IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 October 2023, accepted 20 November 2023, date of publication 28 November 2023,
date of current version 5 December 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3337031*

RESEARCH ARTICLE

# Colorectal Polyp Detection and Comparative Evaluation Based on Deep Learning Approaches

**YAO-TIEN CHEN AND NISAR AHMAD**
International Ph.D. Program in Innovative Technology of Biomedical Engineering and Medical Devices, Ming Chi University of Technology, New Taipei City 243303, Taiwan

Corresponding author: Nisar Ahmad (d101gp006@mail2.mcut.edu.tw)

This work was supported by Ming Chi University of Technology, New Taipei City, Taiwan under Grant VK005-1G00-111.

**ABSTRACT** Colorectal cancer has been one of the leading causes of mortality over the past decade, and colorectal polyps are the leading cause of this disease. Conventional polyp detection techniques are insufficient for proper detection; thus, an efficient detection method is indispensable. In this study, we collected colorectal images from a hospital in Taiwan, annotated the ground truth of polyp locations, and integrated them with a public dataset to create a colonoscopy dataset. Data augmentation techniques are further used to increase the training dataset's diversity to improve the models' detection performance. By developing the comparison system based on the recent state-of-the-art methods (i.e., FasterRCNN, SSD, YOLOv3, and YOLOv4), we compared the measurement metrics and statistically analyzed the performance of the models to identify the significant statistical difference in models' performance. Moreover, we developed and integrated an error handling mechanism with each model to discard the false and null predictions. Finally, our model comparison system selects and proposes the best performing deep learning model to detect and classify colorectal polyps. We expect that the proposed model will accurately locate and classify different types of polyps. Eventually, this approach will ensure a valuable medical aid model.

**INDEX TERMS** Colorectal cancer (CRC), colorectal polyps, polyp detection, deep learning, data augmentation, error handling.

## I. INTRODUCTION

Colorectal cancer (CRC) is one of the most common cancers worldwide in the gastrointestinal tract. It is well known that intestinal polyps increase the possibility of intestinal cancer. Therefore, we need to remove the polyps by surgery to reduce the risk of cancer when detecting them in our intestines and diagnosed by a doctor as a precancerous lesion of colorectal cancer. The CRC patients show a rapid increase trend. According to the American Cancer Society, CRC is the third most common cancer and the second most common cause of cancer deaths in the United States [1]. In Taiwan, the National Health Administration of the Ministry of Health and Welfare reported in 2021 [2] that CRC's annual incidence and mortality rank second and third in all cancer incidence and mortality, respectively.

Polyps are the abnormal growth of tissues that look like tiny mushroom stalks and are most prevalent in the colon. Polyps are benign or harmless but can become malignant if not detected in early stage. Polyps can be classified as hyperplastic (Hp) and Adenomatous (Ad) [3], [4].

In conventional colonoscopy examination, doctors must rely on a series of medical images to speculate the location and shape of polyps in the patient's large intestine (colon) and use this as the basis for diagnosis and treatment. During the colonoscopy, a colonoscope, a flexible tube equipped with a tiny camera at its tip, is inserted into the colon. Polyps can be removed during the procedure if necessary [5]. However, the colonoscopy examination is time-consuming and highly labor-intensive. It is common for polyps to be overlooked during a colonoscopy procedure due to their small size and visual characteristics [6]. Moreover, the endoscopist may make an error in diagnosis due to eyestrain or lack of concentration. Computer science-based methods have been proposed to assist doctors during colonoscopy to address the

The associate editor coordinating the review of this manuscript and approving it for publication was Santosh Kumar.

mentioned issues. These methods use computer technology to detect polyps called computer-aided detection (CAD). According to this, most of these methods and systems are only research-based and not adequately developed for clinical applications [7]. Therefore, based on these factors, we continuously develop advanced CAD incorporating doctor-assisted diagnosis for colon polyp image as an effective medical aid.

When delving into the field of medical diagnosis, AI-based systems can prove advantageous by replicating the functions of the human brain in simpler tasks and formulating innovative solutions for more complex ones [8]. Artificial Intelligence (AI) significantly reduces errors during colonoscopy by highlighting the specific polyp region where the doctor could focus [9]. The rise of AI and machine learning has laid the basis for profound impact in many areas. The development of deep learning provides a technical basis for computer-aided decision-making in medical imaging [10]. AI-based colonoscopy systems can potentially improve the accuracy of lesion detection during clinical procedures. Most applications are based on computer vision to analyze videos and images of the gastrointestinal tract to detect and classify polyps [11].

Object detection is a computer vision technique that predicts objects in an image/video and points out the presence of objects with bounding boxes. It refers to identifying and localizing objects in an image/video that belong to a predefined set of classes [12]. Generally, object detection algorithms can be classified into two categories: one-stage detection models and two-stage detection models. One-stage detection model refers to those that skip the region proposal stage of two-stage models and run detection directly over a dense sampling of locations. These types of models usually have high inference. By contrast, the two-stage model works in two phases: object detection and classification. The basic principle in the two-stage model is that the model first proposes a set of regions of interest by select search or regional proposal network (RPN). The proposed regions are sparse as the potential bounding box candidates can be infinite; the classifier will only process the region candidates.

## II. RELATED WORK

Recently, 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. Pacal et al. [13] developed a modified YOLO (You Only Look Once) algorithm-based automatic polyp detection system by making architecture changes in the original YOLOv4. To enhance the model efficiency, preprocessing and post-processing techniques were adopted. They implemented data augmentation techniques for preprocessing, including flip, rotate, shear, hue, crop, and mosaic operations. NVIDIA TensorRT, a C++ library for high-performance inference on NVIDIA GPUs and deep learning accelerators, was used as a post-processing technique. After implementing architecture changes, their model achieved a higher detection performance.

To prevent overfitting and fine-tuning parameters, deep learning models require larger datasets to reach their full potential. Li et al. [14] built a polyp dataset collecting and integrating all publicly available datasets with the University of Kansas Medical Centre dataset. The images in the dataset contain polyps from different stages and represent different types of polyps. To generate a benchmark dataset for polyp detection, each image was labeled with accurate polyp locations and categories. Using the developed dataset, they evaluated and compared the performance of the state-of-the-art deep learning models for polyp detection and classification. The experiments demonstrated that deep CNN models are promising in CRC screening. This work of constructing polyp benchmark datasets can serve as a baseline for future polyp detection and classification research.

Although many algorithms have been developed to enhance the efficiency of polyp detection, the colon polyp miss detection rate is still high. Further, few are suitable for real-time detection due to their limited computing power; thus, a proposed method [15] used a real-time colonoscopy with the CNN transfer learning approach to solve the problems mentioned above. Pacal et al. [16] proposed another method integrating Cross Stage Partial Network (CSPNet) into the YOLOv3 and YOLOv4 object detection algorithms for real-time polyp detection. Data augmentation techniques and transfer learning were then utilized to improve the performance of polyp detection. The study uses the same data augmentation techniques as in [14]. To further improve the performance of polyp detection using negative samples, the Leaky ReLU and Mish activation functions were substituted by the Sigmoid-weighted Linear Unit (SiLU) activation functions, and Complete Intersection over Union (CIoU) was used as the loss function to provide fast convergence rate in bounding box regression.

Hoang et al. [17] presented a study based on capsule endoscopy for real-time detection of polyps. The magnetic capsule is remotely controlled using the electromagnetic actuation system (EMA) with position recognition and active locomotion. For real-time polyp detection, YOLOv3 was integrated with the system. In YOLOv3, on top of Darknet53, 53 more layers were added for proper object detection, making it a 106-layer convolutional neural network.

Nogueira-Rodriguez et al. [18] also proposed a deep learning model for real-time detection based on a pre-trained YOLOv3 architecture and a post-processing step based on an object-tracking algorithm to reduce false positives. The developed polyp detection method replaced the original class of object in the YOLOv3 model (pre-trained on PASCAL VOC) with a single polyp class; data augmentation techniques were then implemented for better polyp detection. Considering the model's ability to rapidly process every frame in the video and the corresponding results, the authors believe the developed model is valid and can be tested in a real-time environment and integrated into a CAD system.

Using AI to imitate human thinking and apply human intelligence to solve problems, such as deep learning-based methods, can assist doctors in making better diagnoses and accurate examinations of disease symptoms. Quan et al. [19] highlighted the importance of an AI-based polyp detection system. They presented a clinical evaluation of a CAD system based on the Single Shot Detector (SSD) model to detect neoplastic polyps. Elective colonoscopy based on CAD was performed on 300 patients, and the results were compared with non-CAD-based colonoscopy. The CAD-based polyp detection system was observed to have a greater polyp detection rate than those without CAD systems.

Some literature noted that the chances of missed polyps by colonoscopy remain high due to the limitations of diagnostic techniques and data analysis methods. Several approaches can be adopted to overcome these disadvantages, which include (1) expanding the polyp database being used; (2) designing a preprocessing procedure for the image based on polyp-specific features; and (3) improving the original deep learning model architecture. To improve the performance of the automatic polyp detection system, Qian et al. [20] expanded the training dataset using a Generative Adversarial Network (GAN) and modified the architecture of YOLOv4 using dilated convolution. The dataset was first expanded with a GAN, which generates realistic polyp-based images to add more annotated data. These images were later combined with the original dataset. Compared to the baseline models, the proposed model showed better performance regarding average precision (AP) and detection rate (FPS). Since the acquired image data may be of poor quality, such as high noise, low contrast differences, or specular reflections, preprocessing was suggested to obtain processed images more suitable for post-processing.

