# An Automated Framework for Detection, Localization, and Classification of Colonic Polyp using Deep Learning

**Pradipta Sasmal** ( s.pradipta@iitg.ac.in )
Indian Institute of Technology Guwahati
**Avinash Paul**
Indian Institute of Technology Guwahati
**M. K. Bhuyan**
Indian Institute of Technology Guwahati
**Yuji Iwahori**
Chubu University
**Naotaka Ogasawara**
Aichi Medical University
**Kunio Kasugai**
Aichi Medical University

**Research Article**

**Keywords:** Localization, Classification of Colonic Polyp

**Posted Date:** September 21st, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-904695/v1

# An Automated Framework for Detection, Localization, and Classification of Colonic Polyp using Deep Learning

**Pradipta Sasmal[1,*], Avinash Paul[1], M.K. Bhuyan[1], Yuji Iwahori[2], Naotaka Ogasawara[3], and Kunio Kasugai[3]**

[1]Department of Electronics and Electrical Engineering, Indian Institute of Technology, Guwahati, Assam, 781039, India
[2]Department of Computer Science, Chubu University, Kasugai 487-8501, Japan
[3]Department of Internal Medicine, Division of Gastroenterology, Aichi Medical University, School of Medicine, Nagakute 480-1195, Japan
[*]corresponding author: s.pradipta@iitg.ac.in

## ABSTRACT

Colorectal cancer (CRC) in its advanced stage is one of the leading causes of death worldwide. However, early detection of polyps which are the precursor to such cancer can lead to better prognosis and clinical management. This report proposes an automated diagnostic technique to detect, localize, and classify polyps in colonoscopy video frames. Manual detection and localization of polyps on hugely acquired colonic frames have many limitations. Our deep learning-based framework proposes an attention-based YOLOv4 detector for polyp detection and localization. Finally, leveraging a fusion of deep and handcrafted features of the polyps, the detected polyps are classified as benign or malignant. The individual and the cross-database performances on two databases suggest the robustness of our method in polyp localization. The comparison of our approach based on significant clinical parameters with current state-of-the-art methods confirms that our method can be used for automated polyp localization in both real-time and offline colonoscopic video frames. Our method can give an average precision of 0.8971 and 0.9171 and an average IoU of 0.8325 and 0.8179 for the Kvsir-SEG and SUN databases, respectively. Similarly, our proposed classification framework on the detected polyps yields a classification accuracy of 96.66% on a public dataset.

## Introduction

Colorectal cancer (CRC) is one of the major health crises across the globe. The high mortality of CRC contributes significantly to the total deaths, and it is considered to be the third most frequently occurring cancer[1]. Such cancer in its initial state is called polyp and is generally benign. Polyps are abnormal tissues and are usually found in the mucosa of the colon[2]. Colonoscopy is one of the medical procedures adopted in detecting such polyps. Early detection of such polyps is crucial. It helps in a better prognosis and can increase the possibility of survival. During the entire colonoscopy, a considerable number of images of colon regions are captured. Nowadays, wireless capsule endoscopy (WCE) is used, which captures thousands of images of the entire gastrointestinal (GI) tract[3]. The doctors inspect each captured frame for detecting the presence of an anomaly. However, reviewing each frame manually for polyp detection from a hugely acquired colonic frame is very difficult and inefficient. The features of polyps are so indistinctive that sometimes it is challenging to distinguish them from the normal colon tissues. Also, the maximum polyp detection rate, which can be achieved through colonoscopy, is less than 50 % as it is highly operator dependent[4]. Therefore, it is essential to reduce the miss detection rate of polyps.

The application of new technologies in health care applications is on a constant rise. With the advent of new modalities, efforts have been made to enhance the efficiency of colonoscopy. Recently, optical endoscopic modalities using narrow-band imaging (NBI) have been developed for improved colorectal lesions detection[5,6]. NBI imaging enhances the vascular pattern of the lesion, thereby increases their discriminating ability. Blue light imaging, and i-Scan endoscopy are also used for better polyp detection[7,8]. Another endoscopy modality that acquires high definition (HD) images is linked color imaging (LCI). This imaging technique uses colour as an important cue for lesions. Generally, the color of a malignant (Adenomatous) polyp looks reddish, and the colour of a non-adenomatous polyp is whitish[9]. LCI enhances the red and white color and makes the red area look more reddish, and the white area looks more whitish during colonoscopy. Thus, LCI not only helps in lesion detection but also helps in their classification. Other techniques adopted to improve lesion detection include better bowel preparation, use of the broad field of view camera, flattening of colonic folds, etc.[10]. However, the diagnosis using these techniques for better

polyp detection during colonoscopy needs a highly experienced and trained expert in this domain[11].

Another problem that arises during lesion detection in colonoscopy frames is the high variability in the polyp characteristics. Typically, small or serrated polyps, diminutive and isochromatic flat polyps, are missed during manual inspection. Device and patient-specific colonoscopic frames will have different image characteristics[12]. Therefore, the generalization of a particular methodology in colonoscopy image analysis cannot be made. Therefore, all the above-discussed challenges must be considered while proposing an automated polyp detection system. An automatic diagnostic assistant system (DAS) is proposed for polyp detection and localization in this report. Subsequently, the detected polys are classified into benign (hyperplastic) or malignant (Adenomatous) by our proposed classification system. Our method does not need any expert during colonoscopy. Our proposed deep learning-based method can do polyp detection and localization on off-line and real-time colonoscopy frames. First, we'll go through our suggested technique for polyp detection and localization.

Both handcrafted feature learning and deep learning-based methods have been proposed over the years for polyp detection in the literature. Handcrafted-based techniques use different cues from the polyps' image, viz. color, texture, shape, surface properties, etc. On the contrary, deep learning-based methods use the hidden features of the image. Most of the works using handcrafted based feature learning methods are based on supervised learning[13–16]. However, these methods provide inferior performances as features learned during the training of a supervised model may not be sufficient to generalize the test datasets. The huge variation of image features among the acquired data from different endoscopic modalities gives unsatisfactory performances even in the same modality. Sasmal et al.,[17] proposed an unsupervised polyp detection method using saliency map and particle filtering. Recently, deep learning-based automated polyp detection system have been proposed for real-time polyp detection[18–21]. Convolutional neural networks (CNNs) based methods have been deployed in medical imaging for various tasks[22–26]. Shin et al.[27], proposed a transfer learning approach for polyp detection in colonoscopy. They used Inception ResNet and proposed a region-CNN for the task. Shin et al.,[28] proposed a conditional generative adversarial network to generate synthetic colonoscopy images for improved detection performance. Lee et al.,[21] employed YOLO-v2[29] for real-time polyp detection and localization. Yamada et al.,[30] deployed Faster RCNN and VGG-16 to detect and localize lesions in endoscopic video frames. They achieved a real-time detection performance with minimum polyp miss rate in colonoscopy video frames.

Polyp detection systems based on deep learning have improved overall performance in video frames of colonoscopy[19]. One major issue with the deep learning-based techniques is that we cannot establish the generalization ability of these models though they perform better than the handcrafted-based methods. In medical procedures, especially in polyp detection and localization systems, the following features are desired: 1) consistency in performance, i.e., the DAS must reliably produce the performance independent of imaging modalities and patients. 2) minimum polyp miss rate, i.e., a high detection rate, and 3) real-time application, which could help in immediate attention to the patient.

Considering all these requirements for devising an automated polyp detection system, we propose an attention based YOLOv4 detector for these tasks. The main contribution of our proposed method can be summarized as follows. This work presents an attention mechanism in the YOLOv4 framework for improved polyp detection. Our approach proposes to use spatial and channel attention modules in the backbone of the YOLOv4 framework. The attention mechanism gives importance to the region of interest (ROI), i.e., the polyp regions in a colonoscopy frame. A comparison of performance based on important matrices with state-of-the-art methods is presented in this article. The performances evaluated on two databases validates the robustness and generalization capability of our approach.

