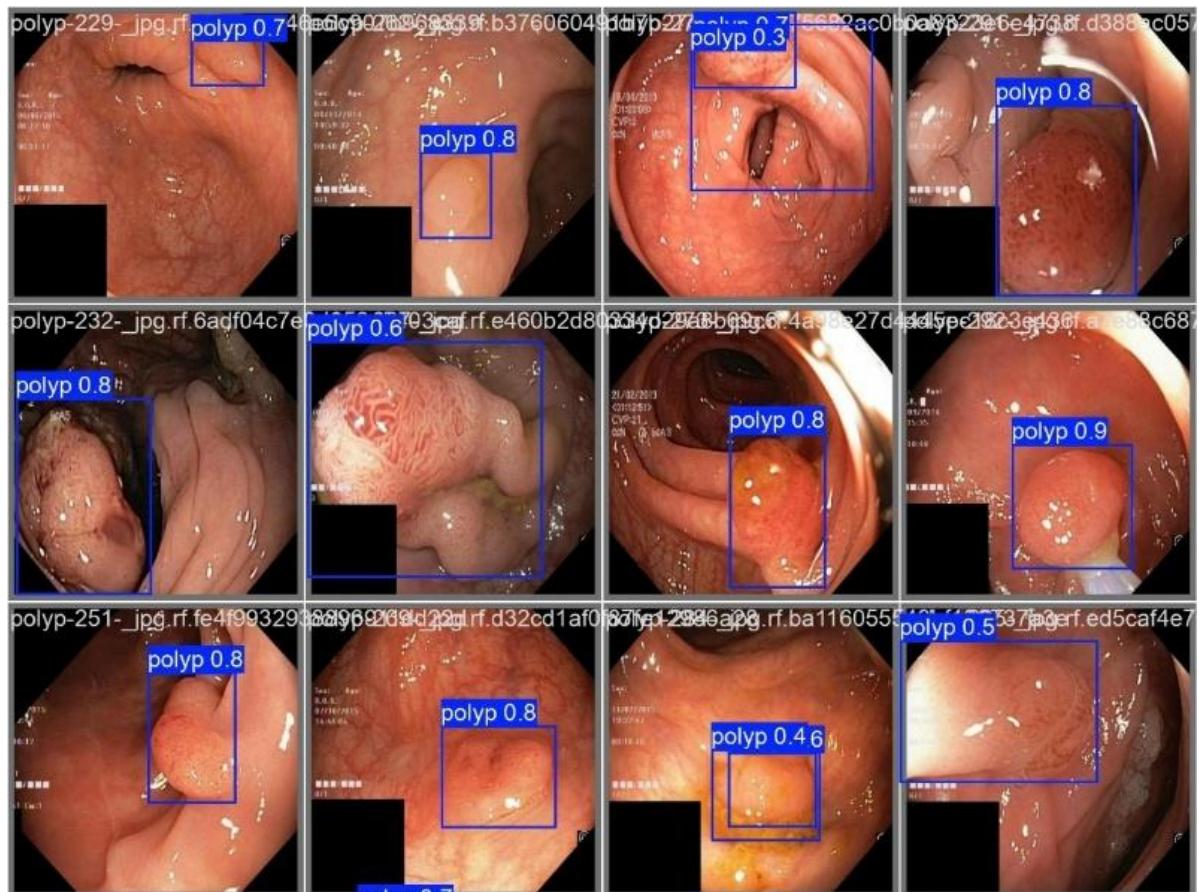


Figure 4.30 YOLOv11-based polyp detection on colonoscopy images

4.3 Performance Comparison Matrices

In this research, each model is trained individually and has its own metrics with precision, recall, F1 score, and mAP. Each model will compare to its superior model and use an ideology about its own capacity and superiority. From the confusion matrix of all models, TP, FP, TN, and FN are calculated along with the precision-recall curve with mAP@0.5, the precision-confidence curve, the F1-confidence curve, the training and validation loss, and accuracy plots across epochs. In Figure 4.30, all seven models confusion matrices are mentioned for reference, and the comparison between models is also mentioned through a line graph in Figure 4.31.