Aside from the above studies, several deliberations are being done to improve the methods for polyp detection further. One research [21] provided a benchmark for polyp detection, localization, and colonoscopy segmentation on recent state-of-the-art deep learning algorithms, including FasterRCNN, RetinaNet, YOLOv3+SPP, YOLOv4, and EfficientDet. It compared the execution performances and the differences between various algorithms on variable polyp sizes and image resolutions. In addition, the author also proposed the ColonSegNet to achieve a better trade-off between an average precision and mean IoU and the fastest detection and localization task. One of the models used in the study was EfficientDet, which is based on EfficientNet backbone architecture. Other models used were Faster R-CNN, YOLOv3, and YOLOv4. The benchmarking of these methods for real-time polyp detection was also presented.

Qian et al. [22] also presented an enhanced FasterRCNN-based system used for polyp detection by using preprocessed data. Specular reflections are one of the causes of false detection of polyps. Therefore, reflective points caused by specular reflections were first removed from the images before model training. Fine-tuning of VGG16 architecture was then performed based on the specific problem. The updated polyp detection model showed better results compared to the original model.

The remaining sections of this paper are organized as follows: Section III introduces the material and methods; Section IV explains the concepts of the proposed approach, which is used to select the proper model; Section V presents the experiments of the polyp detection; Section VI discusses the problems encountered in object detection and demonstrates how to solve them; finally, Section VII reports the conclusions of this study.

## III. MATERIAL AND METHODS

This section contains detailed information about datasets and data augmentation and briefly introduces four object detection and classification models used in this study.

### A. DATASETS

Two major datasets were employed to train and test the models: colorectal images from Chang Gung Memorial Hospital (CGMH) [23] and polyp dataset from Harvard Dataverse [24]. We first obtained colorectal images from CGMH in three sets (as shown in Table 1) based on the available data. These were refined and combined into a single CGMH dataset of 5,773 images. These images were acquired by professional radiologists during colonoscopy. Initially, the dataset comprised 6,436 colorectal images based on the three obtained sets. Later, the datasets were refined by removing blur and out-of-focus images, noisy images, and images having polyps texture the same as the colonic wall to ensure that models could be trained more accurately. The dimension of each image in this dataset is $640 \times 480$. Similarly, polyps with a tiny pixel size were also removed as they led to detection errors [12]. Figs. 1 to 3 show the samples of polyp images from CGMH with the mentioned issues.

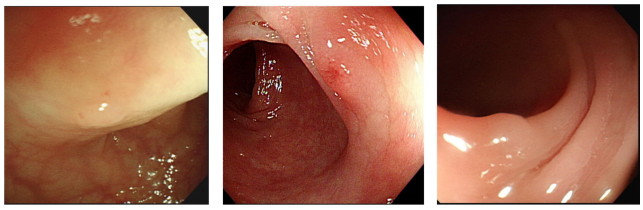


**FIGURE 1.** Polyp texture same as colonic wall.

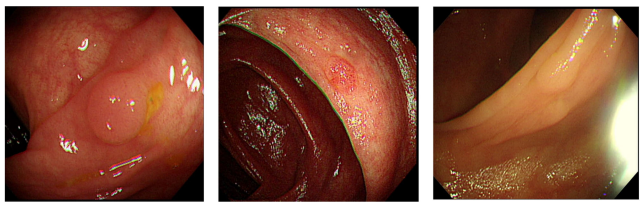


**FIGURE 2.** Polyp images with high noise.

The database, including the class and bounding boxes of the polyps, was annotated with the assistance of endoscopic

**FIGURE 3.** Smaller polyp size.

doctors. All images in the datasets were de-identified without revealing the patient information. Therefore, the CGMH dataset was annotated for ground truth (GT) using the LabelImg tool by generating XML and TXT files for each image containing polyp dimensions, such as top left, top right, bottom left, and bottom right positions. Samples from the CGMH colonoscopy dataset and Harvard Dataverse of colonoscopy are shown in Figs. 4 and 5, respectively. All images contain a single polyp, either an adenomatous or hyperplastic polyp, which professional radiologists took during colonoscopy; the image format is tif.

**TABLE 1.** Details of colonoscopy dataset obtained from CGMH.

| Image Datasets from CGMH | | | |
|---|---|---|---|
| Set1 | Set2 | Set3 | Total Images |
| 2636 | 2302 | 1496 | 6436 |
| Ad / Hp (Refined) | Ad / Hp (Refined) | Ad / Hp (Refined) | 5773 |
| 2448 | 1979 | 1346 | |

The second dataset of colorectal images was obtained from a publicly available polyp dataset on Harvard Dataverse. This dataset contains 7,150 images, each of which has been annotated. The dimension of polyp images in this dataset is 384 × 288, and the image format is jpg. Technical details about the two experimental datasets used in this study are shown in Table 2. These two datasets were combined to create a dataset of 12,923 images.

The more diverse the dataset is, the more likely the model will learn the meaningful features for object detection. The detection performance of the model is highly dependent on its training dataset. Therefore, these two datasets were combined to enhance the input dataset for high learning and improve the models' prediction capabilities. We performed a dataset split of 80% and 20% for training and testing, respectively. The training dataset frames were later shuffled before the training of object detection models.

**TABLE 2.** Information about colonoscopy datasets used in this study.

| Polyp Dataset | Format | No. of Images | Resolution | Available |
|---|---|---|---|---|
| Harvard Dataverse | JPG | 7150 | 384 x 288 | Link |
| CGMH | TIF | 5773 | 640 x 480 | -- |

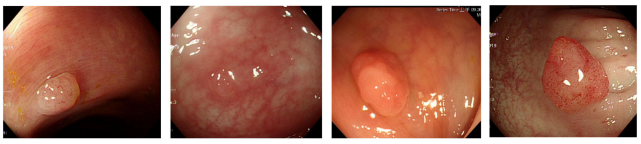


**FIGURE 4.** Four colonoscopy sample images from CGMH.

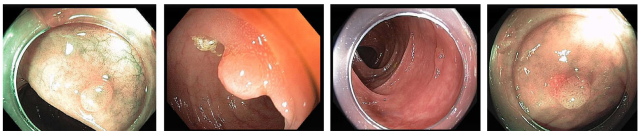


**FIGURE 5.** Colonoscopy sample images from Harvard Dataverse of Colonoscopy.

### B. DATA AUGMENTATION

Data augmentation is a technique used to enhance the dataset's size and increase the training dataset's diversity. Data augmentation is one of the most efficient data preprocessing techniques that can generate promising results [25]. Especially since deep learning-based methods need a vast amount of data for proper detection. Fig. 6 shows some commonly used techniques such as resize, rotation, translation, scale, reflection, shear, cropping, and color jitter. In this study, augmented images are little transformed versions of the original image which are often undetectable to the human visual system [26]. The details and parameters are described as follows. The magnification factor in resize is selected as 0.1. The uniform rotation angle of images is randomly selected from −45° to 45°. The translation is applied to transform image coordinates. Horizontal/vertical translations are applied with a uniform random range from −50 to 50. In scaling, the image was shrunk or enlarged. The scale factor was randomly selected from 1.2 to 1.5. X and Y horizontal reflections were randomly applied with uniform probability. The horizontal and vertical shear in random degrees was applied to a range of −30° to 30°, and the images were cropped to a resolution of 200 × 200.

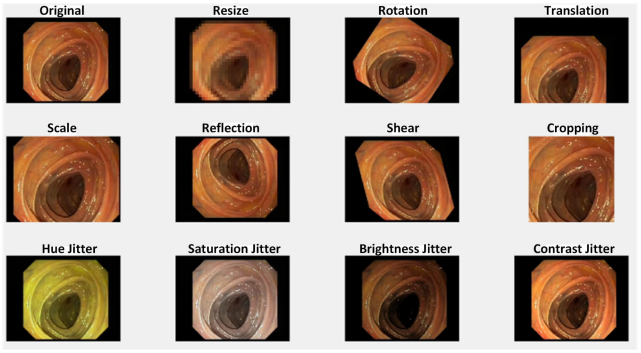



**FIGURE 6.** Sample images for data augmentation techniques.

In our experiments, three augmentation methods, random horizontal flip, random scaling, and color jitter, were employed. The data augmentation techniques are presented

in Algorithm 1. Applying these data augmentation techniques enhanced the training dataset to 31,017 images. Random horizontal flip horizontally flips a given image. Random scaling refers to the image's resizing and making the scaling on the XY axis preserve the original ratio.

**Algorithm 1** Data Augmentation Techniques

**Input:** Labelled CGMH Dataset + Harvard Dataverse
**Function**DataAugmentation(Input)
**Repeat**
  Read each image info (no of images, no of color channels, size, bounding box, labels)
  **If**(no of channels == 3)
    Perform combined data augmentation using the color jitter function: Contrast, Hue, Saturation, Brightness
  **End**
  Randomly Flip image: Perform 2d-Affine transformation (where horizontal reflection is randomly applied)
  Randomly X/Y Scaling: Perform scaling of the image with scale range [1 ∼1.1]
  Apply Geometric transformation of the transformed image
  Warp the resultant image to control the output limits
  Obtain the boundary box information of the warped version in the output
  Update the labeling information
  **If**(all bounding boxes are removed after warping)
    return the original data
  **End**
**Until**(the last image is traversed in the input)
**End Function**

The obtained image that gets scaled outside the original boundary is clipped. Color jitter randomly changes an image's brightness, contrast, and saturation. These three methods were integrated into one image with a predetermined probability in our training data set.

Fig. 7 shows the data augmentation experiment. The images in the left column are the original images, and those in the three right columns are the augmented images. The range of optimization values for data augmentation parameters is shown in Table 3.

### C. MODELS FOR DETECTION AND CLASSIFICATION

Four state-of-the-art object detection models were implemented, and their performances were compared in this study. The following section briefly introduces the object detection models, i.e., FasterRCNN, SSD, YOLOv3, and YOLOv4.

#### 1) FASTERRCNN

FasterRCNN belongs to the RCNN family networks and operates in two stages, detection and classification. It is an updated form of the Fast RCNN model, utilizing a region proposal network instead of a slow selective search algorithm.