## Materials and Methods

### Dataset

We used two databases viz. 1) Kvsir-SEG[31] and 2) SUN Colonoscopy Video Database[32] for detection and localization tasks. The Kvsir-SEG database is a freely available open-access database, whereas the SUN database can be used after registration and agreement from the source. SUN (Showa University and Nagoya University) Colonoscopy Video Database is the colonoscopy-video database designed to evaluate an automated colorectal-polyp detection system. It comprises 49,136 polyp frames taken from 100 different polyps using a high-definition endoscope (CF-HQ290ZI and CF-H290ECI; Olympus, Tokyo, Japan). Similarly, the Kvsir-SEG dataset contains 1000 image frames acquired using ScopeGuide, Olympus Europe, endoscope. Some of the samples from both the data sets are shown in Fig. 1. The details of the datasets are given in Table 1.

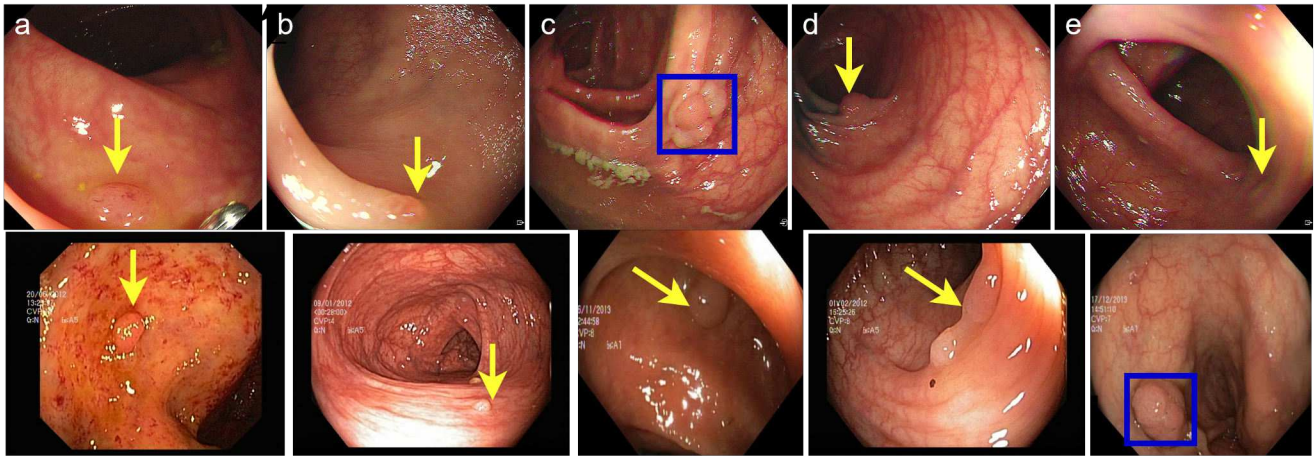


**Figure 1.** Some of the representative images from the trained databases. First-row image samples are from the SUN colonoscopy video database, and second-row images are from the Kvasir-SEG database. (a) 18 mm high-grade adenoma. (b) 2mm hyperplasic diminutive polyp. (c) 10mm low-grade adenoma polypoid polyp. (d) 4 mm distant diminutive polyp. (e) flat polyp.

| Dataset | Organ | Source | Findings | Dataset Contents | Size/ Polyp morphology |
|---|---|---|---|---|---|
| Kvsir-SEG[31] | Large bowel | WL | Polyp | 1000 images | Large polyp: 700 (> 160 × 160 pixels)<br>Median polyp: 323 (> 64 × 64 pixels and ≤ 160 × 160 pixels)<br>Small polyp: 48 (≤ 64 × 64 pixels) |
| SUN Colonoscopy[32] | Large bowel | WL | Polyps/non-polyps | 49,136/109,554 | Median (IQR) mm: 5 (3-7)<br>Diminutive polyp (< 5mm): 60<br>Morphology (Protruded/ flat): 66/ 34 |

**Table 1.** Details of the datasets.

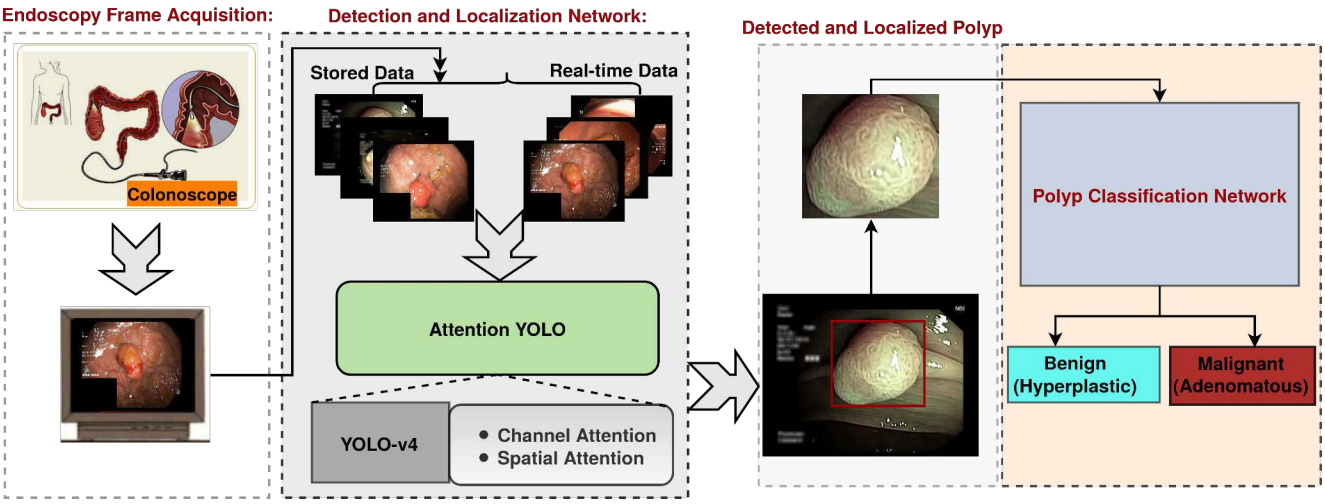


**Figure 2.** Proposed algorithm.

## Method

Fig. 2 shows the overall schema of our proposed methodology. During the colonoscopy, the captured video frames are stored in a computer system for further analysis in the future. However, real-time analysis of colonoscopy video frames can lead to

