Scientific African 20 (2023) e01628

Contents lists available at ScienceDirect

Scientific African

journal homepage: www.elsevier.com/locate/sciaf

Detection of gastrointestinal tract disorders using deep learning methods from colonoscopy images and videos

Salih Aliyi[1], Kokeb Dese[1,2,3,*], Hakkins Raj[1,*]

[1] *School of Biomedical Engineering, Jimma Institute of Technology, Jimma, 378, Ethiopia*
[2] *Artificial Intelligence and Biomedical Imaging Research Lab, Jimma Institute of Technology, Jimma University, Jimma, Ethiopia*
[3] *giCentre, Department of Computer Science, School of Science and Technology, City University of London, London EC1V 0HB, UK*

ARTICLE INFO

Article history:
Received 28 October 2022
Revised 10 January 2023
Accepted 3 March 2023

Editor DR B Gyampoh

Keywords:
Colorectal cancer
Gastrointestinal tract
Object detection
YoloV5
Mean average precision

ABSTRACT

Colorectal cancer (CRC) is the world's third most common cancer, with the second highest fatality rate. It is primarily the result of lower gastrointestinal tract (GI) disorders. The prevention of CRC mainly depends on the early detection and treatment of anomalies in the lower GI tract. Colonoscopy is the gold standard device used for diagnosing abnormalities in the lower GI tract as well as identifying anatomical landmarks and bowel preparation scales. However, it is time-consuming, tedious, and prone to error process, especially for those hospitals in low resource settings. Therefore, in this research, a real-time automated detection, classification, and localization of lower GI tract pre-colorectal cancerous abnormalities were done. The proposed system enables real-time detection, classification, and localization of common pathology, anatomical landmarks, and bowel preparation scale from colonoscopy images. To do the research, data was gathered both online (at hyper k-vasir dataset) and locally from the Yanet Internal Specialized Center and the Ethio-Tebib Hospital. Data augmentation techniques were applied to increase the training dataset. The pre-trained transfer learning SSD, YOLOv4, and YOLOv5 object detection model was used to develop the system with minimal fine-tuning of the hyper parameters and their performance was compared. The Yolo v5 model achieves good precision, recall, and mean average precision (mAP), 99.071%, 98.064% and 98.8%, respectively, on the testing data set. The developed artificial intelligence-based module would have the potential to assist gastroenterologists and general practitioners in decision-making. Even though the proposed work achieved the best performance, further improvement is required by increasing the size of the dataset to include other GI tract disease diagnoses.

Abbreviations: AI, Artificial intelligence; BBPS, Boston bowel preparation scale; CNN, Convolutional neural network; CRC, Colorectal cancer; CT, Computed tomography; DL, Deep learning; GI, Gastro intestinal; GUI, Graphical user interface; IOU, Intersection over union; mAP, Mean average precision; ML, Machine learning; MLP, Multi-layer perceptron; MRI, Magnetic resonance imaging; NN, Neural network; PET, Positron emission tomography; RCNN, Region convolutional neural network; RPN, Region proposal network; SPECT, Single photon emission computed tomography; SSD, Single shot multi box detector.

* Corresponding authors.
E-mail addresses: kokeb-dese.gebremeskel@city.ac.uk (K. Dese), hakkinsbme@gmail.com (H. Raj).

https://doi.org/10.1016/j.sciaf.2023.e01628

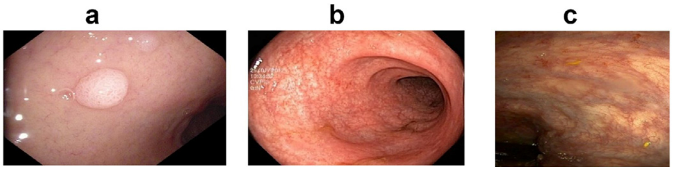


Fig. 1. a) Polyps, b) Ulcerative Colitis, c) hemorrhoid images acquired from Yanet Internal Specialized Center.

Background

The gastrointestinal (GI) tract extends from the esophagus to the anus. The lower GI tract encompasses the area between the anus and the ileum, which is the distal portion of the small intestine. Cancer is the second leading cause of death in humans, following cardiovascular disease [1]. Nowadays, the cancer mortality rate is increasing [2], and colorectal cancer (CRC), which happens in the lower GI tract is the third most diagnosed and second mortality-causing cancer in the world [3,4]. In addition to this, unlike in Western countries, in Ethiopia CRC prevalence is higher in the young population [5]. Moreover, it leads to lower GI tract abnormalities and lesions if not detected at an early stage and proper treatment is not given [1]. For lower GI tract cancer prevention, the detection, treatment, and removal of precancerous lesions in the lower GI tract are very essential. Physicians use a colonoscopy to detect and classify pathology in the lower GI tract, and the disease is overlooked at an early stage and misclassified by physicians. Therefore, medical image processing and artificial intelligence based automatic detection and classification of this abnormality and lesions is a key to overcoming the problem. Moreover, the real-time detection of pathology, anatomical landmarks, and bowel preparation scale from colonoscopy video is essential for saving time, reducing operator effort and misdiagnosis. The common pathologies in the lower GI tract that are screened by colonoscopy are polyps, ulcerative colitis, and hemorrhoids [6,7]. From those pathology polyps and UC are the most common precursors of CRC [8,9,10]. As the study at St. Paul's Hospital Millennium Medical College (SPHMMC) indicates, the most prevalent disease diagnosed by colonoscopy in Ethiopia is hemorrhoid [7,11]. The study also shows the incidence of polyps and UC has significantly increased in our country, which are the major causes of CRC. All these abnormalities and lesions have developed on the mucosal wall of the intestine. To screen the intestine mucosal wall using colonoscopy, the wall needs to be clean. Kutyla et al. (2021) discovered that bowel preparation plays an important role in the success and quality of lower GI tract pathological findings [12]. So good bowel preparation leads to better diagnosis of pathologies in the lower GI tract using colonoscopy.

Patients with lower GI tract disorder manifest abdominal pain, blood on or in the stool that is either bright or dark, distension, diarrhea, constipation, accidental stool leakage, or incontinence [8]. The most effective way to control GI disease and the incidence of CRC is early detection and early treatment of the disease [8,9,10]. This study focused on the diagnosis of three common lower GI tract diseases: polyps, ulcerative colitis, and hemorrhoid. Furthermore, the study includes the detection of essential anatomical landmarks and a bowel preparation scale. The criteria for disease selection were their relevance to form CRC and their prevalence [7,11,13]. Polyps are lesions within the intestine observable as mucosal outgrows. The common type of polyps is premalignant and causes colorectal cancer if it's not removed early (see the Polyps pathology in Fig. 1, a)[14,10]. Ulcerative colitis (UC) on the other hand (see Fig. 1 b) is the other lower GI tract disease that is caused by the mucosal inflammation that starts from the rectum and can extend to the entire colon. UC patients have a high risk of developing colorectal cancer [15,16]. Moreover, hemorrhoids (see Fig. 1 c) are swollen or dilated veins of the anus or rectum. It may be located just inside the anal canal (internal hemorrhoid) or surrounding the anal opening (external hemorrhoids). Hemorrhoids may be present for years but go undetected until bleeding occurs [17,10]. Colonoscopy can even detect before bleeding occurs.

Currently, clinicians may use a mixture of clinical symptoms, laboratory indicators, radiation monitoring, endoscopy, and histological analysis of tissue samples to assess disease occurrence and make treatment decisions in the lower GI tract. Moreover, a combination of nanotechnology and medical imaging methodologies such as computed tomography (CT), single photon emission computed tomography (SPECT), positron emission tomography (PET), and magnetic resonance imaging (MRI) is used for lower GI tract disease diagnosis, disease severity monitoring [18,15]. Nevertheless, colonoscopy can provide in situ imaging of mucosal lesions and abnormalities diagnosis. Capsule endoscopy is a vitamin pill size endoscopy that is used to screen the whole GI tract. The patient swallows the capsule, which moves through the digestive tract taking thousands of pictures and wirelessly sending them to the recording device, and then the physician will examine it [19]. This study aimed to design and develop a real-time detection and classification of precancerous lesions and other abnormalities in the lower GI tract from colonoscopy images using the Yolov5 deep learning model. This is significant especially for developing countries like Ethiopia, where both the experts and the resources are scarce.