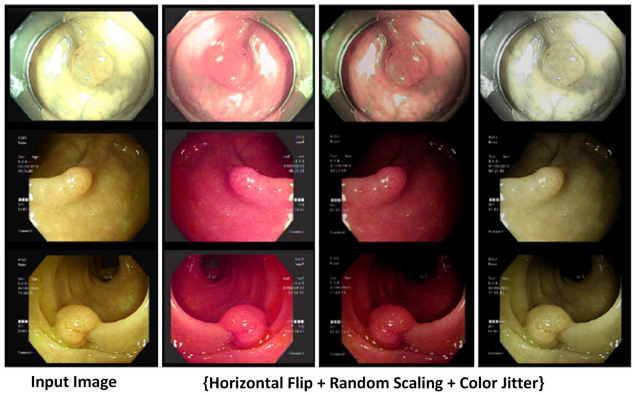



**FIGURE 7.** The training images obtained by the combination of horizontal flip, random scaling, and color jitter.

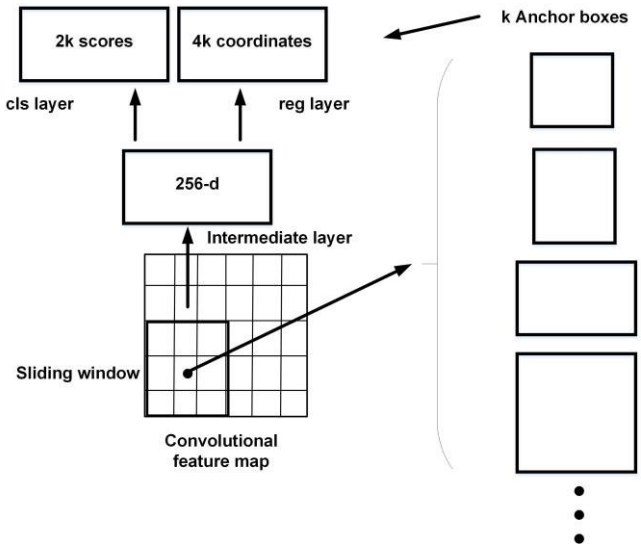



**FIGURE 8.** The RPN module.

This update enhances the detection rate of FasterRCNN. The Region Proposed Network (RPN) can be trained for better detection and to improve the model's performance. It uses the same base Convolutional Neural Network as Fast RCNN [13], [27].

RPN is the first module of FasterRCNN. It proposes regions for the second module, which are then employed for object detection. Regions in this research are only considered rectangular. The RPN takes an image as input to produce the rectangular object proposals. Each of these proposals contains an objectness score. The feature maps obtained from the last convolutional layer are used to generate region proposals by sliding the small network layer, which takes the n x n spatial window of the convolutional feature map. The working mechanism of RPN is presented in Fig. 8. The two fully connected layers, the box regression layer (reg) and box classification layer (cls) are fed with each sliding window mapped to a low dimensional feature. The low dimensional

**TABLE 3.** Data augmentation parameters optimization.

| Parameter | Resize | Rotation | Translation | Scale | Reflection | Shear | Cropping | Hue | Saturation | Brightness | Contrast |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Factor (*from*) | 0.1 | -45 | -50 | 1.2 | Random | -30 | 200 (X- dim) | 0.05 | 0.4 | -0.3 | 1.2 |
| Factor (*to*) | 0.1 | 45 | 50 | 1.5 | Random | 30 | 200 (Y-dim) | 0.15 | -0.1 | -0.1 | 1.4 |

features are 256-d for the Zeiler and Fergus (ZF) model and 255 for VGG with ReLU.

At each sliding window, multiple region proposals are predicted, and the maximum number of possible proposals for each sliding window is represented as k. The cls layer generates 2k scores, and the reg layer outputs 4k coordinates.

The 2k scores generated by the cls layer are used to estimate the probability of whether the object exists. The RPN gets the feature maps extracted by ResNet 101, the backbone network. These feature maps are also used by the classification module of the model. In RPN, the operations of calculating the bounding box locations and object probability score are performed. K anchor boxes are predicted by sliding windows for each location. To attain multi-scale learning, these anchor boxes are based on different sizes and are self-centered.

Fig. 9 shows the basic architecture of FasterRCNN.

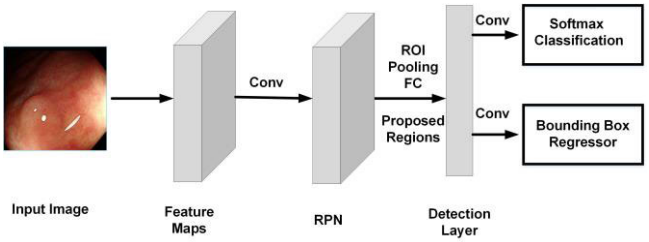



**FIGURE 9.** FasterRCNN architecture.

While performing tests on the PASCAL VOC dataset, K40 GPU, and VGG-16 backbone, the inference time is reduced to 198ms. This improvement is due to the RPN introduction. FasterRCNN is 10 times faster than the selective search. This model achieved high computation speed by reducing the computational time from 1,510ms to 10ms and can perform the computation of 5fps [19].

### 2) SSD

Single Shot Detector (SSD) is a single-stage detector, unlike FasterRCNN. It is based on a simple architecture and utilizes various sizes of feature maps to generate predictions. In the initial stage, convolution operation is performed to extract features from input images [25]. The small convolutional filters are applied to feature maps to predict category scores for bounding boxes. The predictions of different scales from feature maps are generated to achieve high accuracy in terms of detection. For training purposes, SSD requires input images with ground truth boxes for each object. The SSD framework is shown in Fig. 10. Following the convolution style, a small set of default boxes are evaluated with different aspect ratios in different feature maps at each location. For each default box shape, offsets and confidence scores for all categories are predicted. During the training process, the ground truth boxes are matched with default boxes. One example is shown in Fig. 10; a default box is matched with a polyp and is treated as positive.

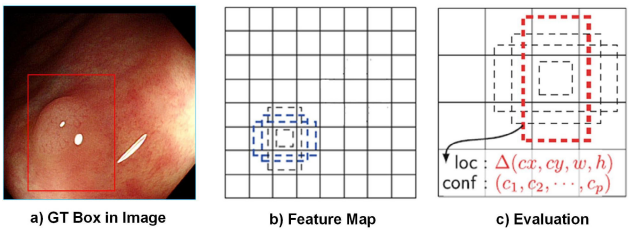



**FIGURE 10.** SSD framework. (a) Input image with ground truth box. (b) Feature map for the input image. (c) Evaluation for prediction.

The mentioned design features play an essential role in proper model training with high accuracy, improving the speed and accuracy and providing a tradeoff of low-resolution input images. The performance of SSD is evaluated and compared with other models in terms of timing and accuracy after training on PASCAL VOC, ILSVRC, and COCO datasets.

SSD is an efficient model that has laid the foundation for other highly efficient object detection models. It uses a feed-forward approach that generates bounding boxes with scores for specific classes. The initial stage of the model is based on the VGG-16 network. The initial network layers are composed of standard architecture used for high-quality image classification. Then, the auxiliary structure in the network produces detections based on multiscale feature maps, convolutional detection predictors, and default boxes and aspect ratios. The SSD architecture is presented in Fig. 11.

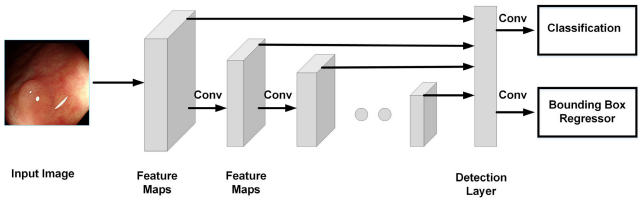



**FIGURE 11.** SSD architecture.

The next stage is based on the auxiliary structure to produce detections that have multiscale feature maps and convolutional predictions for detections, aspect ratios, and default boxes.

Like the YOLO series, SSD divides the input image in m x n grids. For each grid, where the kernel is applied, each class score and bounding box dimension is generated. SSD can achieve a reasonable detection rate with high accuracy [28].

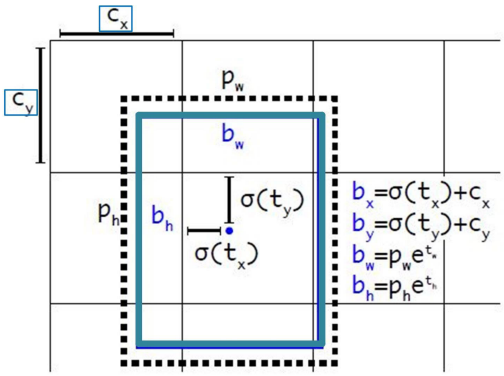



**FIGURE 12.** Bounding boxes with dimension priors and location prediction.

### 3) YOLOV3

YOLOv3 is an improvement of YOLO in terms of model size and accuracy. YOLOv3 makes bounding box predictions by using dimension clusters as anchor boxes. For each bounding box, the network predicts 4 coordinates, such as $t_x$, $t_y$, $t_w$, and $t_h$. The offset for the cell from the top left corner is cx and cy, and the width and height for the bounding box are $p_w$ and $p_h$, respectively. Fig. 12 demonstrates how the predictions can be obtained. The sigmoid function is an activation function in an S-shaped curve and is particularly useful for models where the output needs to be in the range of 0 to 1, making it suitable for binary classification problems. The logistic sigmoid function formula is given as,

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

In YOLOv3, logistic regression predicts the objectness score for each bounding box. If the bounding box prior overlaps the ground truth object by another bounding box prior, the objectness score should be 1. However, predictions are ignored if the bounding box prior overlaps a ground truth object greater than the threshold value.

For an input image with a dimension of 320 × 320, the YOLOv3 processes the image with an mAP of 28.2; it is three times faster than SSD. The significant design changes not present in the previous versions include a new backbone, multiple scale prediction, and updated loss function for class prediction. YOLOv3 architecture is shown in Fig. 13.