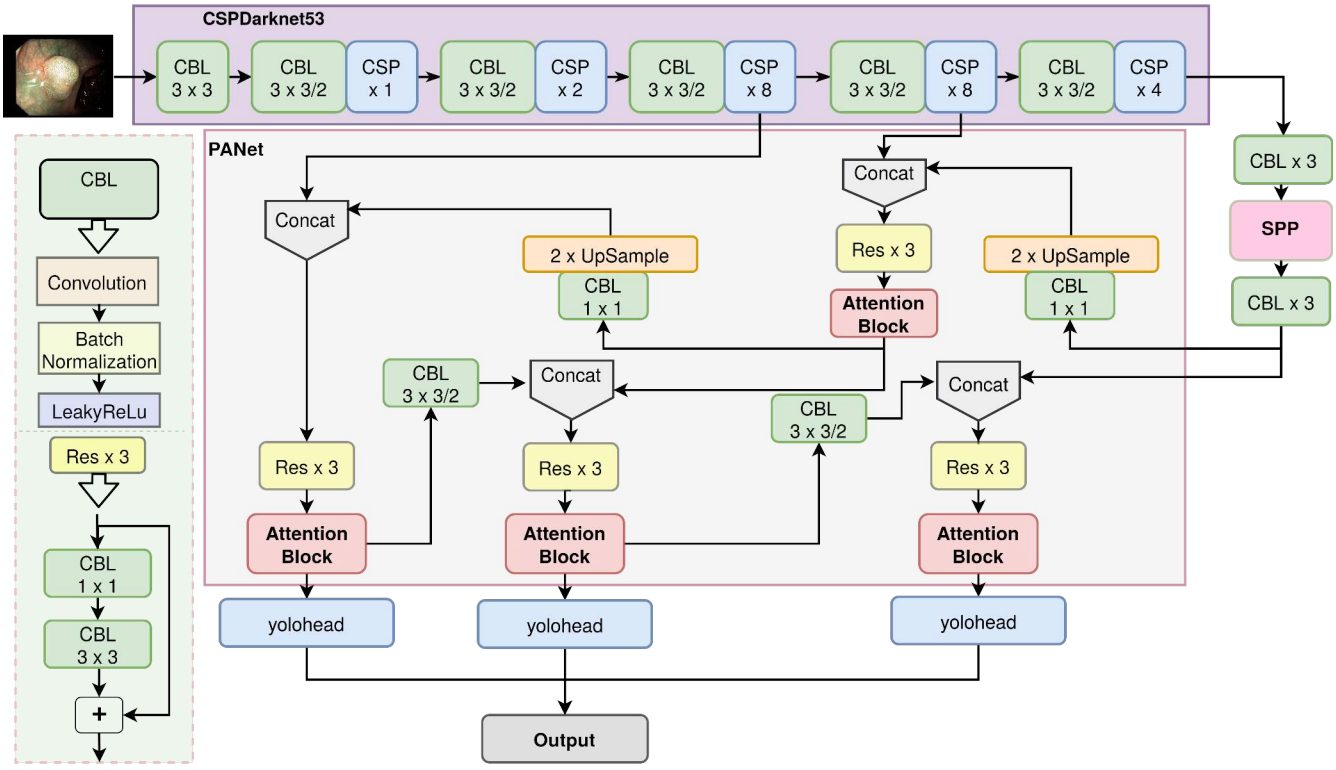


**Figure 3.** The proposed attention YOLOv4 network.

better diagnosis and early treatment. Also, the decision support system must automatically detect any abnormality in the frames. Therefore, the first stage of our current work focuses on handling real-time data for automatic detection and localization of polyps in the colonoscopy frames. For this, a deep learning-based attention YOLOv4 model is proposed in this work. The architecture of the proposed model is shown in Fig. 3. Furthermore, employing our suggested classification network, the localised polyps are classified as benign or malignant. The classification approach is explained in further detail later in this article.

## Attention YOLO

YOLO is a single-step object detection model and is considered superior to other deep learning models owing to its optimal accuracy and detection speed[29]. Further, YOLOv2[33] and YOLOv3[34] were proposed which show improved detection performances. In YOLOv3, a CNN Darknet53 is employed as a backbone of the architecture, efficiently extracting features from the input image. Later, YOLOv4 was proposed by Bochkovskiy et al.[35] to enhance the detection performance and speed. It integrates all the efficient approaches which are employed in different domains. Though it performs well on various datasets, its applicability and generalizability to medical imaging cannot be guaranteed. The medical images, especially endoscopic video frames, are generally of low quality, and they may have high noise, specularity, blur, etc. Also, a lack of annotated data may lead to overfitting the YOLOv4 model and make it less efficient in polyp detection and localization. Therefore, some changes corresponding to the polyp characteristics of endoscopic video frames are made in the existing model for better performances.

Occlusion, clutter, poor image quality, noise, etc., degrade detection performances. Generally, endoscopy videos suffer from such limitations. Also, the bounding box (BBoxes) used to localize the target objects may fit the arbitrary contour of objects. Therefore, various methods are generally adopted to highlight the real target object neglecting the background. The attention mechanism is among the solutions to these problems by enabling the network to focus more on the target object. Attention mechanisms are coupled in deep detection models to learn key features of the object. It mimics the property of the human visual system. Recently, attention mechanism has shown promising performances in various computer vision applications[36–39]. Therefore, the attention module is embedded into the backbone of CSP Darknet to focus more on the ROI of feature maps. This module would enable extraction of the polyp regions' important features, ignoring the non-polyp regions of colonoscopy frames. Our method proposes two attention modules, namely, the channel attention module and spatial attention module, and are incorporated in the backbone of YOLOv4. YOLOv4 extracts feature maps to three different branches to obtain three feature grid maps with various scales for detecting objects of different sizes. The three YOLO heads are then trying to localize the

objects with the BBoxes. Our proposed attention modules are integrated on the feature maps before the three YOLO heads can detect and localize polyps.

### Channel attention block

The channel attention block is proposed to integrate the interaction among the inter-channel feature maps. It is employed to enhance the vital information of a feature map of an object.

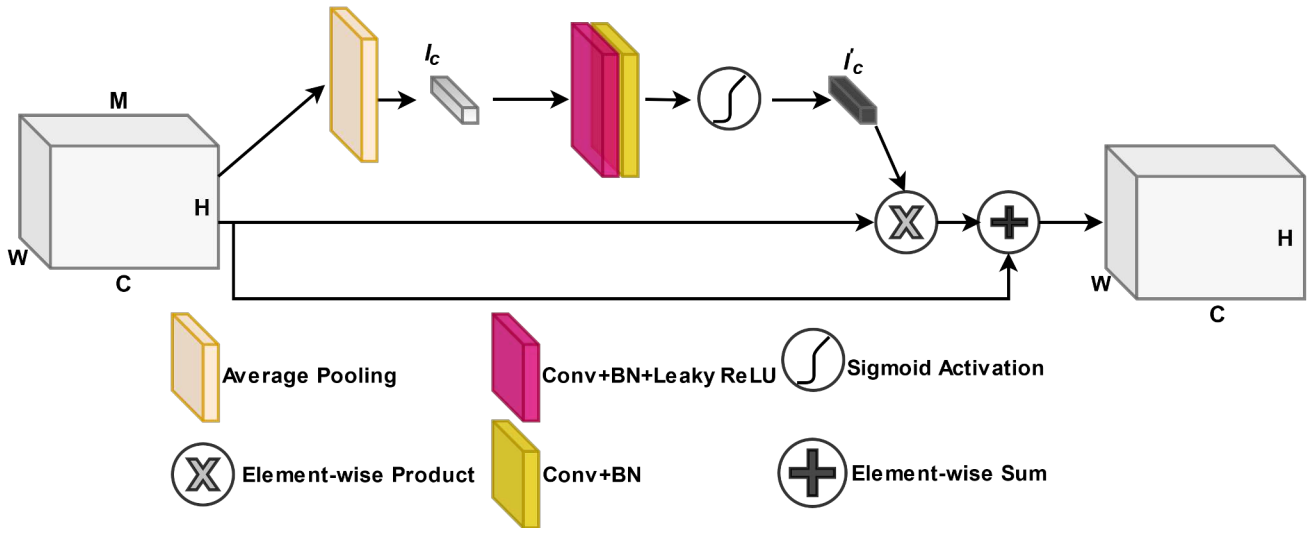


**Figure 4.** Channel attention block; Conv: Convolution, BN: batch normalization.

Let the input feature map be represented as $M \in R^{H\times W\times C}$, where $H$, $W$, and $C$ represent the height, width, and depth of a feature map, respectively. As shown in Fig. 4, a global average pooling operation is employed across all the depth maps to extract contextual information, embedded in the channel descriptor given as $I_c \in R^{1\times 1\times C}$, and the $c$-th element of $I_c$ is given by:

$$i_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} m_c(i,j) \tag{1}$$

$I_c = [i_1, i_2, ...., i_c]$ and $M = [m_1, m_2, ....., m_c]$. Again, to further explore the inter-channel nonlinear relationship among the channel maps, we employ a 2-layers CNN followed by a sigmoid activation function. In order to reduce some parameters overhead, $W_1$ is used as the dimensionality reduction layer with a reduction of factor 16[38]. Similarly, $W_2$ is used to increase the dimensionality again. This process is given as:

$$I_c^{'} = \sigma(W_2\phi(W_1 I_c)) \tag{2}$$

where, $W_1 \in \frac{C}{16}\times C$ and $W_2 \in C\times\frac{C}{16}$.

Finally, an element-wise summation operation is adopted between the input feature map and the generated channel attention map through residual connection to mitigate the incurred information loss. The final feature map is given as: $I_c^{'}\times M + M$. The channel attention module is illustrated in Fig. 4.

### Spatial attention

The spatial attention mechanism focuses on the local regions of a feature map. Thus, this module is employed to preserve the local polyp ROI information in the feature maps. Fig. 5 depicts a spatial attention module. As shown in Fig. 5, a $7\times 7$ convolutional layer is introduced to aggregate the interspatial interaction of maps to produce a one-dimensional spatial descriptor.