So far, a method of automatically detecting and classifying GI tract disease from clinical images has been proposed in several works of literature. The majority of them use machine learning and deep learning techniques to build a model for diagnosing GI tract illnesses from clinical images [20,21,22,23,24]. A research work reported in [25] used Global feature (GF), deep CNN, and transfer learning (Inception-v3) for the classification of GI tract anatomical findings, including esophagitis, polyps, and UC, and anatomical landmarks such as Z-line, pylorus, and cecum. Those methods were developed as baseline

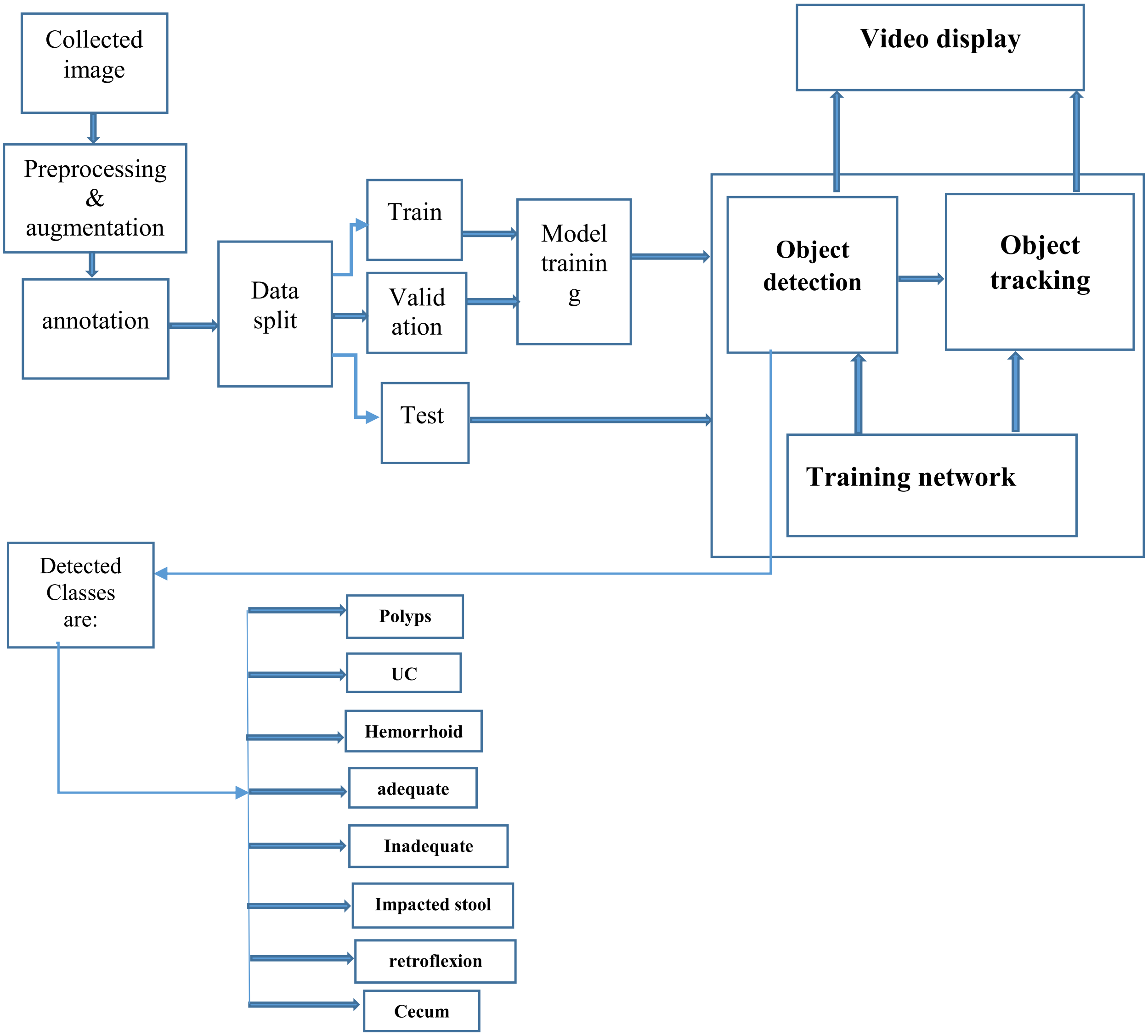


Fig. 2. General block diagram of the system.

performance for the k-vasir dataset. From the system, GF achieved an accuracy of 93.7%, deep CNN achieved 95.9%, and Inception-v3 achieved 92.4%. At the same time, an Inception-based CNN architecture was proposed for the classification of the same classes in [24]. The task's purpose is to classify diseases (as efficiently as feasible) with as little training data as possible by decreasing the parameters in CNN. It uses Google Net as a base and achieves 93.9% accuracy, which is a better achievement than GF and Inception-v3, but less than Deep CNN. Moreover, blurry, cecum, normal, polyps, tumor, and Z-line are five findings from capsule endoscopy with real-time polyp localization that are classified by [26]. They used Darknet-YOLO and TensorBox for polyp localization, which had a greater accuracy of 96.9% but were too sluggish to be deemed real-time. In [22] they used Inception-v4, Inception-ResNet-v2, and NasNet models to classify only 8 classes of GI diseases, which include 3 pathological findings (esophagitis, ulcerative colitis and polyps), 3 anatomical land marks (pylorus, z-line and cecum) and 2 medical procedures (dyed lifted polyps and dyed resection margins). They used the kvasir dataset. This paper used a robust preprocessing method, which leads to better accuracy. Inception-v4, Inception-ResNet-v2, and NasNet each achieve 98.45%, 98.48%, and 97.35% accuracy respectively. In this research, bowel preparation quality, which is very essential for the lower GI tract detection and classification, is not included.

The other related work on the top of the GI tract anatomical and pathological classification is the global feature (GF) approach based on 16 classes. The classified classes are blurry-nothing, colon-clear, dyed-lifted-polyps, dyed-resection-margins, esophagitis, instruments, normal-cecum, normal-pylorus, normal-z-line, out-of-patient, polyps, retroflex-rectum, retroflex-stomach, stool-inclusions, stool-plenty, and UC, whose data is unbalanced from Nerthus and k-vasir datasets. It fetches 6 features from the image, each of which can represent the whole properties of the image and transfer them to the machine learning classifier. The properties include Pyramid Histogram of Oriented Gradients (PHOG), Color Layout, Tamura, Edge Histogram, Auto Color Correlogram, and Joint Composite feature (JCD). Two methods of GFs approach are based on what takes place. The difference is the classifier they use. The first one uses a simple logistic (SL) classifier while the second one uses a logistic model tree (LMT) [27]. Because the data they utilize is imbalanced, both models perform poorly, with an accuracy of 82%. The transfer learning strategy outperforms the GF approach in terms of accuracy and performance. Transfer learning is superior because it employs pre-trained networks that can quickly learn common visual features such as curves, lines, and edges.