YOLOv3 uses a convolutional neural network CNN based on 53 convolutional layers known as Darknet53, an update of

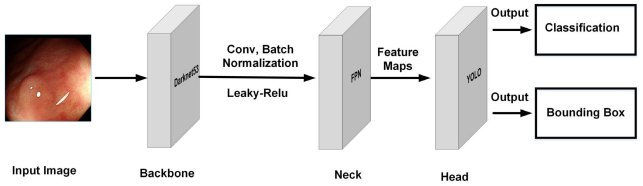



**FIGURE 13.** YOLOv3 architecture.

the Darknet 19 used in YOLOv2. Darknet53 uses 3 × 3 and 1 × 1 convolutional layers. Since the backbone is based on Darknet 53, the neck uses a Feature Pyramid Network (FPN) to generate feature maps for the head for object detection and classification. YOLOv3 uses three detection heads to process the image at different compressions. Its object detection accuracy is better than SSD, and has relatively high performance when it comes to small object detection; however, improvement is required on medium and large objects [29].

### 4) YOLOV4

YOLOv4 is one of the most efficient and powerful models in the YOLO series. To achieve improved detection accuracy, the bag of freebies and bag of specials are adopted. The main idea is to achieve better accuracy of the model without an increase in inference cost. The object detection models implement the ''bag of freebies'' concept as data augmentation. Data augmentation enhances the robustness of the model by training the model on images obtained from different environments. The two major data augmentation techniques used in image processing are geometric distortion and photometric distortion. Different geometric distortion operations are performed on images, such as cropping, scaling, flipping, and scaling. Photometric distortions include brightness, saturation, contrast, and image noise adjustments.

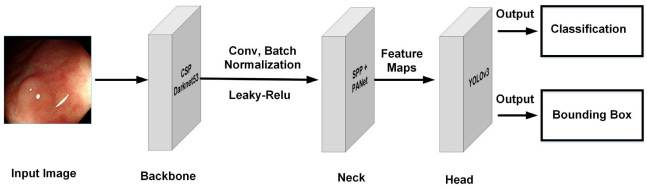



**FIGURE 14.** YOLOv4 architecture. Backbone {CSPDarknet53}, Neck {SPP, PAN}, Head {YOLOv3}.

The optimal classification model is not always suitable for detection. Therefore, the significant requirements for the optimum functionality of the detector are high input resolution, a large number of layers to adjust the high input, and more parameters for the model's high capacity for multiple object detection in a single image [30]. YOLOv4 model architecture is shown in Fig. 14.

As a backbone, CSPDarknet53 is used with an additional SPP block to increase the receptive field. CSPDarknet53 is a Darknet version that uses Cross Stage Partial (CSP) connections [31]. It also separates the most critical context features

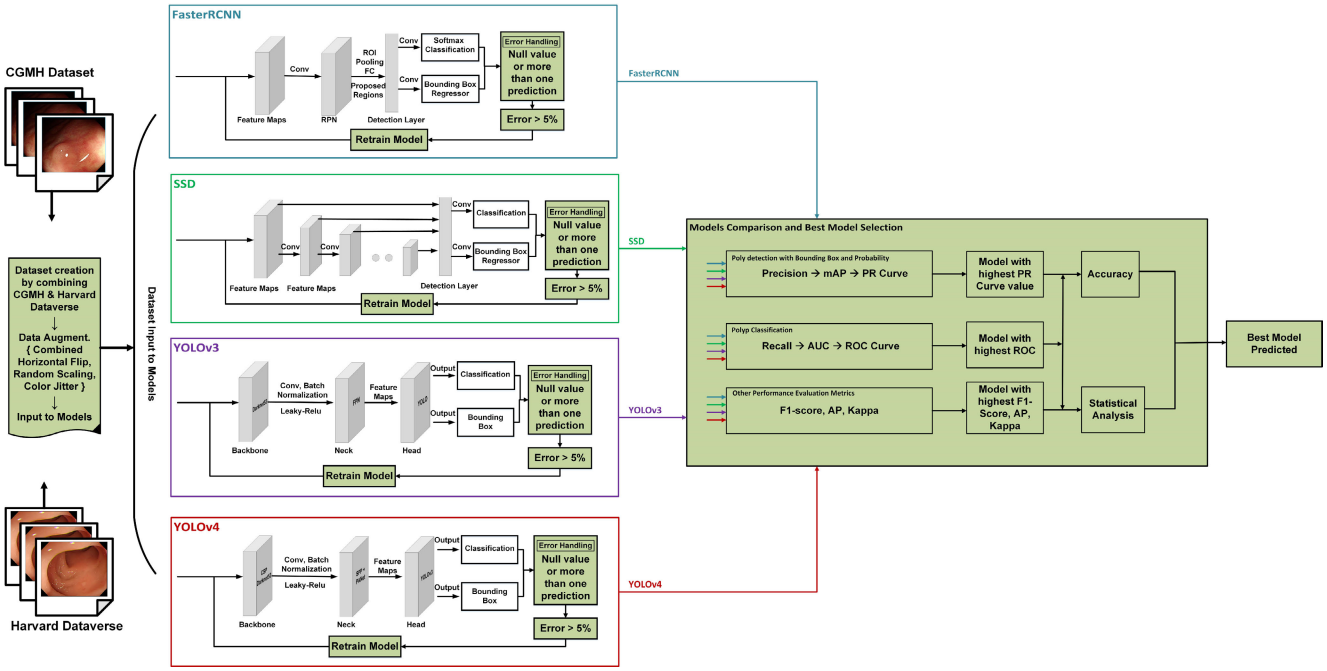



**FIGURE 15.** A detailed overview of the system architecture, dataset creation, data augmentation, models architecture and integration with error handling modules, polyp detection and classification metrics analysis and statistical analysis module.

and does not affect the processing speed of the network. YOLOv3 head and PANet path aggregation are used as the neck in the YOLOv4 model architecture.

## IV. PROPOSED APPROACH

In this section, the proposed approach of our study is presented to identify the best performing object detection model. The models' performances are compared using different metrics, and the best performing model is selected. Fig. 15 shows the complete system architecture, composed of dataset creation, data augmentation, models integration, and comparison system to pick the best performing model for polyp detection and classification. We performed a detailed performance comparison and analysis using different performance metrics and statistical analysis to propose the best performing model for adenomatous and hyperplastic colorectal polyp detection and classification. The major contributions in this study are highlighted in green color.

Since object detection models are trained and tested separately, we created a new dataset by combining CGMH with Harvard Dataverse and provided these colorectal images to the models as input.

Noisy images are removed from the dataset before providing the colonoscopy images to the models. After refining and combining the two datasets, data augmentation techniques are applied to the combined dataset. In the initial stage, to find the most suitable deep learning model among many state-of-the-art methods, the models (FasterRCNN, SSD, YOLOv3, and YOLOv4) separately extract features for training. Then, polyp detection and polyp classification (adenomatous or hyperplastic) are performed. A confidence score with a bounding box is evaluated to inspect the polyp accurately. In the central block of the system architecture, all four models are shown to depict their main features.

The ROC curve is a widely used performance measurement for classification problems at various threshold settings. It is a probability curve, and the AUC represents the degree or measure of separability. The AUC indicates how well the model can distinguish between classes, with higher values indicating better performance. The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis, and FPR is on the x-axis.

On the other hand, the PR curve shows the tradeoff between precision and recall for different threshold values. It provides insights into the relationship between precision and recall, with precision values on the y-axis and recall values on the x-axis. In other words, the PR curve represents TP/(TP+FP) on the y-axis and TP/(TP+FN) on the x-axis.

Although the PR and ROC curves are similar, they serve different purposes. The ROC curve assesses the model's performance without considering class imbalance, while the PR curve is beneficial for evaluating two-class classification algorithms. Both curves offer valuable information about a model's performance and assist analysts in making informed decisions based on their specific objectives and needs.

When drawing ROC curves and calculating the AUC, our approach incorporates specific improvements to address potential problems in object detection results. Object detection outcomes often fall into two scenarios regardless of the model used. First, there may be instances where empty objects are detected due to parameter or threshold settings.

Second, multiple objects might be detected within a single instance. To account for these situations, our proposed method introduces specific criteria during evaluation.

In our method, we set a maximum threshold of 5% for empty object detections within the test dataset. The corresponding records are excluded from the AUC calculation if an empty object is detected. Furthermore, we consider only the object with the highest detection score for AUC calculation when multiple objects are detected.

We employ two main principles to determine the model with the best performance. Firstly, the model should possess highly accurate classification capabilities. Secondly, it should exhibit superior object detection capabilities. The selected model will have the highest AUC value and mAP corresponding to the performance metrics.

The polyp detection results of the models are also discussed with colorectal doctors to verify the detected results for model performance analysis. Based on the performance metrics, the best performing model is identified. For object detection, the mAP values are analyzed to pick the best detection model, and the best classification model for a specific class of polyp is chosen based on the highest AUC (Area Under the Curve) value. The overall classification accuracies of the models are analyzed to select the best classification model with the highest kappa value. Precision, recall, F1-score, AP, and kappa values are presented in Table 11. Class-wise accuracies and overall detection accuracies are shown in Table 8, Table 9, and Table 10.

## V. EXPERIMENTS

All the algorithms were implemented in Matlab programming language. Polyp images provided by CGMH and partial Harvard Dataverse of colonoscopy [20] were used as the training and testing data for the deep learning network. All experiments were executed on a computing system with NVIDIA GRID V100D-16Q GPU and Microsoft Windows 8 operating system for polyp detection and performance metrics evaluation.

We integrated the two datasets to fully utilize deep learning techniques for object detection. These datasets contain only two types of polyps, i.e., hyperplastic and adenomatous polyps. Training a model that can reliably distinguish them into different classes is vital because adenomatous polyps are generally considered precancerous lesions requiring resection, whereas hyperplastic polyps are not.