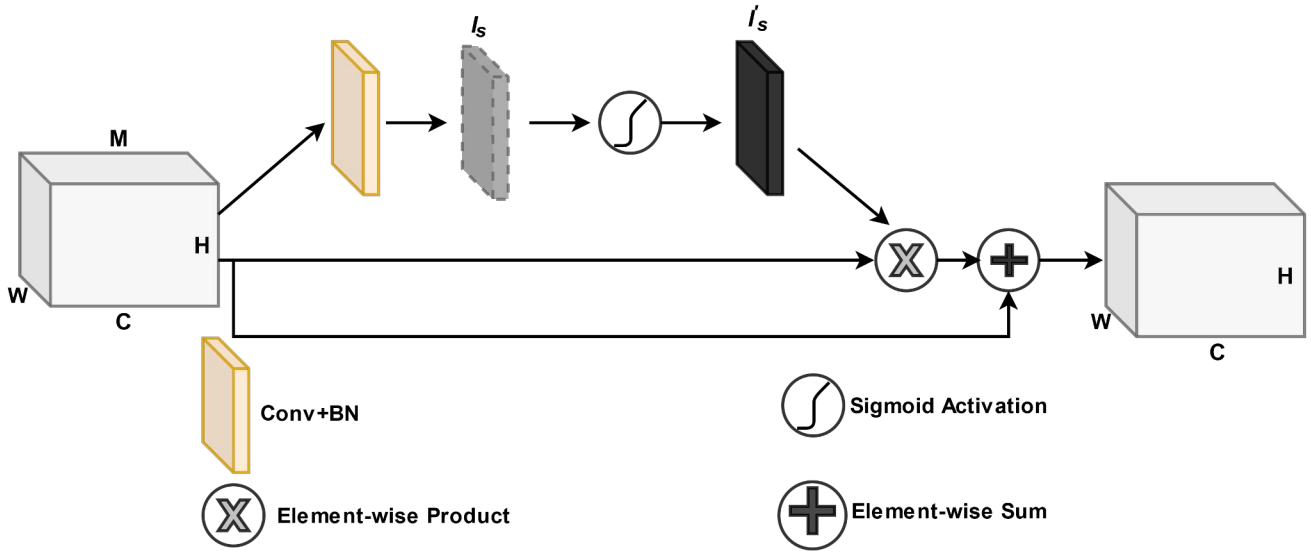


**Figure 5.** Spatial attention block.

Let the input feature map be represented as: $M \in R^{H \times W \times C}$.
Then, the generated feature descriptor $I_s$ is represented as:

$$I_s = conv^{7\times7}(M) \tag{3}$$

where, $I_s \in R^{H \times W \times 1}$ and $conv^{7\times7}(\cdot)$ denotes a $7 \times 7$ convolutional layer. The sigmoid function then activates this feature map to highlight the important regions. Subsequently, it is multiplied and summed up with the input feature maps to produce the final feature map, which is given as:
$I'_s \times M + M$
where, $I'_s = \sigma(I_s)$
The detailed experimental setup, training and performances are discussed in the Results section.

## Classification of Detected Polyps

Following the identification of polyps, endoscopists split off the polyp areas and vividly access them for cancer diagnosis. They do this by analysing several polyp features such as shape, colour, texture, and surface patterns etc. Due to the large medical images acquired during colonoscopy and the similarity in pathological manifestations across ailments, physical inspection and labelling of polyps is tedious and inefficient. The polyp characteristics may not always be visible to the human eye, and diagnostic information may be ignored, making decision-making extremely challenging. An automated polyp classifier for two-class polyp classification, i.e., adenoma (malignant) and hyperplastic (benign), is provided in this report to solve the aforementioned problems.

Hand-crafted feature learning approaches were used in the early research on automatic polyp categorization from colonoscopy frames[40–44]. The inconsistency of these approaches' performance in terms of repeatability is a drawback. Furthermore, the generalizability and robustness of these techniques cannot be guaranteed because pathological situations vary greatly even within the same modality's dataset. Also, a huge domain knowledge is required to characterize the discriminating features of the polyps.

Deep learning-based techniques are better at handling such variances and give a high degree of generalisation. As a result, there has been an increase in interest in using such models in medical image and video processing *especially* in polyp classification[45–49]. However, one of the primary drawbacks of these methods is that they require a large quantity of labelled data during training in order to get relatively good classification performance. Large-scale polyp databases, on the other hand, are harder to achieve by. The wide range of imaging methods and processes, as well as privacy concerns and a lack of medical integration, may provide a number of obstacles in obtaining high-quality, large-scale polyp images. In light of these issues, we suggested a classification technique that does not necessitate the use of large amounts of labelled data. We'll illustrate how the non-linearity of a small, imbalanced dataset may be correctly described by the features learned via our proposed network. As a result, in this work, we present a unique polyp categorization technique to solve some of these problems. For classification of the segmented polys, we propose using the Triplet Network architecture and its related triplet loss to learn

non-linear representations between polyps. We show that the learned features may be used as a highly discriminative basis for machine learning models. We compare our findings to those of prior research and show that the features acquired by a Triple Network can characterize the non-linearity of a small dataset, making them acceptable for use in a linear classifier. In addition, integrating deep and handcrafted features improves polyp classification efficiency. For deep features, we employed a triplet network based on siamese architecture and the handcrafted features were extracted using pyramid histogram of oriented gradient (PHOG). As discussed earlier, texture and shape information of polyps play a vital role while dysplasia grading by the endoscopists. In our proposed framework, the triplet network helps to learn distributed embedding by the notion of similarity and dissimilarity whereas the PHOG extracts the shape and texture information of the polyps[44]. The suggested classification approach is shown schematically in Fig. 7.

## PHOG

A polyp's geometry, texture, and colour provide enough information on its nature. The proposed approach uses a pyramid histogram of oriented gradient (PHOG) characteristics to define the geometry or morphology of a polyp. At each pyramid resolution level, the HOG vector is calculated (L). Finally, the PHOG descriptor is extracted by concatenating all of the HOG vectors. The PHOG descriptor's dimensionality for the full image is provided as: $K * \sum_{l=0}^{L} 4^l$. In this work, $K$ and $L$ values are taken as 8 and 4, respectively. The details of this feature extraction technique can be found in our previous paper[44]. The feature extraction technique using the proposed PHOG is shown in Fig. 6.

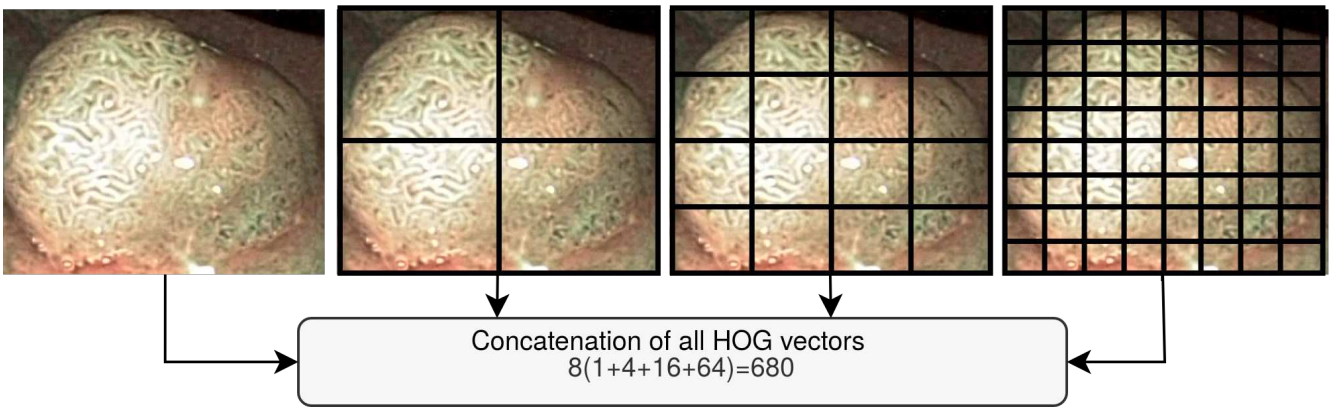


**Figure 6.** PHOG feature extaction.

## Triplet Network

The Siamese network[50] inspired Triplet Network design consists of three identical sub-networks with common parameters. Each sub-network is taught to recognize embedded characteristics in three different samples, the anchor, positive, and negative samples, respectively. A triplet is made up of an anchor, a positive, and a negative sample. The L2 distance between the anchor and the positive sample, as well as the anchor and the negative sample, are the network outputs. The cost function is computed using the triplet loss in Eq. 4, where $f_i^a$ represents the anchor embedding, $f_i^p$ represents the positive embedding, and $f_i^n$ represents the negative embedding.

$$L = \max(0, ||f_i^a - f_i^p||_2^2 - ||f_i^a - f_i^n||_2^2 + \alpha) \tag{4}$$

The value of $\alpha$ was taken 0.5, and the dimensionality of the embedding was set as 256.