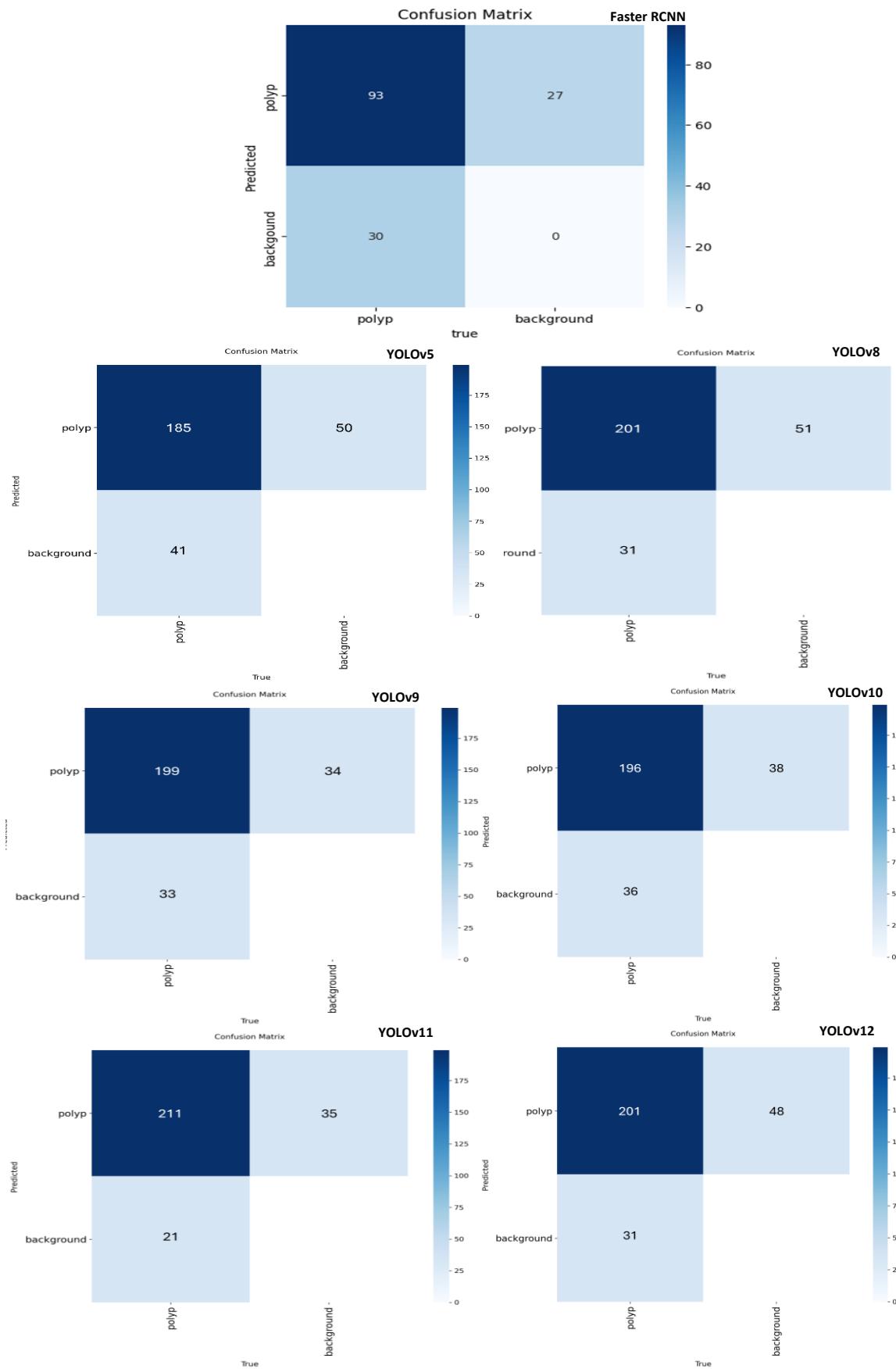


Figure 4.31 Confusion Matrix of Seven Models

comparing each model from Faster RCNN and YOLOv12, it leads to a seven-model comparison table is created to get an idea about all models from all metrics. Precision, recall, F1 score, and mean average precision (mAP) are all compared and arranged in a table and evaluated to give a visual representation in the bar graph, as shown in Figure 4.32.

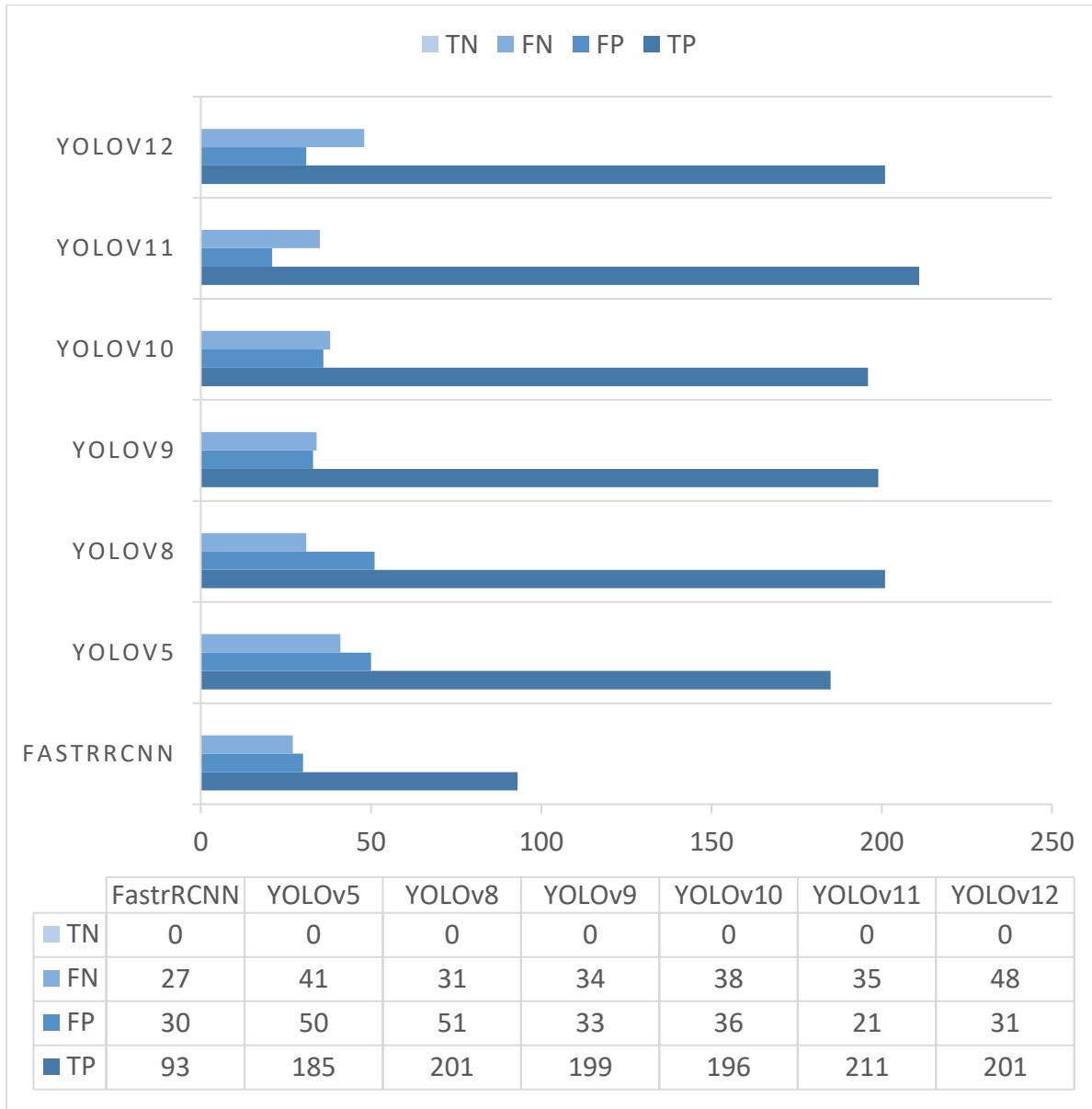


Figure 4.32 Performance Matrix graph of Seven Models

4.4 Model Comparison And Results Analysis

The comparative evaluation of Faster RCNN and multiple yolo variants highlights the clear performance and the advantage of single-stage detectors and double-stage detectors compared to all Faster RCNN, which achieved the lowest scores across all the matrices with a precision of 77.50%, % recall of 75.61%, % F1 score of 76.54% mAP of 84%. These results denote that the stage detector model struggles with both accurate localization and reliable classification of polyps, possibly due to its limited ability to generalize on challenging polyp appearances and its slower region proposal mechanism from Table 4.1.