The experiments consisted of four parts: two-class polyp detection, classification, calculation of the performance metrics and comparison, and performance analysis based on statistical analysis. Detection of colorectal polyps is an essential task in colonoscopy. In addition, accurate polyp classification is required to diminish mortality due to colorectal cancer. Automatic detection of a polyp is, thus, a precious contribution to radiology and medical imaging.

To validate the suitability of different models in object detection, we first used the colonoscopy images from the Harvard Dataverse of colonoscopy to verify the ability of the models. Four different models, including FasterRCNN, SSD, YOLOv3, and YOLOv4, were used to examine and compare the effects of polyp detection. Figs. 16(a) to (h) show the results of adenomatous polyp detection and Figs. 17(a) to (h) show the results of hyperplastic polyp detection. It can be seen that YOLOv4 has the highest confidence score and extracted both different polyps with much more accuracy than the other models. In addition to detecting the polyps from Harvard Dataverse of colonoscopy, we further applied these methods to the colonoscopy image dataset provided by CGMH. The detection results for the adenomatous class from CGMH test images are shown in Fig. 18. Four models achieved better performance for adenomatous polyps, which are larger in size; hence, their shape and texture were easier to distinguish from the colonic wall. The results of CGMH hyperplastic polyp detection are presented in Fig. 19. The qualitative comparison showed that the YOLOv4 can detect the hyperplastic polyp with a more accurate confidence score than the other three models.

To understand the ability of object detection for adenomatous and hyperplastic polyps, the metrics containing precision, recall, F1-score, average precision, and kappa coefficient were used to examine the performance. For completeness, these metrics are explained as follows.

Precision is the measurement of the percentage of correct predictions. Correct predictions of polyps are significant in detecting and treating colorectal cancer. Precision is calculated using the following equation,

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

where True Positive (TP) is a correct prediction of the positive class, and False Positive (FP) is the incorrect prediction of the positive class.

Recall is used to measure the ratio of positive predictions which are correctly identified. It is formulated as,

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

where True Positive (TP) is the correct prediction of the positive class, and False Negative (FN) is the incorrect prediction of the negative class.

The accuracy of the model is calculated as,

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

The confusion matrix is used to measure the model performance by calculating the values of the performance matrices, i.e., precision, recall, F1-score, AP, and kappa values. All of these predictive analytics are calculated using Eqs. 4 to 7. The results obtained in this study are shown in Table 11.

F1-score calculates the test accuracy. It considers precision and recall to measure the model's performance using a relation.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

The average precision is measured as AUC. It measures a precision-recall (PR) curve into one scalar value between 0 and 1. When maximum precision values are dropped, the curve is sampled at recall values ($r1, r2, \ldots$).

When $p(r_i)$ drops, it is sampled, and AP is computed as the sum of rectangular blocks. The value for AP can be defined as

$$AP = \sum (r_{n+1} - r_n)\, p_{interp}\,(r_{n+1}) \tag{6}$$

and

$$p_{interp}\,(r_{n+1}) = max_{r^{\sim} \geq r_{n+1}} p\left(r^{\sim}\right) \tag{7}$$

The main purpose of interpolating the precision-recall curve is to minimize the effect of fluttering in the precision-recall curve. Both methods will diverge as interpolated points that do not cover the precision drop.

The kappa coefficient is an index calculated based on the confusion matrix to assess the model's classification accuracy [32]. The obtained kappa values for each model are presented in Table 11. A higher kappa value indicates a higher classification accuracy. To calculate the kappa value,

$$Kappa = \frac{p_0 - p_e}{1 - p_e} \tag{8}$$

where $p_o$ is the accuracy value, i.e., the sum of diagonal elements divided by the sum of total matrix elements; pe is the sum of products of the actual and corresponding predicted values divided by the total matrix elements.

### A. TWO-CLASS POLYP DETECTION AND CLASSIFICATION

Polyp detection models were set to detect and classify colorectal polyps into two classes, adenomatous and hyperplastic, with a confidence threshold value of 0.5. Models were trained on the combined dataset and tested separately on unseen images based on the two datasets. The detection results with the confidence scores are shown in Figs. 16 to 19. The detection results with a confidence value of over 0.5 are shown in each frame.

#### 1) DETECTION RESULTS ON THE HARVARD DATAVERSE

Figs. 16 and 17 show the detection and classification results for the two classes, i.e., adenomatous and hyperplastic when the object detection models were provided with test images from Harvard Dataverse.

The detection results demonstrated that the models accurately performed the polyp detection and classification within the set confidence scores.

To observe the detection and classification performance of the models, the same test images were provided for each model. As can be seen in Figs. 16 and 17, YOLOv4 had the highest confidence score for polyp detection. The detection and classification accuracies of the models were calculated to compare and identify the best performing model.

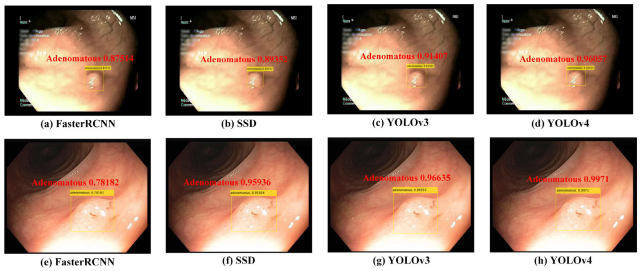



**FIGURE 16.** Detection results of FasterRCNN, SSD, YOLOv3, and YOLOv4 models for adenomatous polyp with the predicted classes and confidence scores. (a) and (e) using FasterRCNN, (b) and (f) using SSD, (c) and (g) using YOLOv3, and (d) and (h) using YOLOv4. Original images were downloaded from Harvard Dataverse of colonoscopy.

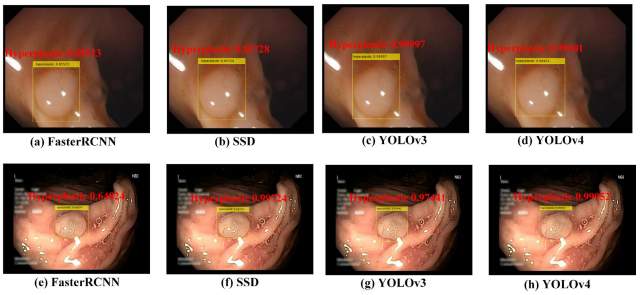



**FIGURE 17.** Detection results of FasterRCNN, SSD, YOLOv3, and YOLOv4 models for hyperplastic polyp with the predicted classes and confidence scores. (a) and (e) using FasterRCNN, (b) and (f) using SSD, (c) and (g) using YOLOv3, and (d) and (h) using YOLOv4. The original images were downloaded from Harvard Dataverse of colonoscopy.

#### 2) DETECTION RESULTS ON THE CGMH DATASET

Figs. 18 and 19 show the detection and classification results for two classes, adenomatous and hyperplastic when the models were provided with the test images from the CGMH dataset. Similar to the detection and classification results in Figs. 16 and 17, YOLOv4 obtained the highest confidence score for both classes, i.e., adenomatous and hyperplastic, on the CGMH dataset.

It can also be observed that on the CGMH test dataset, the models showed higher confidence scores than those on Harvard Dataverse due to higher resolution.

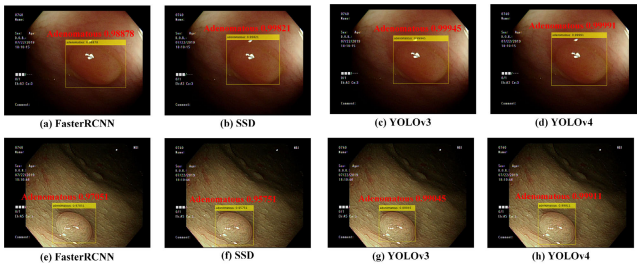



**FIGURE 18.** Detection results of FasterRCNN, SSD, YOLOv3, and YOLOv4 models for adenomatous polyp with the predicted classes and confidence scores. (a) and (e) using FasterRCNN, (b) and (f) using SSD, (c) and (g) using YOLOv3, and (d) and (h) using YOLOv4. The original images were obtained from CGMH.

### B. MODEL COMPARISON AND RESULTS ANALYSIS

The detection results for the two classes were compared and are presented in Table 11. Adenomatous polyps are generally easier to identify than hyperplastic polyps due to their larger

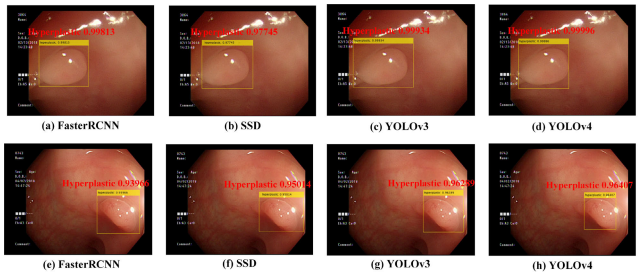



**FIGURE 19.** Detection results of FasterRCNN, SSD, YOLOv3, and YOLOv4 models for hyperplastic polyp with the predicted classes and confidence scores. (a) and (e) using FasterRCNN, (b) and (f) using SSD, (c) and (g) using YOLOv3, and (d) and (h) using YOLOv4. The original images were obtained from CGMH.

size and distinct visual features. As a result, almost all object detection models showed promising results for detecting adenomatous polyps. However, hyperplastic polyps present more of a challenge for object detection models due to their smaller size and subtle visual characteristics. These features make it more difficult for the model to distinguish hyperplastic polyps with similar visual features. Therefore, developing accurate object detection models for hyperplastic polyps requires more research and innovation.