The pre-trained methods that were developed for the classification of anatomical landmarks and pathological findings of the GI tract are Resnet-152 and Densenet-161 and their combinations, which are pre-trained on the ImageNet dataset and selected based on top-one error and top-five errors from pre-trained networks in Pytorch. The method includes Resnet-152 alone, a combination of Resnet-152 + Densenet-161 and Resnet-152 + Densenet-161 + MLP (multilayer perceptron). All of them are used to classify anatomical landmarks and pathological findings, which cover the 16 classes that were mentioned above from the Nerthus [28] and kvasir datasets along the whole GI tract. Additionally, the same methods were selected from nice performing methods presented at the MediaEval Medico Task [29,30] for the classification of 23 classes of the hyper k-vasir dataset as base performance. Image data in the hyper k-vasir dataset is highly unbalanced. collected both from the upper GI tract and the lower GI tract. It includes pathological findings in the upper and lower GI tract, anatomical landmarks in the upper and lower GI tract, quality of mucosal view, and therapeutic interventions. The pathological findings in the upper GI tract are Barrett's esophagus, short-segment Barrett's, esophagitis-a, and esophagitis-b-d. The anatomical landmarks in the upper GI tract are the Z-line, pylorus, and retroflex stomach. The pathological findings in the lower GI tract are polyps, hemorrhoids, UC grade 0-1, UC grade 1-2, UC grade 2-3, UC grade 1, UC grade 2, and UC grade 3. The anatomical landmarks in the lower GI tract are the cecum, retroflex rectum, and ileum. The quality of mucosal views are BBPS-0-1, BBPS-2-3, and impacted stool. The therapeutic interventions are dyed lifted polyps and dyed resection margins. The combination of two pre-trained models with MLP outperforms all other methods because the result is not only the average of the two pre-trained models but MLP is also added to perform complex mathematical formulas for better cumulative decisions. MLP has 32 inputs through which 16 probabilistic outputs of the two pre-trained models enter into it. At the end of the day, MLP processes the outputs for a better decision of the result [21]. Even though this paper covers more classes, the data it utilizes is very imbalanced. More importantly, the classification of pathological abnormalities and anatomical landmarks from colonoscopy images is the focus of everyone. However, because categorization does not show the precise location of a problem on an image, no localization is achievable. Furthermore, there hasn't been enough effort reported on localization and detection of lower GI tract findings from video. The proposed work use augmentation to balance the data and a different annotation file is prepared (to localize exact position of disorders). An XML annotation file and a text annotation file are prepared for 8000 images with collaborators. Overall, 16000-annotation files are prepared to train different pre-trained object detection algorithms. The proposed models are used to detect, classify, and localize pathology, anatomical landmarks, and bowel preparation scale in the lower GI tract from both colonoscopy images and videos. Eventually developed models on custom data are compared and a graphical user interface (GUI) was built for the preferred model of the proposed models.

Materials and methods

This work followed the experimental research type, which includes image data collection from local hospitals and online, preprocessing and augmentation of the data, annotation of data, and the training, validation, and test sets of prepared image data were separated. 70% of the data, which consisted of 5600 images, was utilized for training, 15% for validation, and 15% for the test. Both validation and test data consisted of 1200 images. Brief explanations are provided in the following sections. The following Fig. 2 shows the general block diagram of the developed system.

Data collection

The colonoscopy images were collected from Yanet Internal Specialized Center and Ethio-Tebib Hospital. Fig. 3 below shows data collection and annotation with the help of three clinical collaborators. In addition to this hyper kvasir dataset (online data), which is collected from the Department of Gastroenterology, Baerum Hospital in Norway is also included. Both Yanet Internal Specialized Center and Ethio-Tebib Hospital have a digital colonoscopy that is used to store screened images

Generally, 4508 images were obtained from an online database and 1814 images were collected from local hospitals, where 1000 were collected from Yanet Internal Specialized Center and the remaining 814 were from Ethio-Tebib Hospital. Even though the collected data is not equal over classes, the augmentation method was applied to balance the number of images per class.

Preprocessing and augmentation

Initially, the images were found in different formats, then all of them were converted to the same format before augmentation takes place. To balance the data and increase the amount of training data, we used different augmentation techniques including rotation, vertical flip, and scaling (see Fig. 4) [31]. Image augmentation is the process of applying various changes

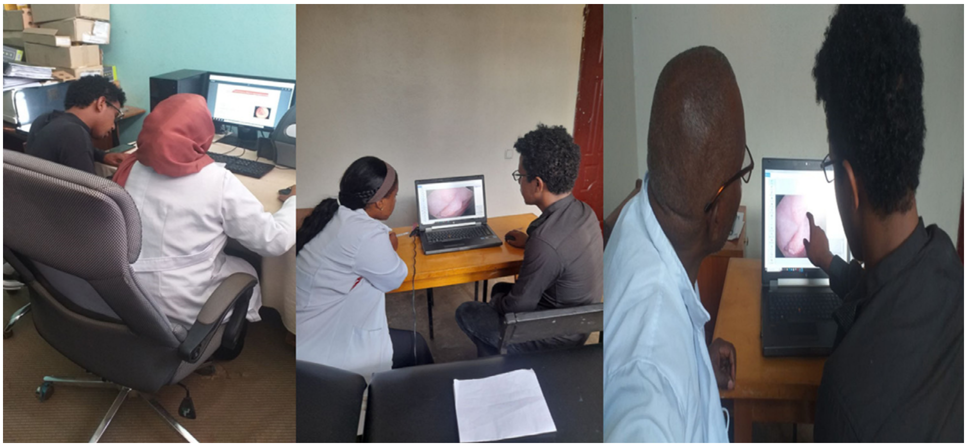


Fig. 3. Data collection and labeling with clinical collaborators.

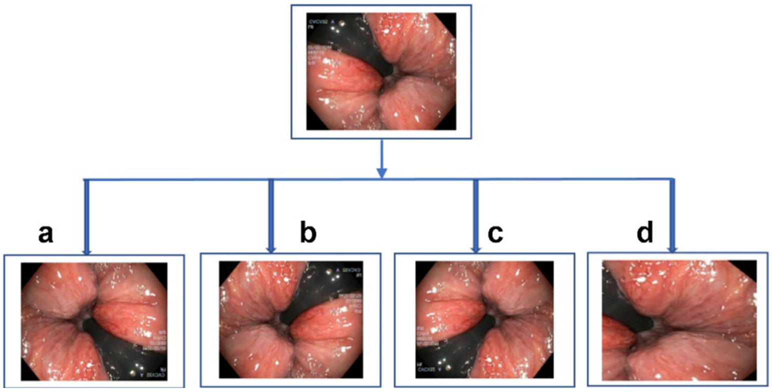


Fig. 4. Different augmentation steps applied. a) rotation with 180°, b) rotation with 90°, c) vertical flip, d) scaling.

Table 1
Summary of image data collected and used for developing the system

No of Item	Name of objects(classes)	No of image data collected from online	No of image data collected from Ethio-Tebib	No of image data collected from Yanet	No of image data produced by augmentation	Total
1	Polyps	1000	0	0	0	1000
2	hemorrhoid	6	294	300	400	1000
3	UC	334	116	250	300	1000
4	cecum	1000	0	0	0	1000
5	Retroflexed rectum	391	100	200	309	1000
6	Impacted stool	131	200	100	569	1000
7	BBPS0-1(Inadequate)	646	104	150	100	1000
8	BBPS2-3(normal)	1000	0	0	0	1000
TOTAL		4508	814	1000	1678	8000

to original images to create many altered copies of the same image. Image augmentation not only increases the size of our dataset but also provides different levels of variance to the training data and allows our model to generalize more effectively to new data. This has increased the training dataset.

The following Table 1 summarizes the total number of data used per class including the data after data augmentations.