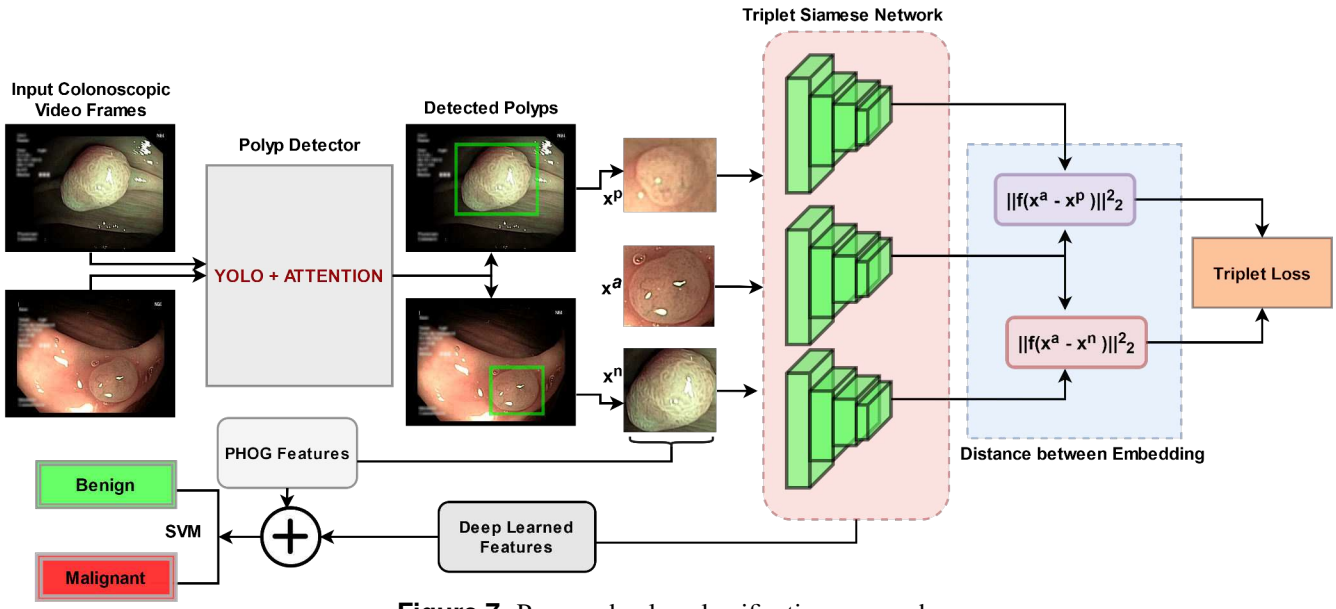


**Figure 7.** Proposed polyp classification approach.

### Training

The anchor and positive samples were labeled as benign polyps, while the negative was labeled as malignant. Three Triplet networks were trained using identical hyperparameters, with Adam as the preferred optimizer and learning rate 0.0001 as the hyperparameters. Each network was started using ImageNet weights and trained from the ground up. For each of the triplet's images, our Siamese Network will generate embeddings. We achieved this by connecting a few Dense layers to a ResNet50 model that has been pre-trained on ImageNet. All of the model's layers' weights will be frozen until the layer conv5_block1_out. The last layers were fine-tuned during training.

## Results

### Evaluation Metrices

In this work, some of the extensively recommended standard metrics are used to evaluate detection and localization performances.[18,51].

- $IoU\,(A,B)=\frac{A\cap B}{A\cup B}$, measures the overlap between two bounding boxes A and B as the ratio between the overlapped area.
- AP: Average precision was computed as an average APs for IoU from 0.25 to 0.75 with a step-size of 0.05.
- FPS=$\frac{\#frames}{sec}$.

Similarly, standard performance indicators like as accuracy, sensitivity, specificity, precision, recall, and F-score are employed for classification.

### Experimental Setup and Configuration

Two databases are used in our experiment for this study. The details of the databases are given in Table 1. 80% of the samples are used for training the YOLO attention model, and the rest 20% samples are used for validation. The size of the images is made $416 \times 416$. The model was tested in Google Colab (cloud GPU) with Nvidia Tesla T4 @585 MHz. The hyperparameters set for the YOLOv4+Attention model are as follows: Learning rate: $1e^{-3}$, batch size: 64, anchors: 8, and threshold: 0.25.

### Detection and Localization Results

Table 2 shows the detection and localization performances by different state-of-the-art methods on the Kvsir-SEG dataset. It can be observed that our method achieves an average precision (AP) of 0.8971, which is the best among all. The APs achieved at multiple IoU threshold i.e $\mathbf{AP_{25}}$, $\mathbf{AP_{50}}$, and $\mathbf{AP_{75}}$ are 0.9485, 0.9279, and 0.7849, respectively. The IoU measures the precision at which the bounding box localizes the target object. From our results, it is clearly observed that our method is better in localizing polyp ROIs compared to the state-of-the-art methods. Also, our method can detect polyps at a real-time speed of 50

| Method | Backbone | AP | IoU | $AP_{25}$ | $AP_{50}$ | $AP_{75}$ | FPS |
|---|---|---|---|---|---|---|---|
| EfficientDet-D0[52] | EfficientNet-b0 | 0.4756 | 0.4322 | 0.6846 | 0.5047 | 0.2280 | 35.00 |
| Faster R-CNN[53] | ResNet50 | 0.7866 | 0.5621 | 0.8947 | 0.8418 | 0.5660 | 8.00 |
| RetinaNet[54] | ResNet50 | 0.8697 | 0.7313 | 0.9395 | 0.9095 | 0.6967 | 16.20 |
| RetinaNet[54] | ResNet101 | 0.8745 | 0.7579 | 0.9483 | 0.9095 | 0.7132 | 16.80 |
| YOLOv3+spp[34] | Darknet53 | 0.8105 | 0.8258 | 0.8856 | 0.8532 | 0.7586 | 45.01 |
| YOLOv4[35] | Darknet53, CSP | 0.8513 | 0.8025 | 0.9348 | 0.9128 | 0.7757 | 66.67 |
| ColonSegNet[18] | - | 0.8000 | 0.8100 | 0.9000 | 0.8166 | 0.6706 | 180 |
| **YOLOv4+Attention** | Darknet53, CSP, Attention | **0.8971** | **0.8325** | **0.9485** | **0.9279** | **0.7849** | 50 |

**Table 2.** Detection and localization performace on Kvsir-SEG dataset.

FPS. Therefore, our method can be employed for accurate polyp detection and localization in real-time colonoscopy video frames.