From the detection results shown in Figs. 16 to 19, it can be seen that bounding boxes were precisely placed around the polyps, which indicates excellent model performance as it is of high importance during colonoscopy. As indicated in Table 11, YOLOv4 achieved the best detection performance in adenomatous and hyperplastic with mAP = 88%. Moreover, YOLOv4 outperformed all other models with a significant margin in the two-class classification of mean precision, mean recall, mean F1-score, and mAP. After YOLOv4, SSD yielded the second highest mean precision, mean recall, and mean F1-score except for mAP, where YOLOv3 performed second best with 77.4%. For adenomatous detection, SSD also performed second best in the recall, F1-score, and AP. YOLOv3 outperformed SSD in adenomatous precision with 87.5% and hyperplastic recall with 90.4%. Therefore, SSD ranked second compared with other detectors obtaining the second highest mean precision, mean recall, and mean F1-score. YOLOv4 obtained a kappa value of 0.828, making it almost a perfect agreement and taking the lead among the other object detection and classification models. SSD ranked second place with a kappa value of 0.628. FasterRCNN and YOLOv3 achieved kappa values of 0.402 and 0.546, respectively.

For training, 80% of the dataset based on CGMH and Harvard Dataverse was utilized, and the remaining 20% was used for testing based on the calculation, i.e., (7150+5773) x 20% = 2585. The confusion matrices for FasterRCNN, SSD, YOLOv3, and YOLOv4 are presented in Tables 4 to 7.

The overall accuracy and class-wise accuracy for FasterRCNN, SSD, and YOLOv3 were calculated and then compared with the accuracy of YOLOv4. The results are presented in Table 8, Table 9, and Table 10.

**TABLE 4.** Confusion matrix for FasterRCNN.

| N=2585 | Adenomatous | Hyperplastic | Sum |
|---|---|---|---|
| Adenomatous | 988 | 440 | 1428 |
| Hyperplastic | 331 | 826 | 1157 |
| Sum | 1319 | 1266 | 2585 |

**TABLE 5.** Confusion matrix for SSD.

| N=2585 | Adenomatous | Hyperplastic | Sum |
|---|---|---|---|
| Adenomatous | 1047 | 207 | 1254 |
| Hyperplastic | 272 | 1059 | 1331 |
| Sum | 1319 | 1266 | 2585 |

**TABLE 6.** Confusion matrix for YOLOv3.

| N=2585 | Adenomatous | Hyperplastic | Sum |
|---|---|---|---|
| Adenomatous | 853 | 122 | 975 |
| Hyperplastic | 466 | 1144 | 1610 |
| Sum | 1319 | 1266 | 2585 |

**TABLE 7.** Confusion matrix for YOLOv4.

| N=2585 | Adenomatous | Hyperplastic | Sum |
|---|---|---|---|
| Adenomatous | 1195 | 100 | 1295 |
| Hyperplastic | 124 | 1166 | 1290 |
| Sum | 1319 | 1266 | 2585 |

### C. STATISTICAL ANALYSIS

The class-wise polyp detection accuracies of FasterRCNN, SSD, and YOLOv3 were compared with the accuracy of YOLOv4. To observe the statistically significant difference between the accuracies, two proportion Z-test [33] was employed. The accuracies for each model were independently tested. The confidence level of 95% was selected with a significance level of $\alpha = 0.05$ (5%), making the critical Z value +/−1.96. The $H_0$ (null hypothesis) in this analysis (for each model comparison with YOLOv4) states that both the accuracies are the same; however, the $H_a$ (alternate hypothesis) states that the accuracies are not the same, and there is a significant difference between them. When we compared the accuracies of any of the above models with the YOLOv4, the null hypothesis was rejected if $Z < -1.96$ or $> 1.96$, and the null hypothesis was accepted if $-1.96 < Z < 1.96$, that is, there was no significant difference between the two accuracies. The Z value calculation formula is given as,

$$Z = \frac{\hat{p}1 - \hat{p}2}{\sqrt{\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{9}$$

where $\hat{p}_1$ is the first proportion, $\hat{p}_2$ is the second proportion, and $\hat{p}$is the overall proportion. The proportions were separately calculated for each class using TP and TN values from

the confusion matrices. The two proportion Z-test results are presented in Table 12.

**TABLE 8. Comparison of polyp detection accuracies for FasterRCNN and YOLOv4.**

| | FasterRCNN | | YOLOv4 | |
|---|---|---|---|---|
| **Polyp type** | Adenomatous | Hyperplastic | Adenomatous | Hyperplastic |
| **No. of images correctly identified** | 988 | 826 | 1195 | 1166 |
| **Class-wise Accuracy** | 74.9% | 65.2% | 90.5% | 92.1% |
| **Model Accuracy** | 70.1% | | 91.3% | |

**TABLE 9. Comparison of polyp detection accuracies for SSD and YOLOv4.**

| | SSD | | YOLOv4 | |
|---|---|---|---|---|
| **Polyp type** | Adenomatous | Hyperplastic | Adenomatous | Hyperplastic |
| **No. of images correctly identified** | 1047 | 1059 | 1195 | 1166 |
| **Class-wise Accuracy** | 79.3% | 83.6% | 90.5% | 92.1% |
| **Model Accuracy** | 81.4% | | 91.3% | |

As shown in Table 12, except for the calculated Z value between YOLOv3 and YOLOv4 for hyperplastic class accuracy, the Z values for both adenomatous and hyperplastic classes were higher than 1.96, which indicates that there is a significant difference in the accuracies of the evaluated models. From this statistical analysis and the above comparative results, we can conclude that the accuracy of YOLOv4 is higher than the other three evaluated models.

Two proportion Z test was performed to analyze the significant difference in accuracy. In adenomatous classification, the vast difference in accuracy of almost 26% between YOLOv4 and YOLOv3 was presented with a Z value of 15.89. This shows that YOLOv3 had the lowest adenomatous polyp detection and classification accuracy. In the hyperplastic polyp class, compared to YOLOv4, FasterRCNN performed with the lowest accuracy, with a difference of almost 27% and a Z value of 16.5.

This difference is 1% more than between the YOLOv3 and YOLOv4 adenomatous class detection accuracy, making the hyperplastic detection by FasterRCNN the overall lowest in this study. The YOLOv3 performed the second best (after YOLOv4) with an accuracy difference of only 1.8%. This difference is depicted in a Z score of 1.67. As the critical Z value for the positive region was 1.96, and the calculated Z value was less than 1.96, the null hypothesis was accepted; this means there is no significant difference between YOLOv3 and YOLOv4 accuracies. The YOLOv3 was the second best after YOLOv4 in hyperplastic detection.

**TABLE 10. Comparison of polyp detection accuracies for YOLOv3 and YOLOv4.**

| | YOLOv3 | | YOLOv4 | |
|---|---|---|---|---|
| **Polyp type** | Adenomatous | Hyperplastic | Adenomatous | Hyperplastic |
| **No. of images correctly identified** | 853 | 1143 | 1195 | 1166 |
| **Class-wise Accuracy** | 64.6% | 90.3% | 90.5% | 92.1% |
| **Model Accuracy** | 70.1% | | 91.3% | |

For adenomatous class detection, SSD came after YOLOv4 with an accuracy difference of 11.2%. The comparison of the models' accuracies is shown in Fig. 20.

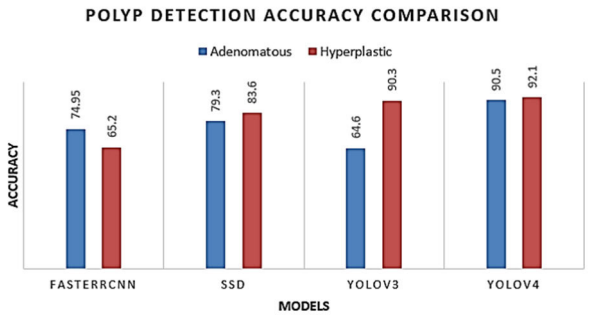



**FIGURE 20. Polyp detection accuracy comparison.**

The relevant Z value for this difference was 8.05, which is significant based on the hypothesis rule. However, in adenomatous detection, SSD was the second best performing model, with an accuracy of 79.3%. In the models' overall detection accuracies comparison, SSD was the second best after YOLOv4 with 81.4%; the difference was only 10%.

In summary, the results suggest that YOLOv4 is the most suitable deep learning model for detecting and classifying polyps from colonoscopic images since it achieved the highest performance in terms of accuracy, precision, recall, F1-score, AP, and kappa value. In the case of YOLOv4, it achieved a mAP = 88.0 and an AUC = 0.8709. Based on these results, we conclude that YOLOv4 is the most efficient and suitable model for polyp detection.AP, and kappa value.

**TABLE 11.** Results for two-class polyp detection.

| Model | Category | Precision | Recall | F1-score | AP | Kappa |
|---|---|---|---|---|---|---|
| **FasterRCNN** | Adenomatous | 69.2 | 74.9 | 71.9 | 72.5 | 0.402 |
| | Hyperplastic | 71.4 | 65.2 | 68.2 | 62.5 | |
| | Mean | 70.3 | 70.1 | 70.1 | 67.5 | |
| **SSD** | Adenomatous | 83.5 | 79.4 | 81.4 | 80.7 | 0.628 |
| | Hyperplastic | 79.6 | 83.6 | 81.6 | 72.2 | |
| | Mean | 81.6 | 81.5 | 81.5 | 76.5 | |
| **YOLOv3** | Adenomatous | 87.5 | 64.7 | 74.4 | 79.9 | 0.546 |
| | Hyperplastic | 71.1 | 90.4 | 79.6 | 74.8 | |
| | Mean | 79.3 | 77.6 | 77.0 | 77.4 | |
| **YOLOv4** | Adenomatous | 92.3 | 90.6 | 91.4 | 90.7 | 0.828 |
| | Hyperplastic | 90.4 | 92.1 | 91.2 | 85.3 | |
| | Mean | 91.4 | 91.4 | 91.3 | 88.0 | |

In the case of YOLOv4, it achieved a mAP = 88.0 and an AUC = 0.8709. Based on these results, we conclude that YOLOv4 is the most efficient and suitable model for polyp detection.