Model training

In any deep learning task, the first and most important step is to prepare the data. Data collection, data preprocessing, data augmentation, and data annotation occurred in this research for data preparation. After the data was prepared, it was divided into train, validation, and test dataset folders, which are created for both images and labels. Google colab was used for training, which is used to train the algorithms online. The object detection pre-trained deep learning models SSD,

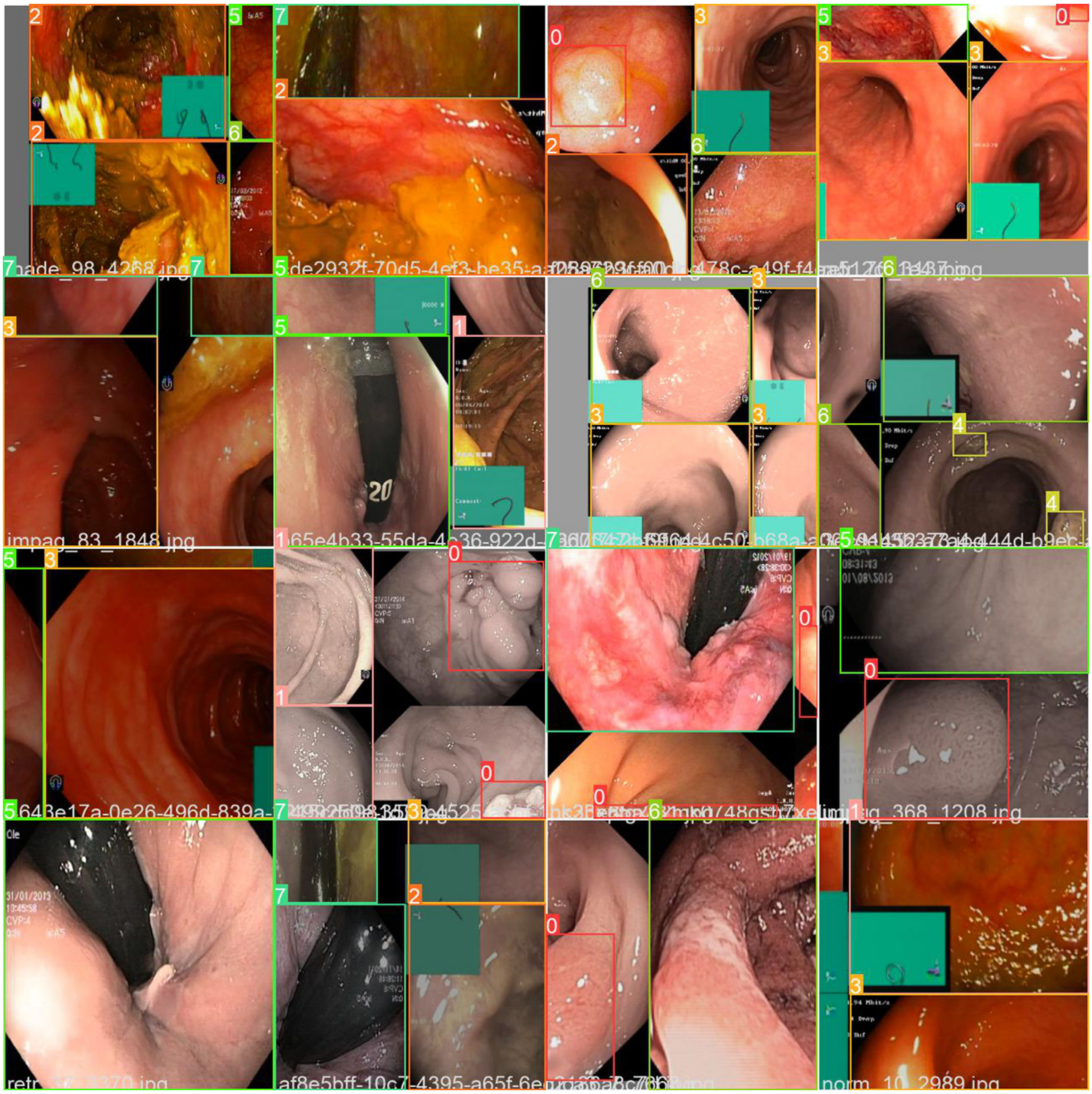


Fig. 5. Sample of training data batch.

YOLOv4, and YOLOv5 were trained on the custom data by adjusting image size, batch size, epoch, and some hyperparameters. The amount of custom data prepared for training was 5600 images for training and 1200 images for validation. An XML file was used to train SSD with image data, while a txt file was used to train YOLOv4 and YOLOv5 with prepared image data. During training, the image size was adjusted to 320×320 and the batch size and epoch were adjusted to 32 and 60, respectively, for all models. In the same way, to achieve better results, the learning rate was decreased to 0.0001 and more filters were changed. Eventually, the ideal model was selected, and an independent Windows application was created for it. The application was developed using Qt Designer and python. Fig. 5 shows a sample of the training batch.

After training of the models on custom data by adjusting hyper-parameters, the developed models were compared and eventually, a GUI was developed for the best model.

The materials used in this study are stated in Table 2 below.

Performance Evaluation Metrics

A detection is represented by three attributes: the object class, the accompanying bounding box, and the confidence score, which is usually a number between 0 and 1 that indicates how sure the detector predicts the class accurately. In the case of object detection, the employed evaluation metrics measure how close the detected (predicted) bounding boxes

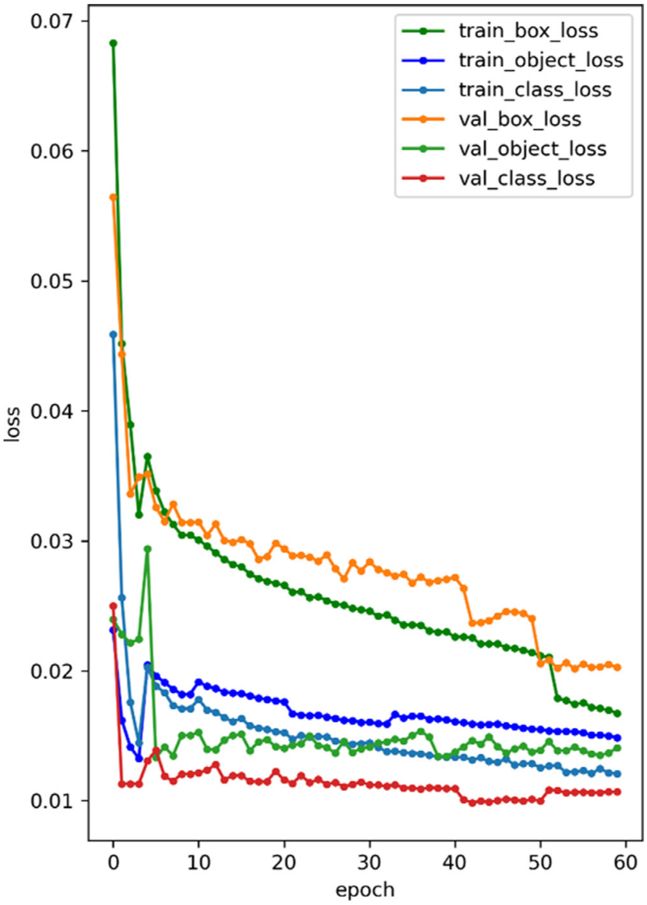
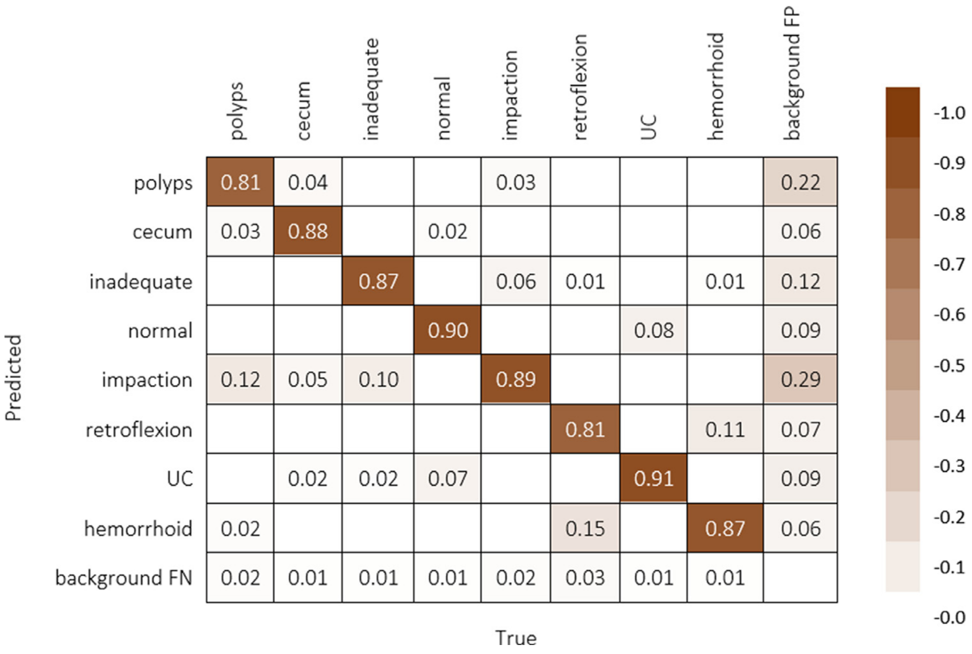


Fig. 6. The training and validation sets loss functions for SSD.



	polyps	cecum	inadequate	normal	impaction	retroflexion	UC	hemorrhoid	background FP
polyps	0.81	0.04			0.03				0.22
cecum	0.03	0.88		0.02					0.06
inadequate			0.87		0.06	0.01		0.01	0.12
normal				0.90			0.08		0.09
impaction	0.12	0.05	0.10		0.89				0.29
retroflexion						0.81		0.11	0.07
UC		0.02	0.02	0.07			0.91		0.09
hemorrhoid	0.02					0.15		0.87	0.06
background FN	0.02	0.01	0.01	0.01	0.02	0.03	0.01	0.01	

Fig. 7. Confusion matrix for SSD.