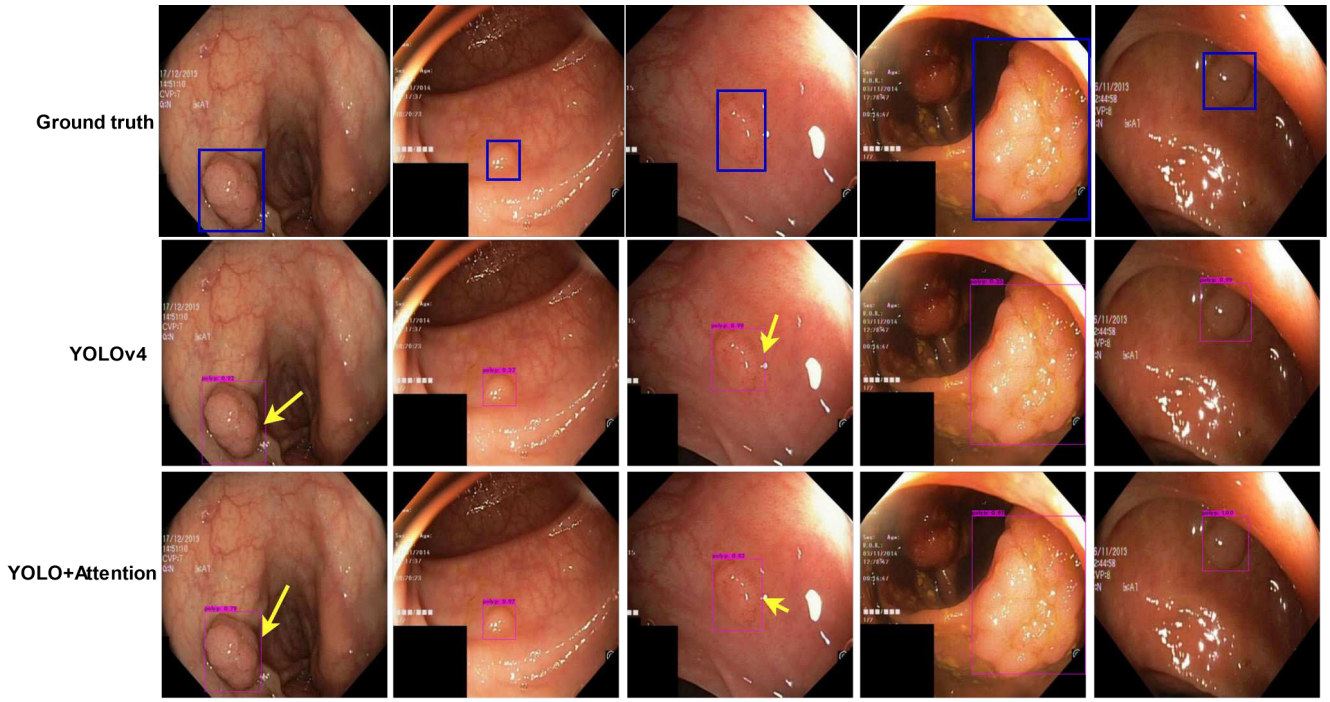


**Figure 8.** Detection and localization results on test dataset: Kvsir-SEG.

Fig. 8 shows qualitative results of some samples from the Kvsir-SEG dataset for the polyp detection and localization task. Results from the recent state-of-the-art method, YOLOv4, and our proposed method are shown in Fig. 8 . From the figures, it can be observed that both YOLOv4 and YOLOv4+Attention can detect and localize polyps with high confidence. Some of the bounding boxes are annotated with the yellow arrows to show that our proposed method is better in localizing the polyps. In YOLOv4, most polyps are localized with wider bounding boxes than the proposed YOLOv4+Attention. This is also validated with the quantitative results, where the average IoU for YOLOv4+Attention is better than YOLOv4, as shown in Table 2. Samples with the blue bounding boxes are ground truths and are available with the data set.

| Method | Backbone | AP | IoU | $AP_{25}$ | $AP_{50}$ | $AP_{75}$ | FPS |
|---|---|---|---|---|---|---|---|
| YOLOv4 | Darknet53, CSP | 0.8597 | 0.8052 | 0.9762 | 0.9621 | 0.6408 | 66.67 |
| **YOLOv4+Attention** | Darknet53, CSP, Attention | **0.9172** | **0.8179** | **0.9868** | **0.9721** | **0.7328** | **50** |

**Table 3.** Detection and localization performace on SUN data set.

| Method | Backbone | AP | IoU | $AP_{25}$ | $AP_{50}$ | $AP_{75}$ | FPS |
|---|---|---|---|---|---|---|---|
| YOLOv4 | Darknet53, CSP | 0.8597 | 0.7240 | 0.8789 | 0.7945 | 0.5287 | 66.67 |
| **YOLOv4+Attention** | Darknet53, CSP, Attention | **0.9172** | **0.7667** | **0.9231** | **0.8600** | **0.6144** | **50** |

**Table 4.** Cross dataset detection and localization performace: Trained on SUN Colonoscopy database and tested on Kvsir-SEG dataset.

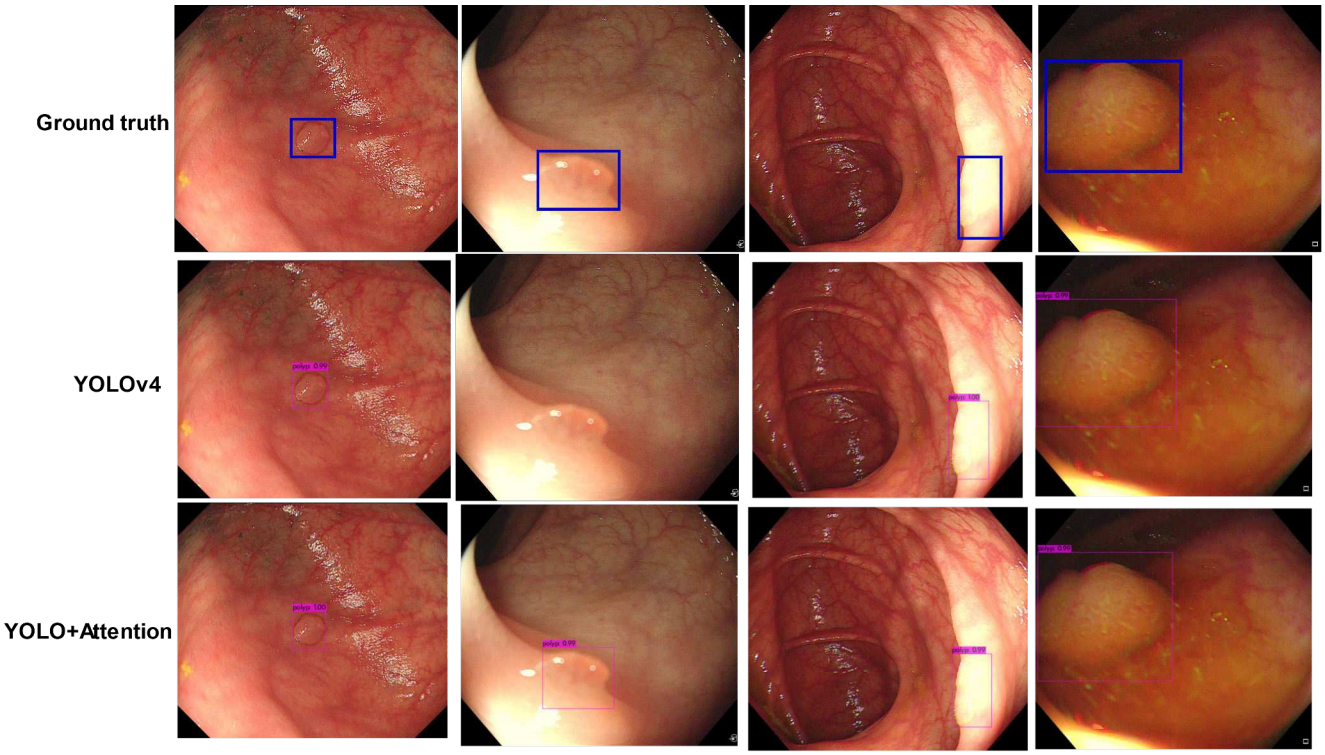


**Figure 9.** Detection and localization results on test data set: SUN Colonoscopy database.

The performances on the SUN database with YOLOv4 and the proposed method are also shown in Table 3. The qualitative localization performances on some of the samples of the SUN database are shown in Fig. 9. It is observed that similar performances are also achieved on this data set. YOLOv4 model did not detect the second image of the second-row polyp, but our proposed model could detect and localize it. Further to validate the robustness of our model, we also cross-validated the performances. We evaluated the performance using the test data from the Kvasir-SEG dataset while the model was trained with the SUN database. The performance of the cross dataset is shown in Table 4.