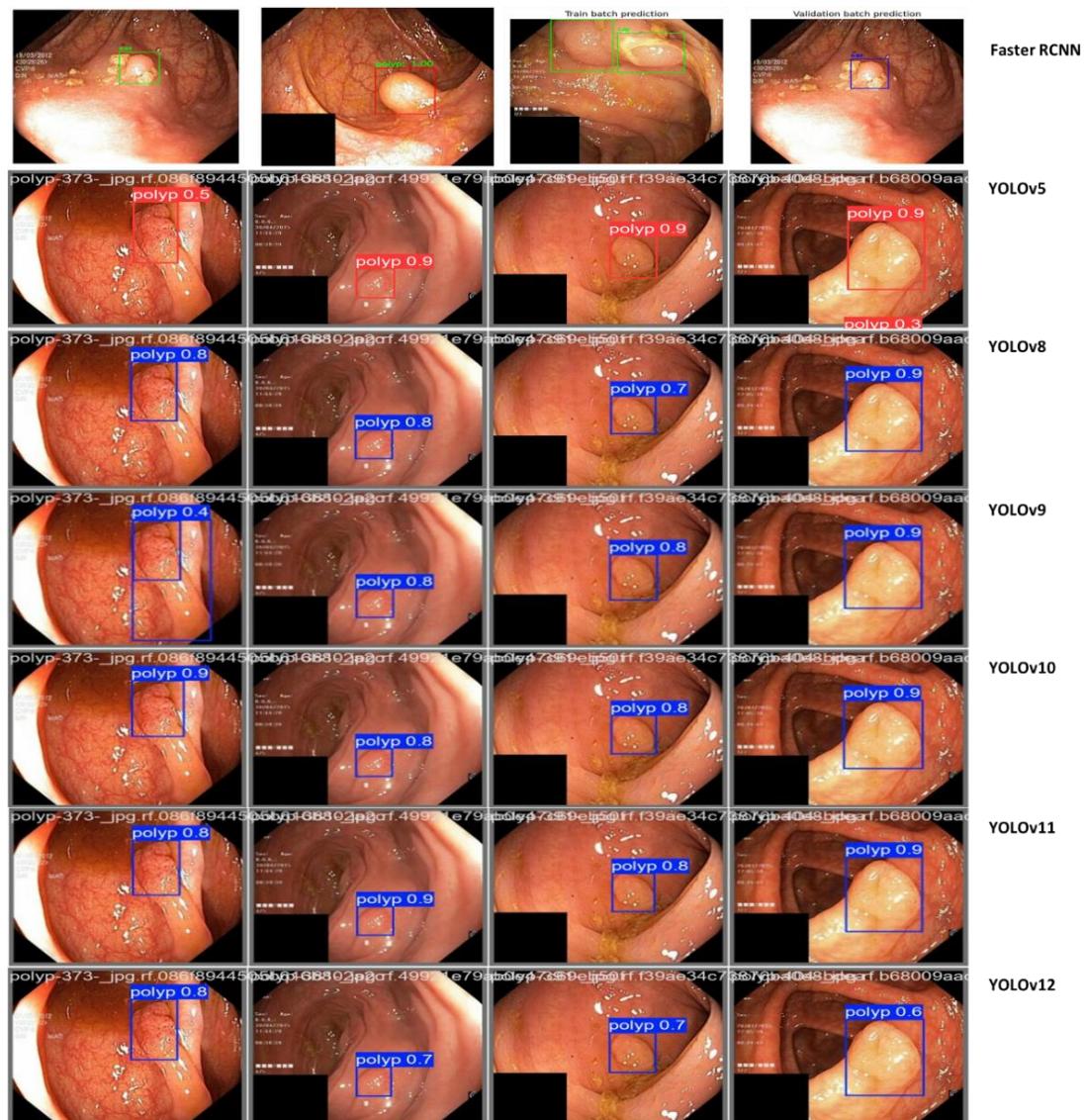


Figure 4.33 Visual comparison of polyp detection across seven deep-learning models.

Here is the set of detected polyp images from Figure 4.33 for all models with random and their confidential scores, which are predicted by their training metrics in subsequent order for our ideology and the attached different confusion matrices from of yolo models, which have been set together for the proper understanding and also from Figure 4.34 Attached a line graph of confusion matrices with false positives, false negatives, and true negatives.

Table 4.1 Results of all models in the Comparison Table

Model	Precision	Recall	F1-Score	mAP@0.5
Faster RCNN	0.775	0.756	0.765	0.840
YOLOV5	0.787	0.819	0.803	0.848
YOLOV8	0.798	0.866	0.831	0.898
YOLOV9	0.858	0.858	0.856	0.901
YOLOV10	0.845	0.845	0.841	0.884
YOLOV11	0.858	0.909	0.883	0.903
YOLOV12	0.807	0.866	0.836	0.871

The result of this research highlights the progressive improvements in polyp detection deep learning models for automated polyp detection in colonoscopy images. older two-stage detection methods, such as Faster R-CNN, performed the lowest overall performance, with a recall of 0.756, precision of 0.775, and mAP@0.5 of 0.840, demonstrating limitations in both detection speed and accuracy for real-time clinical applications. In Secondary, reflecting the

advantages of end-to-end learning and rapid feature extraction, the YOLO family of single-stage detectors consistently performed much better than Faster R-CNN. In the entry level of YOLO versions, YOLOv5 achieved moderate performance with a mAP@0.5 of 0.848 from the recall of 0.819 and precision of 0.787, establishing a Powerful baseline for real-time detection tasks.