Table 2
List of software and hardware materials used in the study

Software materials	Hardware materials
Labelimg annotation tool Python 3.8 Qt designer	Colonoscopy machine (Olympus and Karl Storz models) HP,Processor: Intel(R) Core(TM) i7-2670QM CPU @ 2.20GHz, 2201 Mhz, 4 Core(s), 8 Logical Processor(s), 12 GB RAM, 16 bit operating system, window 10

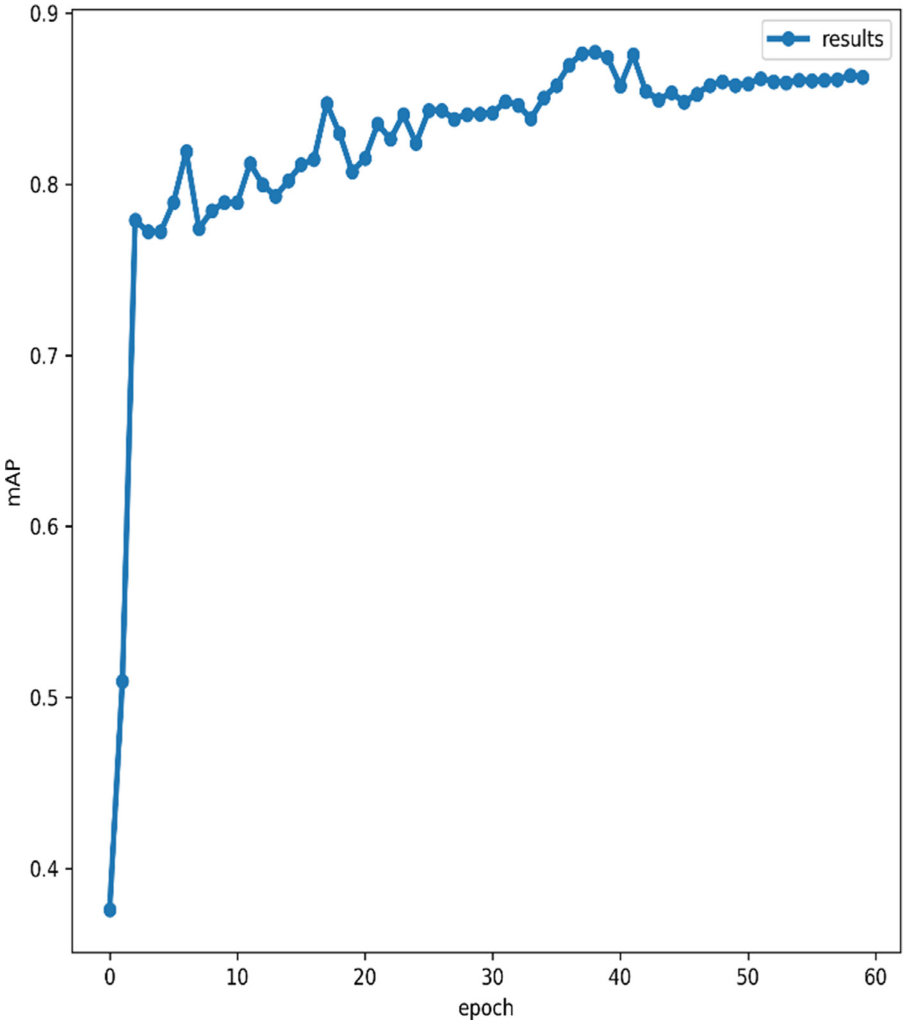


Fig. 8. mAP for SSD model over 60 epochs.

by the model was to the ground truth (hand-labeled) bounding boxes. This is done for each object class separately, by calculating intersection over union (IOU), or the amount of overlap between the predicted and ground-truth areas. Let the bounding box predicted by the model be Bp and the ground truth bounding box be Bgt. Then IOU is the bounding box intersection of Bp and Bgt divided by the bounding box union of Bp and Bgt, which is expressed as the following Eq.1[32].

$$\mathrm{IOU} = \frac{\mathrm{Bp} \cap \ \mathrm{Bgt}}{\mathrm{Bp} \cup \ \mathrm{Bgt}} \tag{1}$$

The better the detection, the closer the IOU becomes to 1.

On the other hand, the confidence score represents how confident the model is that the box includes an object as well as how accurate the box's predictions are. The method for finding the confidence score is as described in the following Eq.2 [32,31]:

$$\mathrm{C} = \mathrm{pr(obj)} * \mathrm{IOU} \tag{2}$$

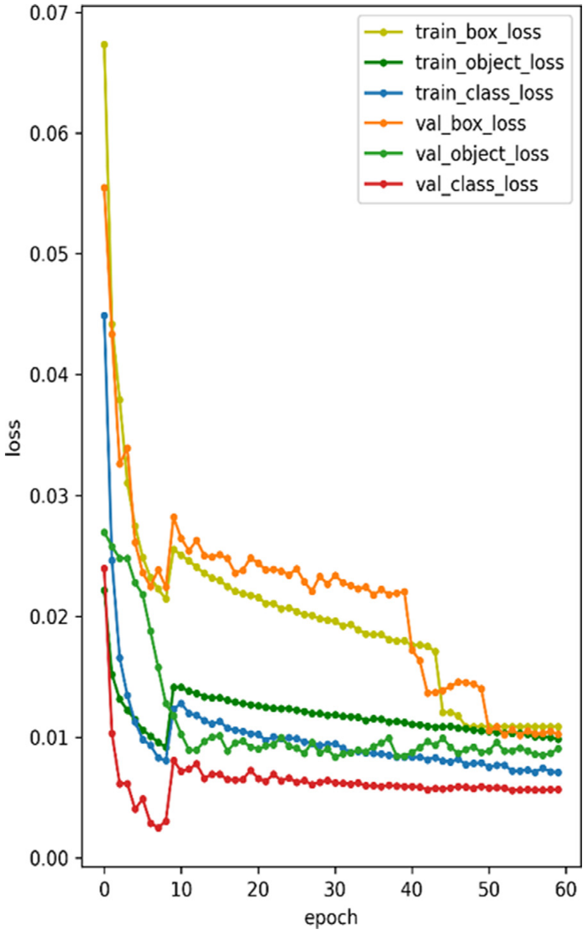


Fig. 9. The training and validation sets loss for YOLOv4.

Where C represents confidence and Pr (Object) denotes the probability that the cell contains an object in the expected bounding box. The IOU is used to define the most important object detection metrics like precision, recall, f1 score, average precision (AP), and mean average precision (mAP). The proportion of correct positive classification from positive prediction is measured by precision, while the proportion of correct positive prediction from the real truth is measured by a recall. The F1 score is a comparison indicator between precision and recall numbers that determine the most ideal confidence score threshold where precision and recall produce the maximum F1 score. The average precision (AP) is a method of condensing the precision-recall curve into a single number that represents the average of all precisions. The mean of the APs for all classes is called the mean average precision (mAP). Eq.3, Eq.4, Eq.5, Eq.6, and Eq.7, below depict the mathematical representation for precision, recall, f1 score, Ap, and mAP respectively. A confusion matrix is a N x N matrix that is used to evaluate the performance of a classification or detection model, with N denoting the number of target classes. On the other hand, it determines how much the model is confused while detecting different classes[32,24].