## Classification Results

The proposed method is validated on the publicly available a labeled polyp dataset for colorectal polyps classification[40]. The dataset is available at url: http://www.depeca.uah.es/colonoscopy_dataset/. It contains video sequences using narrow-band imaging (NBI) and White light (WL) imaging. The dataset contains video sequences for 21, 15, and 40 hyperplastic (benign), serrated, and adenoma (malignant) polyps. Fig. 10 shows some of the samples from both the classes of the dataset. The video sequences are converted to frames, and from each frames, the polyps are segmented out using the

proposed YOLOv4 attention network. Subsequently, these polyps are fed to the triplet network for classification. In this work, only NBI image frames from hyperplastic and adenoma classes are considered. Three-fold cross-validation was employed as a validation method for our approach. Extracted features are analyzed using linear SVMs to classify polyps between benign and malignant. A classification accuracy of 90.16% is achieved. The embedded features of the image samples of the database are analyzed using t-SNE and are shown in Fig. 11. Further, the PHOG features are also fused with the embedded features extracted from the triplet network to enhance the classification performances. The fusion of these features increases the dimensionality and non-linearity in the feature space. Therefore, an RBF kernel SVM was used for classification of the fused features. It was also varified from the experiments that the RBF SVM performs better as compared to other classifiers. Table 5 shows the classification accuracies of some of the handcrafted-based methods on the same dataset. Similarly, Table 6 shows the classification accuracies on the same dataset using the transfer learning approaches. Finally, the results are compared with the state-of-the-art methods, and it is clearly seen that our method gives better performances in a limited data environment. The results are shown in Table 7.

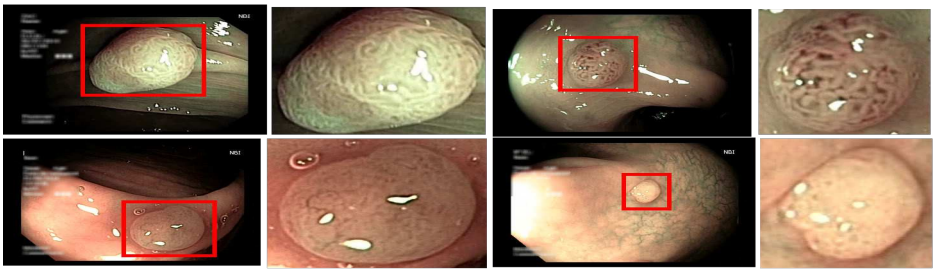

**Figure 10.** Sample frames from both the classes. First row samples are of malignant type and the bottom row images are of benign type. The polyps are segmented out by the YOLOv4 attention model.

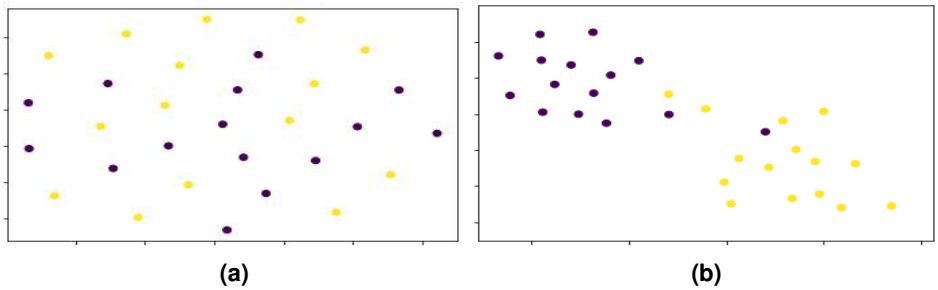


**Figure 11.** t-Distributed Stochastic Neighbor Embedding (t-SNE) is employed to decrease the dimensionality of the feature embedding into a 2D representation, . (a) Initial features after the first training epoch, illustrating the embedding space mixture of classes with no distinction. (b): the final embedding once the model has converged to an optimum solution. Malignant polyps are shown by yellow dots, while benign polyps are represented by violet dots.

| **Method** | **Accuracy %** |
|---|---|
| LBP | 72.29 |
| HOG | 68.93 |
| GLCM | 70.52 |
| Curvelet features[43] | 70.90 |
| LETRIST[55] | 80.27 |
| LGONBP[56] | 83.20 |
| FWLBP[57] | 80.25 |
| PHOG | 85.20 |
| PHOG+FWLBP[44] | 90.16 |

**Table 5.** Comparision of classification performance with differfent texture descriptors using SVM classifier.

| Method | **Accuracy** |
|---|---|
| **Proposed** | **96.66** |
| VGG16[58] | 82 |
| VGG16 fine-tuned | 90 |
| VGG19[58] | 70 |
| VGG19 fine-tuned | 89 |
| MobileNet[59] | 90.11 |
| ResNet50[60] | 68 |
| ResNet50 fine-tuned | 72 |
| Inception v3[61] | 90.02 |

**Table 6.** Comparison of classification accuracy between the baseline deep learning models and our method.

| | Accuracy | Sensitivity | Specificity | Precision | Recall | F-Score |
|---|---|---|---|---|---|---|
| **Proposed** | **96.66** | 93.33 | 93.75 | 93.33 | 100.00 | 96.54 |
| **2D Texture+3D Features**[40] | 89.47 | 94.55 | 76.19 | 91.23 | 94.55 | 92.86 |
| **LoG**[41] | 84.21 | 90.91 | 66.67 | 87.72 | 90.91 | 89.29 |
| **Color GLCM+SVM**[42] | 64.47 | 74.55 | 38.10 | 75.93 | 74.55 | 75.23 |
| **BoW+SPM**[43] | 73.68 | 98.18 | 90.52 | 73.97 | 98.18 | 84.38 |
| **Triplet Network**[62] | 90.16 | 92.50 | 85.71 | 92.50 | 98.25 | 92.50 |

**Table 7.** Comparison with the existing works.

## Discussion

This paper presents a framework for analysis of colonic polyps using colonoscopy video frames. A deep attention based YOLOv4 network is proposed to detect and localise polyps in the first step of the study. The performance of the suggested algorithm outperforms state-of-the-art approaches by a significant margin. The generalizability and robustness of our method are also demonstrated by the consistency of results across datasets and between datasets. Following that, the localised polyps are classified, which is crucial for better prognosis. We propose a triplet network based on siamese architecture, followed by SVM, to achieve this. Additionally, local polyp features are extracted and fused with deep features, resulting in improved classification results. The effectiveness of our strategy in a limited data environment is demonstrated by its classification performance on a relative small dataset. We hope to improve polyp detection and localization in future by training the network with features that best characterize the polyp clinical manifestations. Further, grading of dysplasia in polyps could also allow practitioners better comprehend pathological situations.

### Data availability

The Kvsir-SEG and the SUN databases are publicly available. The dataset used for the classification is also an open access dataset.