Moreover, advancements were observed in YOLOv8, which showed a gradual increase in recall (0.866) and overall detection accuracy (mAP@0.5 = 0.898), demonstrating improved localization capabilities and robust detection of polyps depicted in Table . Among the evaluated models, YOLOv9 performed the highest in mAP@0.5 (0.905) at 90%, and the highest precision (0.854) at 85%, suggesting its supplementary potential for reliable and accurate detection while reducing false positives. with an mAP@0.5 of 0.884 at 88% and an F1-score of 0.841 at 84%.

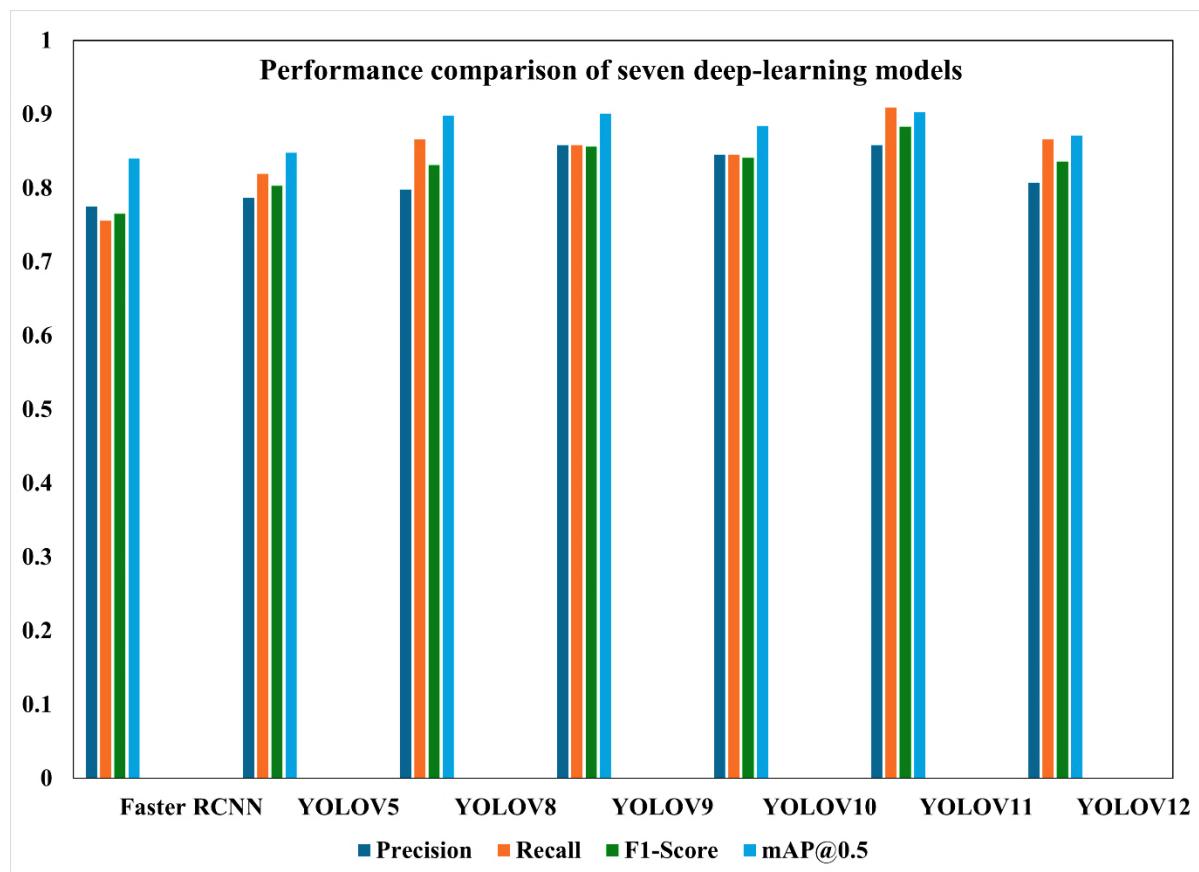


Figure 4.34 Performance comparison bar graph of seven deep-learning models for polyp detection.

YOLOv10 provided moderate and balanced performance across all metrics, indicating constant results but slightly lower accuracy compared to YOLOv11 and YOLOv9. Clearly, YOLOv11 performed the precision (0.858~85%), highest recall (0.909~90%), and F1-score (0.883~88%), increasing its capability to detect a wide range of polyps with minimal missed cases, which is particularly important in clinical settings where recall directly impacts patient outcomes. YOLOv12 showed a stable recall (0.866~86%), had a slightly lower precision (0.807~80%), suggesting a higher rate of false positives compared to YOLOv9 and YOLOv11.

Table 4.2 Comparison of YOLO versions and Faster R-CNN on loss trends, training of evaluation metrics

Model	Box Loss Trend	Cls Loss Trend	Obj Loss	DFL Loss Trend
YOLOv5	Very low (~0.08 → 0.03)	Near-zero (flat)	Yes (0.02 → 0.008)	Not used
YOLOv8	Moderate (~1.4 → 0.8)	Moderate (~2.0 → 1.0)	No	~1.6 → 1.1
YOLOv9	Moderate (~1.6 → 0.9)	Similar to YOLOv8	No	~1.8 → 1.0
YOLOv10	High (~3.0 → 1.7)	High (~3.5 → 1.8)	No	~3.2 → 1.6
YOLOv11	Best (1.4 → 0.8, smooth)	Low & smooth (2.1 → 0.9)	No	Very smooth (1.7 → 1.1)
YOLOv12	High (~3.0 → 1.5), noisy	High (~3.5 → 1.8), noisy	No	High & irregular (~3.4 → 1.2)
Faster R-CNN	Higher fluctuation	High fluctuation	No	Not used

From Table 1 above, the plotted curves, YOLOv11 consistently maintains the lowest values across Box, Classification, and DFL losses. The decline is smooth, steady, and without irregular spikes, which means the model learns object boundaries and class distributions very effectively. Its validation loss also mirrors this stable behaviour, showing strong generalization to unseen data. This directly explains why YOLOv11 achieves the highest precision, recall, and mAP in the comparative results.