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4}$$

$$\mathrm{F1} = 2*\left(\frac{\mathrm{P} * \mathrm{R}}{\mathrm{P} + \mathrm{R}}\right) \tag{5}$$

$$\mathrm{AP@}\alpha = \int_0^1 \mathrm{p(r)dr} \tag{6}$$

$$\mathrm{mAP} \ @ \ \alpha = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{APi} \tag{7}$$

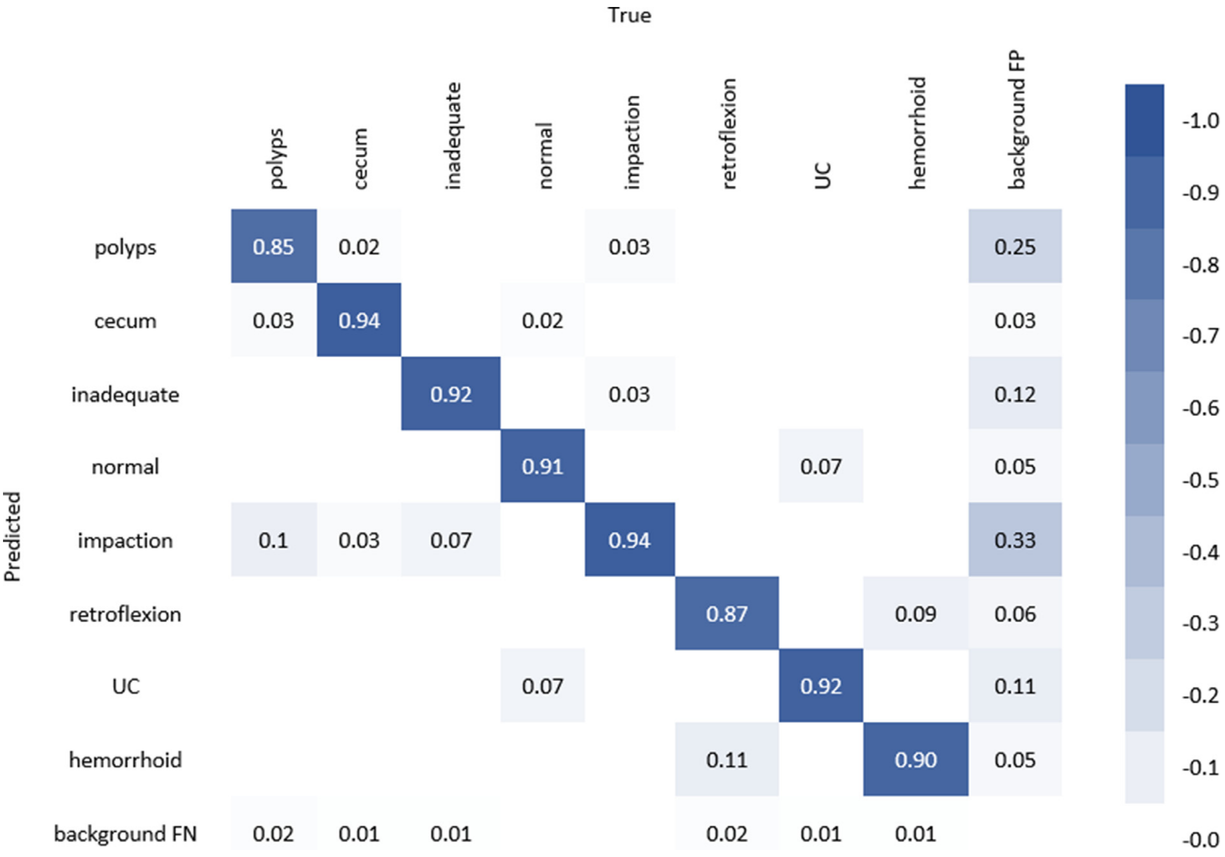


Fig. 10. Confusion matrix for YOLOv4.

utilized Terms are defined as follows:

- True Positive (TP) – The model correctly detects the disease.
- False Positive (FP) – The model incorrectly detects the disease.
- False Negative (FN) – An object detector misses (does not detect) a ground truth.
- True Negative (TN) —This is a background region that the model has appropriately missed. Because such locations are not explicitly labeled when producing the annotations, this measure is not employed in object detection.

Results

For all models, the total number of clinical images utilized for 1 training, validation, and testing, as well as the accompanying annotation files, was 5600, 1200, and 1200, respectively. Each models result was discussed below.

Single-shot detector (SSD) result

The pre-trained SSD model was trained using clinical images and XML files for 60 epochs. The model achieves 87.5% mAP at an IOU threshold of 0.5. The lowest train box loss, train object loss, and train class loss for the model are 0.016721, 0.013232, and 0.01208, respectively. Similarly, 0.020179, 0.013311, and 0.009852 are the lowest validation box loss, validation object loss, and validation class loss, respectively. Fig. 6 below shows the loss of train and validation over 60 epochs.

The model also achieves 90.42% precision and 91.32% recall results on unseen image test data. Furthermore, eight videos each including all instances of the class are also used to test the model. The model correctly detects and localizes 59 objects from an instance of 64 objects in eight test videos. A confusion matrix compares actual target objects to model predictions. The model scores polyps (81%), UC (91%), hemorrhoids (87%), normal (90%), inadequate (87%), impacted stool (89%), retroflexed rectum (81%), and cecum (88%). The developed SSD model's confusion matrix is shown in Fig. 7 below.

Fig. 8 below presents the experimental result output in terms of the mAP of the model over a given epochs.

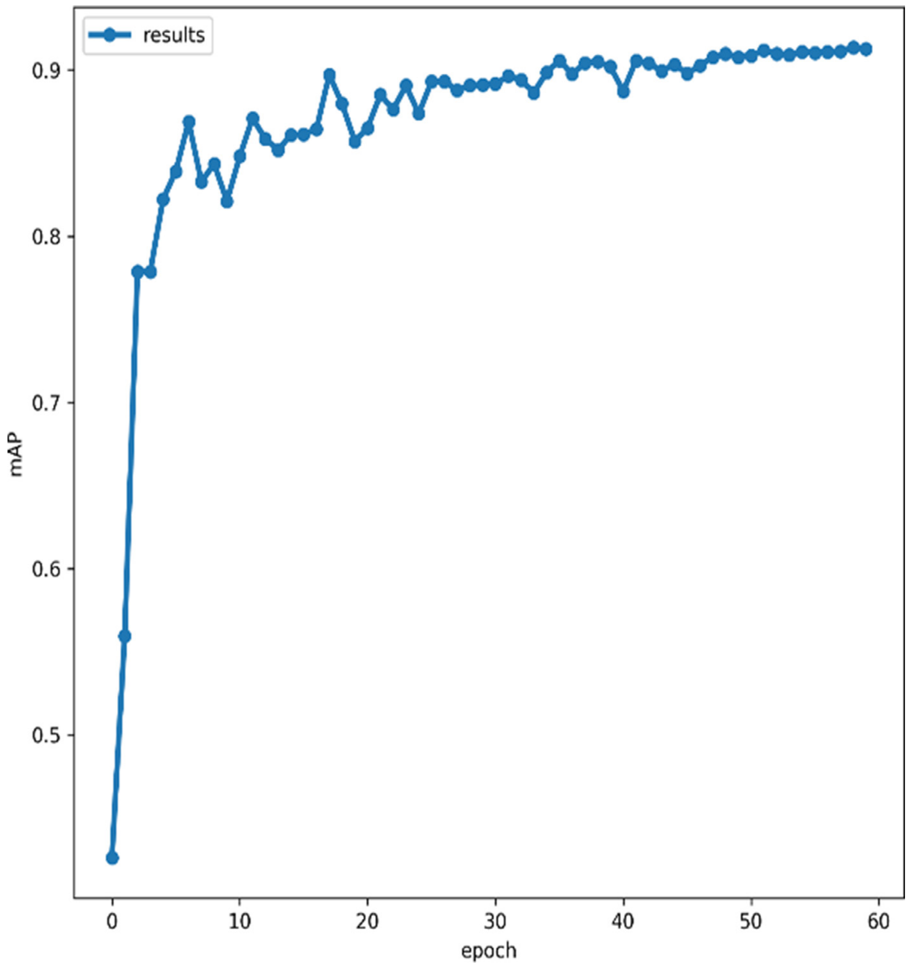


Fig. 11. mAP for YOLOv4 over 60 epochs.