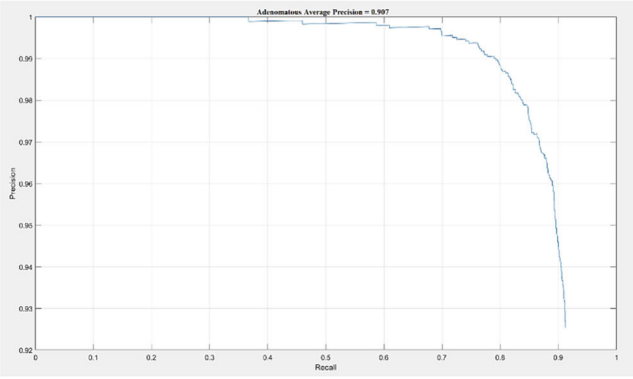



**FIGURE 21.** YOLOv4 PR curve for adenomatous polyp class.

Figs. 21 and 22 show the PR curves where the AP values are 0.907 and 0.853 for adenomatous and hyperplastic polyps, respectively.

The results indicated that adenomatous classification is more accurate than hyperplastic classification when the YOLOv4 model is used. The ROC curves for the four models implemented are shown in Fig. 23. The performance comparison was made based on the highest AUC value after acquiring the respective ROC curves. YOLOv4 achieved the highest AUC value of 0.8709 (SSD=0.826; YOLOv3=0.801, and FasterRCNN=0.779).

## VI. DISCUSSION

This comparative polyp detection and classification study concluded YOLOv4 as the best performing model with accuracies of 90.5% for adenomatous and 92.1% for hyperplastic.

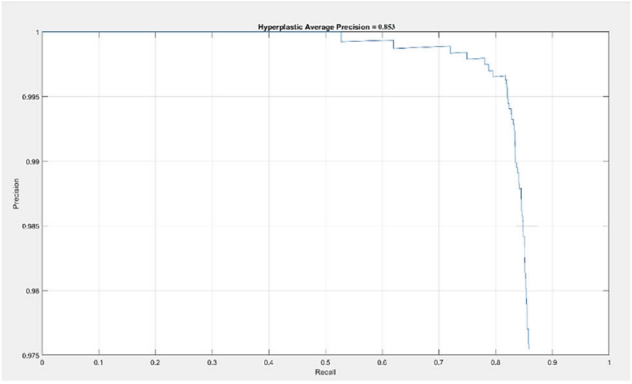



**FIGURE 22.** YOLOv4 PR curve for hyperplastic polyp class.

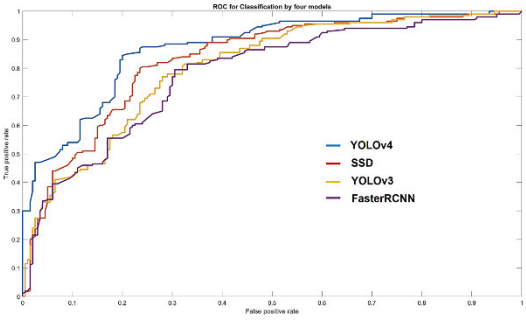



**FIGURE 23.** Combined ROC curves for the four models, i.e., YOLOv4, YOLOv3, SSD, and FasterRCNN.

Data preprocessing with image annotations and data augmentation techniques were implemented to enhance the training dataset to train the models with more features and make them more robust. Due to proper results, only three data augmentation techniques, i.e., random horizontal flip, random scaling, and color jitter were adopted to enhance the dataset

**TABLE 12.** Statistical analysis for models accuracies comparison.

| Comparison | $\hat{p}_1$ | | $\hat{p}_2$ | | $\hat{p}$ | | Z value | |
|---|---|---|---|---|---|---|---|---|
| | Ad | Ad | Ad | Hp | Ad | Hp | Ad | Hp |
| FasterRCNN vs YOLOv4 | 0.749 | 0.652 | 0.905 | 0.921 | 0.827 | 0.786 | 10.695 | 16.501 |
| SSD vs YOLOv4 | 0.793 | 0.836 | 0.905 | 0.921 | 0.858 | 0.878 | 8.052 | 6.534 |
| YOLOv3 vs YOLOv4 | 0.647 | 0.902 | 0.905 | 0.921 | 0.776 | 0.911 | 15.891 | 1.676 |

three times. The other data augmentation techniques were not adopted due to improper results.

Since the test images were based on both datasets (CGMH and Harvard Dataverse), it is imperative to mention the causes of misjudgment in detection that lead to detection errors. As mentioned in Section III, unclear and noisy images were manually removed from the CGMH dataset; similarly, while refining the datasets, another cause that may lead to low detection accuracy is blur images. In both datasets, images were acquired from colonoscopy videos; therefore, during the colonoscopy, the video/image captured in different angles and positions led to blur frames, as shown in Fig. 24. Other than polyps with a similar texture as the colonic wall, images with high noise due to high brightness, small polyps, and images with unclear polyps (during colonoscope movement) were also removed before the training process.

**FIGURE 24.** Samples of colonoscopy images with unclear polyps. These samples are related to Harvard Dataverse of colonoscopy.

In this study, our objective was to detect and classify a single polyp class within a single image. However, we encountered a few issues as we faced only in a few images, i.e., null value (empty values in class types and confidence score) and two-class identification in a single image. Dealing with these two significant issues is vital to avoid false prediction and polyp miss rate. Images with a high level of noise within the dataset can result in errors when it comes to detecting and classifying objects. The misidentification of polyps and classification errors can occur due to non-apparent polyps and polyp textures that closely resemble those of the intestinal wall [12].

It is imperative to deal with these two major issues to avoid false prediction and polyp miss rate. To address these issues, we developed specific functions that effectively eliminate the null values resulting from the detection process. Additionally, we implemented a mechanism to discard low probability values in the multi-class probabilities, retaining only the higher value. These functions are integrated with each model (fasterRCNN, SSD, YOLOv3, AND YOLOv4) as part of the model at the end as a separate module to deal with the mentioned issues as shown in Fig 15. Algorithm 2 is designed and shown below to illustrate the sequence of steps followed by the error handling module. These measures helped us improve the accuracy and reliability of our classification results. samples of two-class identification in a single

**Algorithm 2** Polyp Detection Errors Handling (Null Value and Multiple Object Detection in Predicted Images)

**Input**: Predicted images (predicted output data from each model with the labelled information)
**Function** DataHandling(Input)
**Repeat**
  Read each image info. in the predicted dataset (no. of images, bounding boxes, labels)
  **If** (no. of Bounding box == 1)
    **If** (confidence score >= 0.5)
      Store information (Confidence score, Polyp class)
    **Else**
      Count the images with less confidence score and discard the image
    **End**
  **Else If** (no. of Bounding box == 0)
    Count and discard the image
  **Else If** (no. of Bounding box >1)
    Read and compare confidence score information of bounding boxes
    Discard the less confidence value (bounding box)
    **If** (remaining bounding box confidence score >= 0.5)
      Store information (Confidence score, Polyp class)
    **Else**
      Count and discard the image
    **End**
  **End**
  **If** (no. of discarded images > 5% of the total predicted images)
    Retrain the model
  **End**
**Until** (the last image is traversed in the predicted dataset)
**End Function**

image and null detection are shown in Fig. 25 and Fig. 26, respectively.

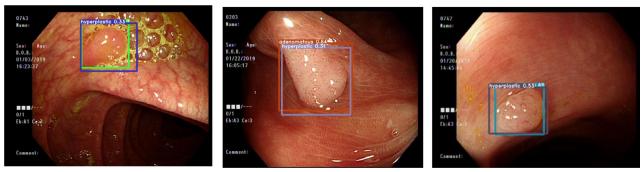


**FIGURE 25.** Two-class identification in single images. These test images are samples from the CGMH dataset.

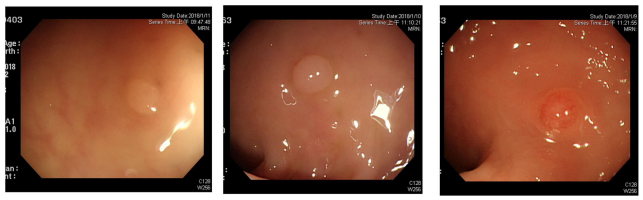


**FIGURE 26.** Samples of null detection. Test images were selected from the CGMH dataset to test the model's polyp detection and classification ability.

## VII. CONCLUSION

This study developed, evaluated, compared, and analyzed the performance of polyp detection systems that can detect colon polyps with the aim of early resection of colon polyps. In this research, a large dataset was obtained from a hospital and combined with the Harvard Dataverse of colonoscopic images. The CGMH dataset was then refined to remove noisy, unclear images and similar images. Finally, a diverse colonoscopy dataset was developed for polyp detection and classification. Three data augmentation techniques were applied to the dataset to enhance the models' training and prediction performances. Polyps were classified as either adenomatous or hyperplastic. Four state-of-the-art deep learning models, i.e., FasterRCNN, SSD, YOLOv3, and YOLOv4, were evaluated using the dataset. Their performance was finally compared and analyzed in detail using different performance metrics to identify the best performing model for each polyp class based on the results and analysis. We developed and integrated a polyp detection error handling module with each model to avoid false predictions and maintain high accuracy. A significant statistical difference was observed between the performances of the object detection models. In class detection and classification for both polyp classes, YOLOv4 ranked the best in this study, obtaining the highest precision, recall, F1-score, AP, kappa score, and accuracy. It also outperformed all other evaluated models.

## ACKNOWLEDGMENT

The authors would like to thank Chen-Ming Hsu (M.D.) with the Department of Gastroenterology and Hepatology, Chang Gung Memorial Hospital (CGMH), Taoyuan, Taiwan, for providing the polyp database and for helping them to identify the diagnostic diseases in medical images.