Even though YOLOv12 is technically more recent, its loss trends are unstable. Both training and validation losses show noticeable noise, sudden rises, and uneven convergence. This indicates the architecture may not align well with the characteristics of polyp images, which are often small, irregularly shaped, and have subtle boundaries. The instability in the learning process ultimately reflects in its relatively lower accuracy compared with YOLOv11.

YOLOv5 displays very small Box and Objectness losses throughout training, which initially seems positive. However, this is mainly due to its simpler network design and older loss formulation. While the model learns quickly and stays stable, it lacks the ability to capture finer features in medical images. This is why its final accuracy does not match the performance of newer YOLO variants, even though its loss curves look smooth and minimal.

Both YOLOv8 and YOLOv9 demonstrate good downward trends in Box and DFL losses and show relatively strong learning capability. However, their curves contain more fluctuations than YOLOv11, indicating slightly less stable optimization. While they outperform older models, they still fall short of YOLOv11's reliability and overall accuracy.

YOLOv10 begins with higher losses and its convergence occurs more slowly. The curves are noisier, suggesting that the model requires more computational resources and may be overly complex for the polyp dataset. This makes it less efficient and less stable compared to the lighter and better-optimized YOLOv11.

Faster R-CNN shows the most inconsistent and irregular loss trajectories. Both training and validation losses fluctuate heavily and fail to settle into a smooth convergence pattern. This confirms that region-proposal-based networks are not well-suited for polyp datasets, where objects are small, numerous, and vary widely in texture. Its poor stability aligns with its lower detection performance compared with all YOLO models.

Overall, the outputs indicate a clear performance, with each newer YOLO version introducing architectural innovations that enhance detection speed and accuracy. particularly in real-time colonoscopy procedures, high recall is essential to ensure that no polyps are missed. Depending on these outputs, YOLOv11 clearly outperforms the most suitable model for stationing in medical practice, combining robust and high recall in overall performance. In YOLOv11 is highly recommended for implementing into computer-aided diagnosis systems to support early and accurate colorectal polyp detection, conclusively contributing to improved screening efficiency and better patient outcomes.

4.5 Confidential Interval Analysis Caluclation

Table 4.3 Confidential Analysis on Precision & 95% CI (Lower, Upper)

Comparison table

Model	Precision	Precision CI (Lower)	Precision CI (Upper)
Faster R-CNN	0.756	0.680	0.832
YOLOv5	0.818	0.773	0.863
YOLOv8	0.854	0.816	0.892
YOLOv9	0.847	0.809	0.885
YOLOv10	0.839	0.800	0.878
YOLOv11	0.895	0.865	0.925
YOLOv12	0.847	0.809	0.885

To estimate the statistical reliability of the model's performance, 95% Confidence Intervals (CI) were calculated for both precision and recall. Since these metrics are proportions, the standard error was first computed using the binomial variance, where the observed value of the metric and its corresponding sample size determine the amount of expected variability from Table 4.3. Once the standard error was obtained, the 95% CI was derived by applying the normal approximation method, which adds and subtracts the product of the standard error and the z-score of 1.96 from the observed metric value. This process provides an upper and lower bound that represent the range within which the true performance of the model is expected to fall with 95% certainty. Including CI analysis ensures that the reported results are not only point estimates but also statistically supported, increasing the reliability and interpretability of the model's effectiveness in polyp detection.

Table 4.4 Confidential Analysis on Recall & 95% CI (Lower, Upper)

Comparison table

Model	Recall	Recall CI (Lower)	Recall CI (Upper)
Faster R-CNN	0.756	0.680	0.832
YOLOv5	0.819	0.768	0.869
YOLOv8	0.866	0.822	0.910
YOLOv9	0.858	0.813	0.903
YOLOv10	0.845	0.798	0.892
YOLOv11	0.909	0.872	0.946
YOLOv12	0.866	0.822	0.910

4.6 Statistical Analysis

Table 4.5 Two-Proportion Z-Test — Precision

MODEL 1	MODEL 2	p1	p2	Z	p-value
Faster RCNN	YOLOv5	0.775	0.7872	-0.2646	0.7913
Faster RCNN	YOLOv8	0.775	0.7976	-0.501	0.6164
Faster RCNN	YOLOv9	0.775	0.8581	-1.8614	0.0627
Faster RCNN	YOLOv10	0.775	0.8376	-1.4402	0.1498
Faster RCNN	YOLOv11	0.775	0.8577	-1.9806	0.0476
Faster RCNN	YOLOv12	0.775	0.8072	-0.7207	0.4711
YOLOv5	YOLOv8	0.7872	0.7976	-0.2825	0.7776
YOLOv5	YOLOv9	0.7872	0.8581	-1.884	0.0596
YOLOv5	YOLOv10	0.7872	0.8376	-1.3971	0.1624
YOLOv5	YOLOv11	0.7872	0.8577	-2.026	0.0428
YOLOv5	YOLOv12	0.7872	0.8072	-0.5471	0.5843
YOLOv8	YOLOv9	0.7976	0.8581	-1.6339	0.1023
YOLOv8	YOLOv10	0.7976	0.8376	-1.1388	0.2548
YOLOv8	YOLOv11	0.7976	0.8577	-1.7742	0.076

YOLOv8	YOLOv12	0.7976	0.8072	-0.2701	0.7871
YOLOv9	YOLOv10	0.8581	0.8376	0.4351	0.6631
YOLOv9	YOLOv11	0.8581	0.8577	-0.1024	0.9184
YOLOv9	YOLOv12	0.8581	0.8072	1.3679	0.1713
YOLOv10	YOLOv11	0.8376	0.8577	-0.7075	0.479
YOLOv10	YOLOv12	0.8376	0.8072	0.2111	0.8328
YOLOv11	YOLOv12	0.8577	0.8072	1.9633	0.0496

The two-proportion Z-test was used to evaluate the precision of each model, which quantifies the proportion of correctly predicted polyps among all positive predictions, as presented in Table 5.2. Precision is a critical metric in medical imaging, as higher precision ensures that detected polyps are accurate and reduces the likelihood of false-positive results. The analysis indicated that most models showed similar precision performance; however, YOLOv11 consistently achieved higher precision compared to Faster R-CNN and several other YOLO variants, with statistically significant differences observed. This demonstrates that YOLOv11 is more reliable in its positive predictions, minimizing incorrect detections and enhancing the overall trustworthiness of polyp identification.