Using YOLOv4 model result

The pre-trained YOLOv4 model was trained by using clinical images and text files for 60 epochs. The model achieves 92.7% mAP at an IOU threshold of 0.5. The lowest train box loss, train object loss, and train class loss for the model are 0.010866, 0.009193, and 0.00708, respectively. In the same way, 0.016179, 0.008343, and 0.002506 are the lowest validation box loss, validation object loss, and validation class loss, respectively. Fig. 9 below shows the loss of train and validation over 60 epochs. The model also achieves precision and recall results of 93.02% and 92.83%, respectively, on unseen image test data.

A confusion matrix compares actual target objects to model predictions. The YOLOv4 model scores polyps (85%), UC (92%), hemorrhoids (90%), normal (91%), inadequate (92%), impaction (94%), retroflexed rectum (87%), and cecum (94%). The developed YOLOv4 model's confusion matrix is shown in Fig. 10 below. Additionally, eight videos each including all instances of the class are also used to test the model. The model correctly detects and localizes 62 objects from an instance of 64 objects in eight testing videos.

Fig. 11 depicts the mAP over epochs experimental result of YOLOv4 model.

Using YOLOv5 model result

The pre-trained YOLOv5 model was trained by using clinical images and text files for 60 epochs. The model achieves 98.8% mAP at an IOU threshold of 0.5. This best mean average precision was achieved with the lowest losses. The lowest train box loss, train object loss, and train class loss for the model are 0.009721, 0.004852, and 0.00208, respectively. In the same way, 0.015179, 0.003343, and 0.000602 are the lowest validation box loss, validation object loss, and validation class loss, respectively. Fig. 12 below shows the loss of train and validation over 60 epochs. The model also achieves the ideal precision and recall results on unseen image test data with 99.071% and 98.064%, respectively. The model is also

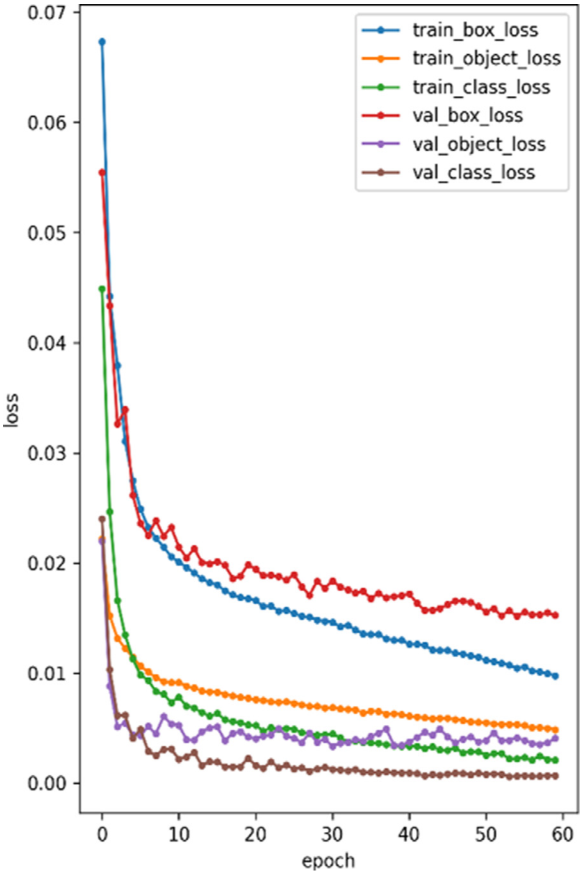


Fig. 12. The training and validation sets loss functions are converged for YOLOv5.

tested using eight videos, each of which contains all instances of the class. All 64 test objects in test videos are successfully detected and localized by the model.

A confusion matrix compares actual target objects to model predictions. The model scores polyps (90%), UC (97%), hemorrhoids (100%), adequate (99%), inadequate (99%), impacted stool (99%), retroflexed rectum (97%), cecum (99%). The developed YOLOv5 model's confusion matrix is shown in Fig. 13. The values of the 8 numbers on the diagonal of the confusion matrix are all greater than 90%, and a very small number of values of the diagonal indicates that the model is confused between objects very little.

Fig. 14 shows the experimental results in terms of the mAP of the model over epochs. Yolov5 developed on custom data for the detection of lower GI tract disorders has a low latency and detects every frame of test video at a speed of 24 FPS.

Algorithm demonstration

From images and videos, the developed GUI allows the user to detect and localize common pathological findings, anatomical landmarks, and bowel preparation scales in the lower GI tract. By using the load file button, the user can choose a video or image from the directory. The user then presses the play button to see the result of the detection and localization as an image or video on the screen. The bounding box, object type, and confidence are all included in the detection result. Fig. 15 shows the pictures that were taken while the objects were being found in real-time processing from the video.

As depicted in Fig. 15 above, the designed model returns a diagnosis result by using the object's bounding box, putting its class on top of the object in word along with predicting accuracy (confidence score). The confidence score ranges from 0–1, which indicates how confident the model is that the box contains an object and also how accurate it realizes the box is predicted.

Discussions

Even though there are more tools available for diagnosing the lower GI tract, the process is time-consuming, tedious, and complicated [33]. Furthermore, since those technologies are operator-dependent, a gastroenterologist's lack of experience,