## REFERENCES

[1] American Cancer Society. *Key Statistics for Colorectal Cancer*. Accessed: Sep. 14, 2022. [Online]. Available: https://www.cancer.org/cancer/types/colon-rectal-cancer/about/key-statistics.html

[2] Ministry of Health and Welfare National Health Service. (2021). *General Situation of Prevention and Treatment of Colorectal Cancer*. Accessed: Sep. 14, 2022. [Online]. Available: https://www.hpa.gov.tw/Pages/Detail.aspx?nodeid=615&pid=1126

[3] Q.-X. Huang, G.-S. Lin, and H.-M. Sun, "Classification of polyps in endoscopic images using self-supervised structured learning," *IEEE Access*, vol. 11, pp. 50025–50037, 2023, doi: 10.1109/ACCESS.2023.3277029.

[4] M. S. Hossain, M. M. Rahman, M. M. Syeed, M. F. Uddin, M. Hasan, M. A. Hossain, A. Ksibi, M. M. Jamjoom, Z. Ullah, and M. A. Samad, "DeepPoly: Deep learning-based polyps segmentation and classification for autonomous colonoscopy examination," *IEEE Access*, vol. 11, pp. 95889–95902, 2023, doi: 10.1109/access.2023.3310541.

[5] K. Elkarazle, V. Raman, P. Then, and C. Chua, "Improved colorectal polyp segmentation using enhanced MA-NET and modified mix-ViT transformer," *IEEE Access*, vol. 11, pp. 69295–69309, 2023, doi: 10.1109/ACCESS.2023.3291783.

[6] Y. Wen, L. Zhang, X. Meng, and X. Ye, "Rethinking the transfer learning for FCN based polyp segmentation in colonoscopy," *IEEE Access*, vol. 11, pp. 16183–16193, 2023, doi: 10.1109/ACCESS.2023.3245519.

[7] A. Krenzer, M. Banck, K. Makowski, A. Hekalo, D. Fitting, J. Troya, B. Sudarevic, W. G. Zoller, A. Hann, and F. Puppe, "A real-time polyp-detection system with clinical application in colonoscopy using deep convolutional neural networks," *J. Imag.*, vol. 9, no. 2, p. 26, Jan. 2023, doi: 10.3390/jimaging9020026.

[8] W. Siłka, M. Wieczorek, J. Siłka, and M. Woźniak, "Malaria detection using advanced deep learning architecture," *Sensors*, vol. 23, no. 3, p. 1501, Jan. 2023, doi: 10.3390/s23031501.

[9] M. Stan-Ilie, V. Sandru, G. Constantinescu, O.-M. Plotogea, E. M. Rinja, I. F. Tincu, A. Jichitu, A. E. Carasel, A. C. Butuc, and B. Popa, "Artificial intelligence—The rising star in the field of gastroenterology and hepatology," *Diagnostics*, vol. 13, no. 4, p. 662, Feb. 2023, doi: 10.3390/diagnostics13040662.

[10] D. Wang, X. Wang, S. Wang, and Y. Yin, "Explainable multitask Shapley explanation networks for real-time polyp diagnosis in videos," *IEEE Trans. Ind. Informat.*, vol. 19, no. 6, pp. 7780–7789, Jun. 2023, doi: 10.1109/TII.2022.3208364.

[11] I. Vilkoite, I. Tolmanis, H. A. Meri, I. Polaka, L. Mezmale, L. Anarkulova, M. Leja, and A. Lejnieks, "The role of an artificial intelligence method of improving the diagnosis of neoplasms by colonoscopy," *Diagnostics*, vol. 13, no. 4, p. 701, Feb. 2023, doi: 10.3390/diagnostics13040701.

[12] C.-M. Hsu, C.-C. Hsu, Z.-M. Hsu, F.-Y. Shih, M.-L. Chang, and T.-H. Chen, "Colorectal polyp image detection and classification through grayscale images and deep learning," *Sensors*, vol. 21, no. 18, p. 5995, Sep. 2021, doi: 10.3390/s21185995.

[13] I. Pacal and D. Karaboga, "A robust real-time deep learning based automatic polyp detection system," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104519, doi: 10.1016/j.compbiomed.2021.104519.

[14] K. Li, M. I. Fathan, K. Patel, T. Zhang, C. Zhong, A. Bansal, A. Rastogi, J. S. Wang, and G. Wang, "Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations," *PLoS ONE*, vol. 16, no. 8, Aug. 2021, Art. no. e0255809, doi: 10.1371/journal.pone.0255809.

[15] C.-P. Tang, K.-H. Chen, and T.-L. Lin, "Computer-aided colon polyp detection on high resolution colonoscopy using transfer learning techniques," *Sensors*, vol. 21, no. 16, p. 5315, Aug. 2021, doi: 10.3390/s21165315.

[16] I. Pacal, A. Karaman, D. Karaboga, B. Akay, A. Basturk, U. Nalbantoglu, and S. Coskun, "An efficient real-time colonic polyp detection with YOLO algorithms trained by using negative samples and large datasets," *Comput. Biol. Med.*, vol. 141, Feb. 2022, Art. no. 105031, doi: 10.1016/j.compbiomed.2021.105031.

[17] M. C. Hoang, K. T. Nguyen, J. Kim, J.-O. Park, and C.-S. Kim, "Automated bowel polyp detection based on actively controlled capsule endoscopy: Feasibility study," *Diagnostics*, vol. 11, no. 10, p. 1878, Oct. 2021, doi: 10.3390/diagnostics11101878.

[18] A. Nogueira-Rodríguez, R. Domínguez-Carbajales, F. Campos-Tato, J. Herrero, M. Puga, D. Remedios, L. Rivas, E. Sánchez, Á. Iglesias, J. Cubiella, F. Fdez-Riverola, H. López-Fernández, M. Reboiro-Jato, and D. Glez-Peña, "Real-time polyp detection model using convolutional neural networks," *Neural Comput. Appl.*, vol. 34, no. 13, pp. 10375–10396, Jul. 2022, doi: 10.1007/s00521-021-06496-4.

[19] S. Y. Quan, M. T. Wei, J. Lee, R. Mohi-Ud-Din, R. Mostaghim, R. Sachdev, D. Siegel, Y. Friedlander, and S. Friedland, "Clinical evaluation of a real-time artificial intelligence-based polyp detection system: A US multi-center pilot study," *Sci. Rep.*, vol. 12, no. 1, p. 6598, Apr. 2022, doi: 10.1038/s41598-022-10597-y.

[20] Z. Qian, W. Jing, Y. Lv, and W. Zhang, "Automatic polyp detection by combining conditional generative adversarial network and modified you-only-look-once," *IEEE Sensors J.*, vol. 22, no. 11, pp. 10841–10849, Jun. 2022, doi: 10.1109/JSEN.2022.3170034.

[21] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *IEEE Access*, vol. 9, pp. 40496–40510, 2021, doi: 10.1109/ACCESS.2021.3063716.

[22] Z. Qian, Y. Lv, D. Lv, H. Gu, K. Wang, W. Zhang, and M. M. Gupta, "A new approach to polyp detection by pre-processing of images and enhanced faster R-CNN," *IEEE Sensors J.*, vol. 21, no. 10, pp. 11374–11381, May 2021, doi: 10.1109/JSEN.2020.3036005.

[23] Chang Gung Memorial Hospital CGMH. (2015). *Colon and Rectal Surgery*. Accessed: Nov. 15, 2022. [Online]. Available: https://www1.cgmh.org.tw/branch/lnk/2016/en/dept2.aspx?deptId=32800

[24] G. Wang, "Replication data for: Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations," Harvard Dataverse, Harvard Univ., Cambridge, MA, USA, Tech. Rep. Version 01, May 2021. [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FCBUOR

[25] J. Chaki and M. Woźniak, "Deep learning for neurodegenerative disorder (2016 to 2022): A systematic review," *Biomed. Signal Process. Control*, vol. 80, Feb. 2023, Art. no. 104223, doi: 10.1016/j.bspc.2022.104223.

[26] J. Chaki and M. Woźniak, "A deep learning based four-fold approach to classify brain MRI: BTSCNet," *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, Art. no. 104902, doi: 10.1016/j.bspc.2023.104902.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 1, 2015, pp. 91–99.

[28] W. Liu, "SSD: Single shot MultiBox detector," in *Proc. Comput. Vis. (ECCV)*, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[29] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

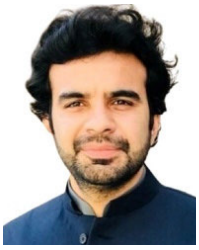
[30] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[31] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1571–1580.

[32] K. Wang, C. Xu, G. Li, Y. Zhang, Y. Zheng, and C. Sun, "Combining convolutional neural networks and self-attention for fundus diseases identification," *Sci. Rep.*, vol. 13, no. 1, p. 76, Jan. 2023, doi: 10.1038/s41598-022-27358-6.

[33] R. A. Johnson, I. Miller, and J. E. Freund, *Miller & Freund's Probability and Statistics for Engineers*, 9th ed. Boston, MA, USA: Pearson, 2016.

**YAO-TIEN CHEN** received the Ph.D. degree in computer science and information engineering from National Central University, Jhongli, Taiwan, in June 2007. He is currently an Assistant Professor with the International Ph.D. Program in Innovative Technology of Biomedical Engineering and Medical Devices, Ming Chi University of Technology, New Taipei City, Taiwan. His current research interests include computer vision, image processing, and virtual reality, especially in medical image segmentation, wavelet-based image processing, and object detection in medical images.



**NISAR AHMAD** received the master's degree in electronics design from Mid Sweden University, Sundsvall, Sweden. He is currently pursuing the Ph.D. degree with the International Ph.D. Program in Innovative Technology of Biomedical Engineering and Medical Devices, Ming Chi University of Technology, New Taipei City, Taiwan. His research interests include image processing, computer vision, deep learning-based object detection, classification, and the segmentation of medical images.