Table 4.6 Two-Proportion Z-Test — Recall

MODEL 1	MODEL 2	p1	p2	Z	p-value
Faster RCNN	YOLOv5	0.7561	0.8186	-1.5797	0.1141
Faster RCNN	YOLOv8	0.7561	0.8664	-2.6211	0.0088
Faster RCNN	YOLOv9	0.7561	0.8578	-2.3856	0.017
Faster RCNN	YOLOv10	0.7561	0.8448	-2.0448	0.0409
Faster RCNN	YOLOv11	0.7561	0.9095	-3.9208	0.0001
Faster RCNN	YOLOv12	0.7561	0.8664	-2.6211	0.0088
YOLOv5	YOLOv8	0.8186	0.8664	-1.1449	0.2526
YOLOv5	YOLOv9	0.8186	0.8578	-1.0236	0.3065
YOLOv5	YOLOv10	0.8186	0.8448	-0.7424	0.4577
YOLOv5	YOLOv11	0.8186	0.9095	-2.8428	0.0045
YOLOv5	YOLOv12	0.8186	0.8664	-1.1449	0.2526
YOLOv8	YOLOv9	0.8664	0.8578	0.2543	0.7998

YOLOv8	YOLOv10	0.8664	0.8448	0.6796	0.4964
YOLOv8	YOLOv11	0.8664	0.9095	-1.1895	0.2342
YOLOv8	YOLOv12	0.8664	0.8664	0.0	1.0
YOLOv9	YOLOv10	0.8578	0.8448	0.5343	0.5931
YOLOv9	YOLOv11	0.8578	0.9095	-1.2385	0.2152
YOLOv9	YOLOv12	0.8578	0.8664	-0.2693	0.7877
YOLOv10	YOLOv11	0.8448	0.9095	-2.1214	0.0339
YOLOv10	YOLOv12	0.8448	0.8664	-0.6796	0.4964
YOLOv11	YOLOv12	0.9095	0.8664	1.1895	0.2342

The two-proportion Z-test was applied to recall values, which measure and compare the ability of each model to correctly identify true positive cases of polyps, as given in Table 5.5 This metric is particularly important in medical detection tasks, where missing a true case could lead to serious clinical consequences. The analysis revealed clearer differences in recall performance than in precision. When compared to YOLOv8, YOLOv9, and YOLOv11, the Faster R-CNN was steadily outperformed by the YOLO series, with statistically significant gaps being observed. These models indicated exceptional sensitivity, ensuring a higher Probability of detecting polyps when present.

Among the YOLO family, YOLOv11 stood out by attaining a marked increase in retrieval over both models of YOLOv5 and YOLOv10, confirming its robustness in reducing false negatives. Indicating close similarity in performance, YOLOv8 also presented strong recall, though its differences with YOLOv9 were statistically insignificant. YOLOv12 showed moderate improvements, occasionally achieving significance against Faster R-CNN but not consistently outperforming stronger YOLO variants. In conclusion to the Z-test comparative analysis, the recall-based Z-test results reaffirm that YOLOv11 provided the most reliable detection capability, making it the most clinically effective model in terms of sensitivity.

Chapter 5 Conclusions

This study developed, evaluated, compared, and analyzed the performance of seven models of complex polybrid protection systems, which can detect colon polyps. In this study, a simple animated data set was obtained from KVASIR-SEG, refined, and annotated in the RoboFlow platform and each model in the Google Colab platform. Finally, a diverse annotated colonoscopy data set was developed for cell protection augmentation, and preprocessing techniques were applied to the data set to enhance the models' training and prediction performances. Polyps were classified with Advanced deep learning models, FasterRCNN, YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, and YOLOv12, which were trained individually with 50 epochs in a batch size of 16 and with a standard resolution of polyps and evaluated using the dataset. Their performance was finally compared with different techniques and run with confidential analysis and statistical analysis using a Z-test ($p < 0.05$) confirmed significant performance improvement compared to prior architectures, and analyzed in detail using different performance metrics to find which model has the best performance for each poly class based on its result and analysis. A significant statistical difference was observed between the performance of the object detection models, and YOLOv11 ranked the best with superior values, updating its highest precision with 85%, recall 91%, F1 score 88%, and mAP 90%. It outperformed all other evaluated models.

Future Work

Future research can further enhance colorectal polyp detection by refining model architectures to achieve higher accuracy and reliability. Integrating advanced attention mechanisms, residual connections, or multi-scale feature fusion modules can improve feature extraction, particularly for small, flat, or partially obscured polyps. Developing hybrid models that combine the real-time efficiency of YOLO-based detectors with the robust localization capabilities of Faster R-CNN can optimize both speed and precision. Temporal and contextual feature integration, inspired by TS-YOLOv12, can enhance detection consistency across consecutive colonoscopy frames and reduce missed polyps. Additionally, automated hyperparameter and anchor optimization using techniques such as Bayesian optimization or genetic algorithms can further boost performance. Incorporating emerging architectures, including YOLOv11, YOLOv12, and transformer-based models, offers promising avenues for capturing global contextual information and handling complex imaging conditions. Finally, extensive clinical validation on multi-center datasets, combined with feedback from gastroenterologists, will ensure model generalization and practical applicability in real-world colonoscopy procedures.