## References


1. Arnold, M. *et al.* Global patterns and trends in colorectal cancer incidence and mortality. *Gut* **66**, 683–691 (2017).
2. Messmann, H. *Atlas of Colonoscopy: Techniques-Diagnosis-Interventional Procedures.* (Thieme, 2006).
3. Iddan, G., Meron, G., Glukhovsky, A. & Swain, P. Wireless capsule endoscopy. *Nature* **405**, 417–417 (2000).
4. Kahi, C. J., Hewett, D. G., Norton, D. L., Eckert, G. J. & Rex, D. K. Prevalence and variable detection of proximal colon serrated polyps during screening colonoscopy. *Clin. Gastroenterol. Hepatol.* **9**, 42–46 (2011).
5. Rex, D. K. Narrow-band imaging without optical magnification for histologic analysis of colorectal polyps. *Gastroenterology* **136**, 1174–1181 (2009).
6. Wada, Y. *et al.* Diagnosis of colorectal lesions with the magnifying narrow-band imaging system. *Gastrointest. endoscopy* **70**, 522–531 (2009).
7. Guo, C.-G., Ji, R. & Li, Y.-Q. Accuracy of i-scan for optical diagnosis of colonic polyps: a meta-analysis. *PloS one* **10**, e0126237 (2015).
8. Pohl, J. *et al.* Computed virtual chromoendoscopy for classification of small colorectal lesions: a prospective comparative study. *Am. J. Gastroenterol.* **103**, 562–569 (2008).
9. Min, M. *et al.* Computer-aided diagnosis of colorectal polyps using linked color imaging colonoscopy to predict histology. *Sci. reports* **9**, 1–8 (2019).
10. Bond, A. & Sarkar, S. New technologies and techniques to improve adenoma detection in colonoscopy. *World journal gastrointestinal endoscopy* **7**, 969 (2015).
11. Kuiper, T. *et al.* Accuracy for optical diagnosis of small colorectal polyps in nonacademic settings. *Clin. Gastroenterol. Hepatol.* **10**, 1016–1020 (2012).
12. Jha, D. *et al.* Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, 225–2255 (IEEE, 2019).

**13.** Ganz, M., Yang, X. & Slabaugh, G. Automatic segmentation of polyps in colonoscopic narrow-band imaging data. *IEEE Transactions on Biomed. Eng.* **59**, 2144–2151 (2012).

**14.** Bernal, J., Sánchez, J. & Vilarino, F. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognit.* **45**, 3166–3182 (2012).

**15.** Iwahori, Y. *et al.* Automatic detection of polyp using hessian filter and hog features. *Procedia computer science* **60**, 730–739 (2015).

**16.** Iakovidis, D. K., Maroulis, D. E., Karkanis, S. A. & Brokos, A. A comparative study of texture features for the discrimination of gastric polyps in endoscopic video. In *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, 575–580 (IEEE, 2005).

**17.** Sasmal, P., Bhuyan, M., Gupta, S. & Iwahori, Y. Detection of polyps in colonoscopic videos using saliency map based modified particle filter. *IEEE Transactions on Instrumentation Meas.* (2021).

**18.** Jha, D. *et al.* Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *Ieee Access* **9**, 40496–40510 (2021).

**19.** Wang, P. *et al.* Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat. biomedical engineering* **2**, 741–748 (2018).

**20.** Urban, G. *et al.* Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology* **155**, 1069–1078 (2018).

**21.** Lee, J. Y. *et al.* Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets. *Sci. reports* **10**, 1–9 (2020).

**22.** Li, Q. *et al.* Colorectal polyp segmentation using a fully convolutional neural network. In *2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)*, 1–5 (IEEE, 2017).

**23.** Vázquez, D. *et al.* A benchmark for endoluminal scene segmentation of colonoscopy images. *J. healthcare engineering* **2017** (2017).

**24.** Bernal, J. *et al.* Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE transactions on medical imaging* **36**, 1231–1249 (2017).

**25.** Shin, H.-C. *et al.* Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging* **35**, 1285–1298 (2016).

**26.** Tajbakhsh, N. *et al.* Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging* **35**, 1299–1312 (2016).

**27.** Shin, Y., Qadir, H. A., Aabakken, L., Bergsland, J. & Balasingham, I. Automatic colon polyp detection using region based deep cnn and post learning approaches. *IEEE Access* **6**, 40950–40962 (2018).

**28.** Shin, Y., Qadir, H. A. & Balasingham, I. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. *IEEE Access* **6**, 56007–56017 (2018).

**29.** Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788 (2016).

**30.** Yamada, M. *et al.* Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. *Sci. reports* **9**, 1–9 (2019).

**31.** Jha, D. *et al.* Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, 451–462 (Springer, 2020).

**32.** Misawa, M. *et al.* Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointest. Endosc.* **93**, 960–967 (2021).

**33.** Redmon, J. & Farhadi, A. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271 (2017).

**34.** Redmon, J. & Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

**35.** Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).

**36.** He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis machine intelligence* **37**, 1904–1916 (2015).

**37.** Faster, R. Towards real-time object detection with region proposal networks. *Adv. neural information processing systems* **9199** (2015).

**38.** Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141 (2018).

**39.** Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19 (2018).

**40.** Mesejo, P. *et al.* Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE transactions on medical imaging* **35**, 2051–2063 (2016).

**41.** Riaz, F., Vilarino, F., Ribeiro, M. D. & Coimbra, M. Identifying potentially cancerous tissues in chromoendoscopy images. In *Iberian Conference on Pattern Recognition and Image Analysis*, 709–716 (Springer, 2011).

**42.** Engelhardt, S., Ameling, S., Wirth, S. & Paulus, D. Features for classification of polyps in colonoscopy. In *Bildverarbeitung für die Medizin*, vol. 574, 350–354 (2010).

**43.** Aman, J. M., Summers, R. M. & Yao, J. Characterizing colonic detections in ct colonography using curvature-based feature descriptor and bag-of-words model. In *International MICCAI Workshop on Computational Challenges and Clinical Opportunities in Virtual Colonoscopy and Abdominal Imaging*, 15–23 (Springer, 2010).

**44.** Sasmal, P., Bhuyan, M., Iwahori, Y. & Kasugai, K. Colonoscopic polyp classification using local shape and texture features. *IEEE Access* **9**, 92629–92639 (2021).

**45.** Fonollá, R., van der Sommen, F., Schreuder, R. M., Schoon, E. J. & de With, P. H. Multi-modal classification of polyp malignancy using cnn features with balanced class augmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 74–78 (IEEE, 2019).

**46.** Ribeiro, E., Uhl, A. & Häfner, M. Colonic polyp classification with convolutional neural networks. In *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*, 253–258 (IEEE, 2016).

**47.** Ribeiro, E., Uhl, A., Wimmer, G. & Häfner, M. Exploring deep learning and transfer learning for colonic polyp classification. *Comput. mathematical methods medicine* **2016** (2016).

**48.** Byrne, M. F. *et al.* Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. *Gut* **68**, 94–100 (2019).

**49.** Golhar, M. *et al.* Improving colonoscopy lesion classification using semi-supervised deep learning. *IEEE Access* **9**, 631–640 (2020).

**50.** Koch, G., Zemel, R., Salakhutdinov, R. *et al.* Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, vol. 2 (Lille, 2015).

**51.** Everingham, M. *et al.* The pascal visual object classes challenge: A retrospective. *Int. journal computer vision* **111**, 98–136 (2015).

**52.** Tan, M., Pang, R. & Le, Q. V. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10781–10790 (2020).

**53.** Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988 (2017).

**54.** Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. neural information processing systems* **28**, 91–99 (2015).

**55.** Song, T., Li, H., Meng, F., Wu, Q. & Cai, J. Letrist: Locally encoded transform feature histogram for rotation-invariant texture classification. *IEEE Transactions on circuits systems for video technology* **28**, 1565–1579 (2017).

**56.** Song, T., Feng, J., Luo, L., Gao, C. & Li, H. Robust texture description using local grouped order pattern and non-local binary pattern. *IEEE Transactions on Circuits Syst. for Video Technol.* **31**, 189–202 (2020).

**57.** Roy, S. K., Bhattacharya, N., Chanda, B., Chaudhuri, B. B. & Ghosh, D. K. Fwlbp: a scale invariant descriptor for texture classification. *arXiv preprint arXiv:1801.03228* (2018).

**58.** Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

**59.** Howard, A. G. *et al.* Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

**60.** He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

**61.** Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018).

**62.** Fonollà, R. *et al.* Triplet network for classification of benign and pre-malignant polyps. In *Medical Imaging 2021: Computer-Aided Diagnosis*, vol. 11597, 1159731 (International Society for Optics and Photonics, 2021).

## Acknowledgements

Iwahori's research is supported by JSPS Grant-in-Aid Scientific Research (C)(#20K11873).

## Author contributions

P.S, A.P and M.K.B designed the entire study. P.S and A.P simulated the algorithm. P.S and M.K.B wrote the manuscript. Y.I was involved during formulation of the problem. K.K and N.O analyzed the datasets and validated the performance of the algorithm. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.