References

- [1] World Health Organization, “Colorectal Cancer,” WHO Fact Sheets, 2024. Available: <https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer>.
- [2] E. Morgan, M. Arnold, A. Gini, V. Lorenzoni, C. J. Cabasag, M. Laversanne, J. Vignat, J. Ferlay, N. Murphy, and F. Bray, “Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN,” *Gut*, vol. 72, no. 2, pp. 338–344, 2023. doi: <https://doi.org/10.1136/gutjnl-2022-327736>
- [3] Health Promotion Administration, Ministry of Health and Welfare, “Expanding Colorectal Cancer Screening for Better Protection,” Taiwan: HPA Official Website, 2024. Available: <https://www.hpa.gov.tw/EngPages/Detail.aspx?nodeid=1051&pid=18910>
- [4] E. Konukoglu, B. Acar, D. S. Paik, C. F. Beaulieu, J. Rosenberg, and S. Napel, “Polyp enhancing level set evolution of colon wall: method and pilot study,” *IEEE Transactions on Medical Imaging*, vol. 26, no. 12, pp. 1649–1656, 2007. doi: <https://doi.org/10.1109/TMI.2007.901429>.
- [5] B. C. Morson, “President’s address: The polyp-cancer sequence in the large bowel,” *Proceedings of the Royal Society of Medicine*, vol. 67, no. 6 Pt 1, pp. 451–457, 1974. doi: <https://doi.org/10.1177/00359157740676P115>.
- [6] H. Kanavati, T. Tsuneki, and K. Tsunoda, “Deep learning models for polyp detection,” *Medical Image Analysis*, vol. 65, 101713, 2020. doi: <https://doi.org/10.1016/j.media.2020.101713>.
- [7] Y. -T. Chen and N. Ahmad, “Colorectal polyp detection and comparative evaluation based on deep learning approaches,” *IEEE*

- Access, vol. 11, pp. 135074–135089, 2023. doi: <https://doi.org/10.1109/ACCESS.2023.3337031>.
- [8] Li, J. Xu, P. Wang, et al., “A colonic polyps detection algorithm based on an improved YOLOv5s,” *Scientific Reports*, vol. 15, 13456, 2025. doi: <https://doi.org/10.1038/s41598-025-91467-1>
- [9] Lalinia, M., and S. Sahafi, “Colorectal polyp detection in colonoscopy images using YOLO-V8 network,” *Signal, Image and Video Processing*, vol. 18, pp. 2047–2058, 2024. doi: <https://doi.org/10.1007/s11760-023-02835-1>.
- [10] Zhang and S. Tian, “TS-YOLOv12: Accurate and real-time colorectal polyp detection in complex medical images,” *SSRN preprint*, 2024. doi: <https://dx.doi.org/10.2139/ssrn.5380011>.
- [11] D. Urban et al., “Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets,” *Gastroenterology*, vol. 158, no. 1, pp. 202–214, 2020. doi: 10.1053/j.gastro.2020.02.046.
- [12] D. Jha et al., “Real-time polyp detection, localization and segmentation in colonoscopy using deep learning,” *IEEE Access*, vol. 9, pp. 40496–40510, 2021. doi: <https://doi.org/10.1109/ACCESS.2021.3066281>.
- [13] A. R. Sahoo, S. S. Sahoo, and P. Chakraborty, "Polyp detection in colonoscopy images using YOLOv11," *arXiv preprint arXiv:2501.09051*, 2025.
- [14] M. Zhang, X. Li, and Y. Chen, "Real-time colorectal polyp detection via YOLOv10," *Proceedings of SPIE*, vol. 13561, pp. 135610O, 2025.
- [15] P. Ghose, "Improved polyp detection from colonoscopy images using fine-tuned YOLO-v5 model," *SpringerLink*, 2024.
- [16] J. Wan, "Polyp detection from colorectal images by using attentive YOLOv5," *Diagnostics*, vol. 11, no. 12, pp. 2264, 2021.

- [17] H. Borgli, "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Scientific Data*, vol. 7, p. 283, 2020.
- [18] M. Lalinia, "Colorectal polyp detection in colonoscopy images using YOLO-V8 network," *Signal, Image and Video Processing*, vol. 18, no. 4, pp. 1021–1034, 2024.
- [19] M. Gelu-Simeon, "Deep learning model applied to real-time delineation of polyps," *PMC*, 2025.
- [20] M. Spadaccini, "AI and polyp detection during colonoscopy," *PMC*, 2025.
- [21] Pacal and D. Karaboga, "A robust real-time deep learning based automatic polyp detection system," *Computers in Biology and Medicine*, vol. 134, 104519, 2021. doi: <https://doi.org/10.1016/j.compbioemed.2021.104519>.
- [22] H. A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham, "Polyp detection and segmentation using Mask R-CNN: does a deeper feature extractor CNN always perform better?" 2019 13th International Symposium on Medical Information and Communication Technology (ISMICT), Oslo, Norway, pp. 1–6, 2019. doi: <https://doi.org/10.1109/ISMICT.2019.8743694>
- [23] S. Wang, S. Lin, F. Sun, and X. Li, "Enhanced YOLOv8 with attention mechanisms for accurate detection of colorectal polyps," *Biomedical Signal Processing and Control*, vol. 99, p. 106942, 2025. doi: 10.1016/j.bspc.2024.106942.
- [24] K. Redjimi and M. Redjimi, "Polyps Detection Using Ultralytics YOLOv8," *Journal of Electrical Systems*, vol. 20, no. 3, pp. 12–20, 2024. [Online]. Available: <https://journal.esrgroups.org/jes/article/view/7650>

- [25] L. Zhang, Y. Chen, and F. Guo, “CRH-YOLO for precise and efficient detection of gastrointestinal polyps,” *Frontiers in Medicine*, vol. 11, 2024. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/39627309>
- [26] S. Kumar, A. Verma, and R. Rao, “Generalized Polyp Detection from Colonoscopy Frames Using Proposed EDF-YOLO8 Network,” in *Proc. CaPTION 2024: Cancer Prevention, Detection, and Intervention*, Springer, 2024, pp. 143–154. doi: 10.1007/978-3-031-73376-5_12.
- [27] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2016). Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1), 142–158. <https://doi.org/10.1109/TPAMI.2015.2437384>.
- [28] M. ElKarazle et al., “Automatic colorectal polyp detection using YOLOv5,” *Sensors*, vol. 23, no. 3, 1225, 2023. doi: <https://doi.org/10.3390/s23031225>.
- [29] G. Jocher, A. Chaurasia, and J. Qiu, “YOLOv5 (Version 5.0),” Zenodo, 2020. doi: <https://doi.org/10.5281/zenodo.3908559>.
- [30] Ultralytics, “YOLOv8,” Zenodo, 2023. doi: <https://doi.org/10.5281/zenodo.13769834>.
- [31] M. Yaseen, “What is YOLOv9: An in-depth exploration of the internal features of the next-generation object detector,” arXiv preprint arXiv:2409.07813, 2024. doi: <https://doi.org/10.48550/arXiv.2409.07813>.
- [32] M. Mehta, C. Gunavathi, N. Mathews, and D. Ruby, “Can programmable gradient information enhance polyp segmentation?”

- IEEE Access, vol. 12, pp. 198816–198827, 2024. doi: <https://doi.org/10.1109/ACCESS.2024.3496999>.
- [33] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, “YOLOv10: Real-time end-to-end object detection,” arXiv preprint arXiv:2405.14458, 2024. doi: <https://doi.org/10.48550/arXiv.2405.14458>.
- [34] J. Mei and W. Zhu, “BGF-YOLOv10: Small object detection algorithm from unmanned aerial vehicle perspective based on improved YOLOv10,” Sensors, vol. 24, no. 21, 6911, 2024. doi: <https://doi.org/10.3390/s24216911>.
- [35] G. Zhao, K. Zhao, and Y. Zhang, “Research on object detection and recognition in remote sensing images based on YOLOv11,” Scientific Reports, vol. 15, no. 1, 96314, 2025. doi: <https://doi.org/10.1038/s41598-025-96314-x>.
- [36] L. He, Y. Zhou, L. Liu, and J. Ma, “Research and application of YOLOv11-based object segmentation in intelligent recognition at construction sites,” Buildings, vol. 14, no. 12, 3777, 2024. doi: <https://doi.org/10.3390/buildings14123777>.
- [37] Y. Tian, Q. Ye, and D. Doermann, “YOLOv12: Attention-centric real-time object detectors,” arXiv preprint arXiv:2502.12524, 2025. doi: <https://doi.org/10.48550/arXiv.2502.12524>.
- [38] M. Yamada, S. Hirota, T. Kudo, Y. Kono, S. Tanaka, M. Fujishiro, and K. Otani, “Artificial intelligence for detection of colorectal neoplasia during colonoscopy,” Scientific Reports, vol. 9, 20254, 2019. doi: <https://doi.org/10.1038/s41598-019-50857-z>.
- [39] S. Ali, P. H. Smedsrød, M. A. Riegler, P. Halvorsen, and D. Johansen, “An objective comparison of deep learning methods for polyp

- detection and segmentation,” IEEE Reviews in Biomedical Engineering, vol. 14, pp. 206–221, 2021. doi: <https://doi.org/10.1109/RBME.2021.3062865>.
- [40] Y. Ou, Z. Yan, S. Liu, and Y. Li, “Polyp detection with improved YOLO models,” 2021 6th International Conference on Intelligent Computing and Information Processing (ICICIP), pp. 370–374, 2021. doi: <https://doi.org/10.1109/ICIBA52610.2021.9688145>.
- [41] B. Reddy, M. M. Kumar, R. Gowda, and J. P. Singh, “Colorectal polyp detection using YOLOv5 and DeepSort,” 2022 4th International Conference on Power, Control and Electrical Systems (PCEMS), pp. 1–6, 2022. doi: <https://doi.org/10.1109/PCEMS55161.2022.9807988>.
- [42] R. Krenzer et al., “Polyp detection benchmarks,” Journal of Imaging, vol. 9, no. 2, 26, 2023. doi: <https://doi.org/10.3390/jimaging9020026>.
- [43] O. Urban, T. M. Berzin, R. Shinozaki, S. Shinozaki, J. Oh, T. N. Tan, and T. N. Sano, “Deep learning-based real-time detection of colorectal polyps during colonoscopy,” Gastroenterology, vol. 155, no. 3, pp. 1069–1078, 2018. doi: <https://doi.org/10.1053/j.gastro.2018.06.037>.
- [44] J. Bernal et al., “Comparative validation of polyp detection methods in colonoscopy,” IEEE Transactions on Medical Imaging, vol. 36, no. 6, pp. 1231–1243, 2017. doi: <https://doi.org/10.1109/TMI.2017.2664042>.
- [45] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: unified, real-time object detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788, 2016. doi: <https://doi.org/10.1109/CVPR.2016.91>.
- [46] H. Kanavati, T. Tsuneki, and K. Tsunoda, “Deep learning models for polyp detection,” Medical Image Analysis, vol. 65, 101713, 2020. doi: <https://doi.org/10.1016/j.media.2020.101713>.

- [47] F. Vazquez, J. A. Nunez, X. Fu, P. Gu, and B. Fu, “Exploring Transfer Learning for Deep Learning Polyp Detection in Colonoscopy Images Using YOLOv8,” *arXiv preprint*, arXiv:2502.00133, 2025.
- [48] M. Lalinia and A. Sahafi, “Colorectal polyp detection in colonoscopy images using YOLO-V8 network,” *Signal, Image and Video Processing*, vol. 18, pp. 2047–2058, 2024. doi: 10.1007/s11760-023-02835-1.
- [49] Rainio, O., Teuho, J., & Klén, R. (2024). *Evaluation metrics and statistical tests for machine learning*. *Scientific Reports*, 14(1), 15724. DOI: 10.1038/s41598-024-66611-y [PubMed+2PMC+2](#)
- [50] Takahashi, K., Yamamoto, K., Kuchiba, A., & Koyama, T. (2021). Confidence interval for micro-averaged F_1 and macro-averaged F_1 scores. *Applied Intelligence*, 52(5), 4961–4972. DOI: 10.1007/s10489-021-02635-5 [PMC](#)
- [51] Shang, H., Langlois, J.-M., Tsoutsouliklis, K., & Kang, C. (2023). Precision/Recall on Imbalanced Test Data. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics* (PMLR, Vol. 206, pp. 9879–9891). DOI: 10.48550/arXiv.2301.04216 (or use the PMLR link)
- [52] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, “Kvasir-SEG: A segmented polyp dataset,” *Proceedings of the International Conference on Multimedia Modeling (MMM)*, pp. 451–462, 2020. DOI: https://doi.org/10.1007/978-3-030-37734-2_42