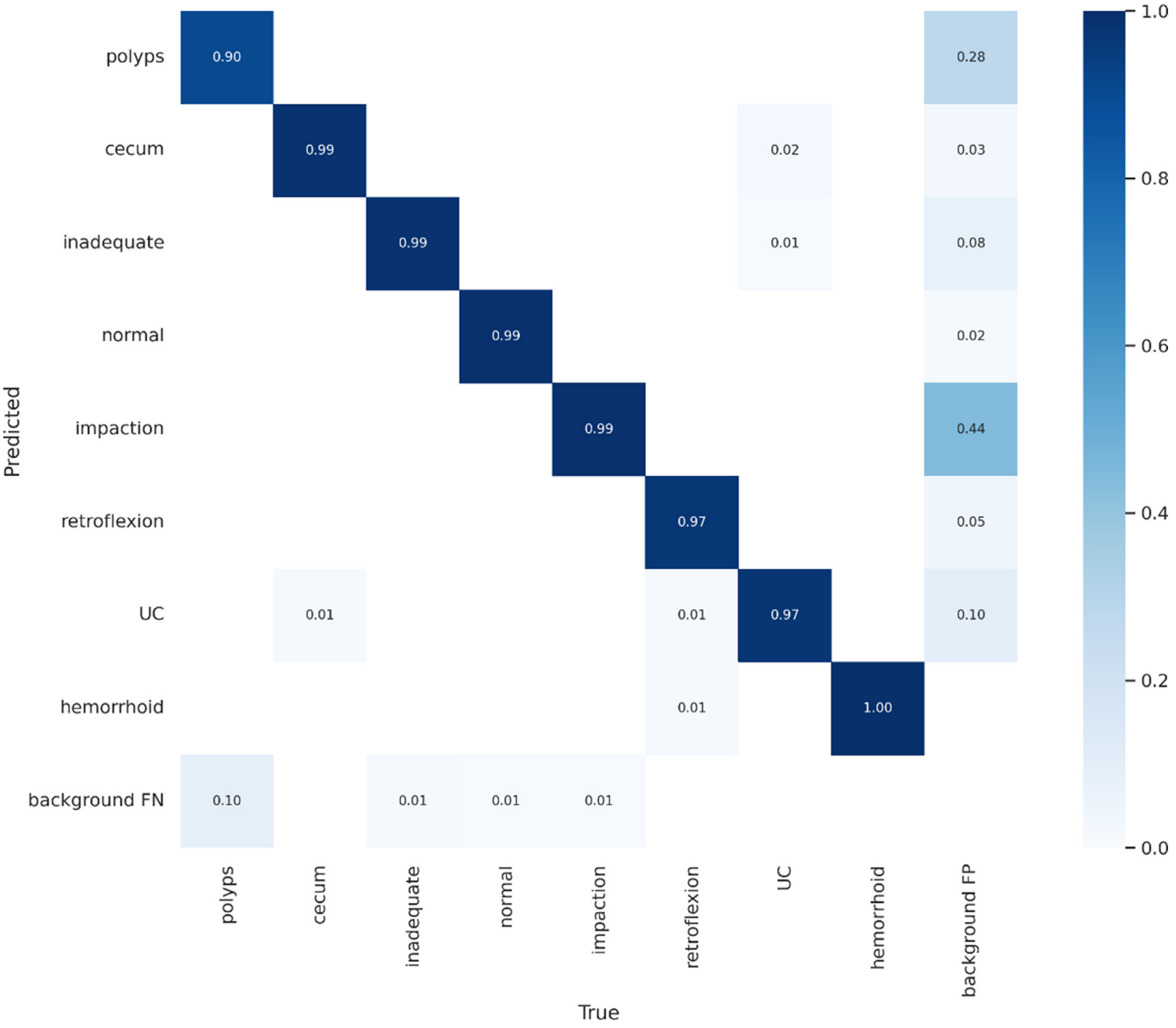


Fig. 13. Confusion matrix for YOLOv5.

fatigue, and loss of concentration during the examination leads to the missing and misclassification of disorders in the gastrointestinal tract[21,22,34]. The goal of this study was to design and develop an automatic detection and classification system for problematic abnormalities in the lower GI tract, as well as anatomical landmarks and a bowel preparation scale from image and video. The number of image data obtained online was 4508, with 1814 data collected locally (1000 from Yanet internal specialized center and 814 from Ethio-Tebib hospital). The total number of images obtained was 6322, which was letter boosted to 8000 to balance the data via augmentation.

The proposed approach detects and localizes common pathology in the lower GI tract in automatically from image and video, allowing the disease to be treated sooner rather than later before it progresses to CRC. The two most important factors in detecting pathology in a colonoscopy image or video are the quality of the mucosal view and the identification of anatomical landmarks in the lower GI tract so that the models can also detect and localize anatomical landmarks and bowel preparation scale in the lower GI tract images and videos.

Pretrained SSD, YOLOv4, and YOLOv5 were trained on the custom data (5600 images with equal label files for training and 1200 images with equal label files for validation) by adjusting image size, batch size, epoch, and some hyperparameters. The image size was adjusted to 320×320, and the batch size and epoch were adjusted to 32 and 60, respectively, for every model.

The pre-trained SSD model was trained using clinical images and XML files for 60 epochs. The model achieves 87.5% mAP at an IOU threshold of 0.5. The lowest train box loss, train object loss, and train class loss for the model are 0.016721, 0.013232, and 0.01208, respectively. Similarly, 0.020179, 0.013311, and 0.009852 are the lowest validation box loss, validation

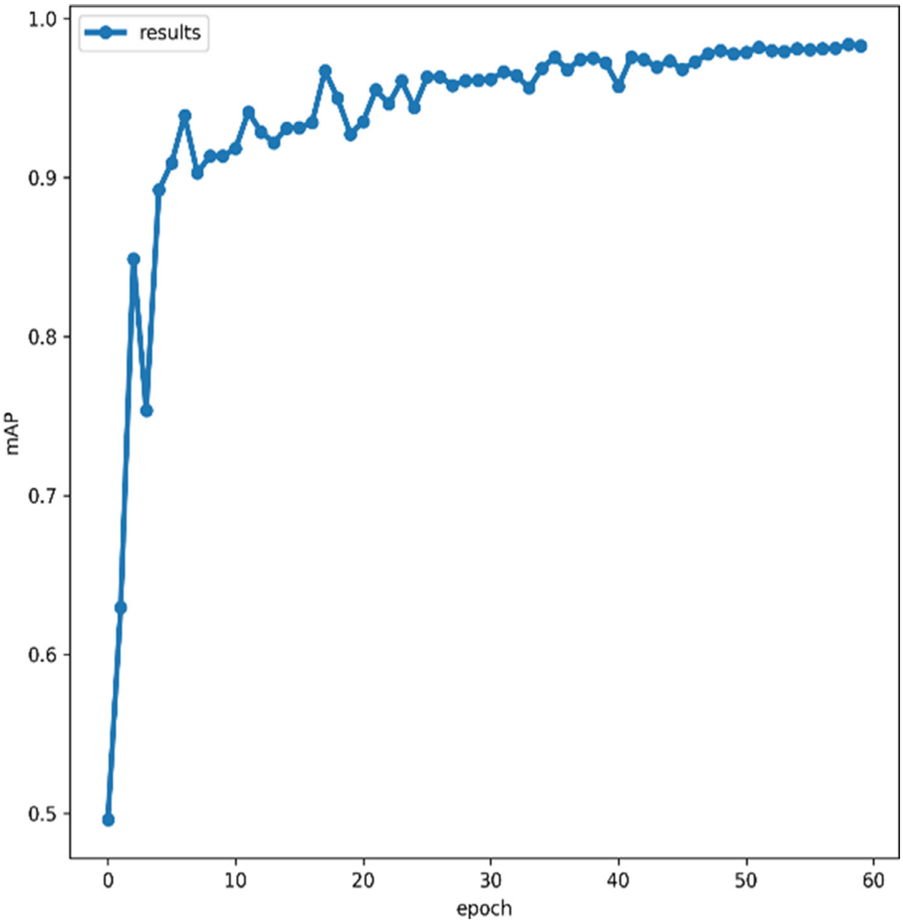


Fig. 14. mAP for YOLOv5 over 60 epochs.

Table 3
Comparison of train loss and validation loss for developed models

Models	Train box loss	Train object loss	Train class loss	Validation box loss	Validation object loss	Validation class loss
SSD	0.01672	0.01323	0.01208	0.02018	0.01331	0.00985
YOLOv4	0.01087	0.00919	0.00708	0.01618	0.00834	0.00251
YOLOv5	**0.00972**	**0.00485**	**0.00208**	**0.01518**	**0.00334**	**0.0006**

object loss, and validation class loss, respectively. The model also achieves 90.42% precision and 91.32% recall results on unseen test image data. Furthermore, the model correctly detects and localizes 59 objects from an instance of 64 objects in eight test videos.

The pre-trained YOLOv4 model was trained by using clinical images and text files for 60 epochs. The model achieves 92.7% mAP at an IOU threshold of 0.5. The lowest train box loss, train object loss, and train class loss for the model are 0.010866, 0.009193, and 0.00708, respectively. In the same way, 0.016179, 0.008343, and 0.002506are the lowest validation box loss, validation object loss, and validation class loss, respectively. The model also achieves precision and recall results on unseen test image data with 93.02% and 92.83%, respectively. Moreover, the model correctly detects and localizes 62 objects out of an instance of 64 objects in eight test videos.

The ideal performance was achieved using the YOLOv5 model. The error of the model training and validation fell dramatically for all boxes, objects, and classes, indicating that the model is strongly converged. 0.009721, 0.004852, and 0.00208 are the lowest train box, train object, and train class losses, respectively. Similarly, the lowest validation box loss, validation object loss, and validation class loss are 0.015179, 0.003343, and 0.000602, respectively. Table 3 below shows the comparison of losses for developed models in this thesis. The model achieves the ideal precision and recall results of 99.071% and 98.064%, respectively, on unseen test image data. The widely used metrics for assessing object detection algorithms are average precision (AP) and mean average precision (mAP). The model score was 94.6% for polyps, 99.4% for cecum, 99.3% for

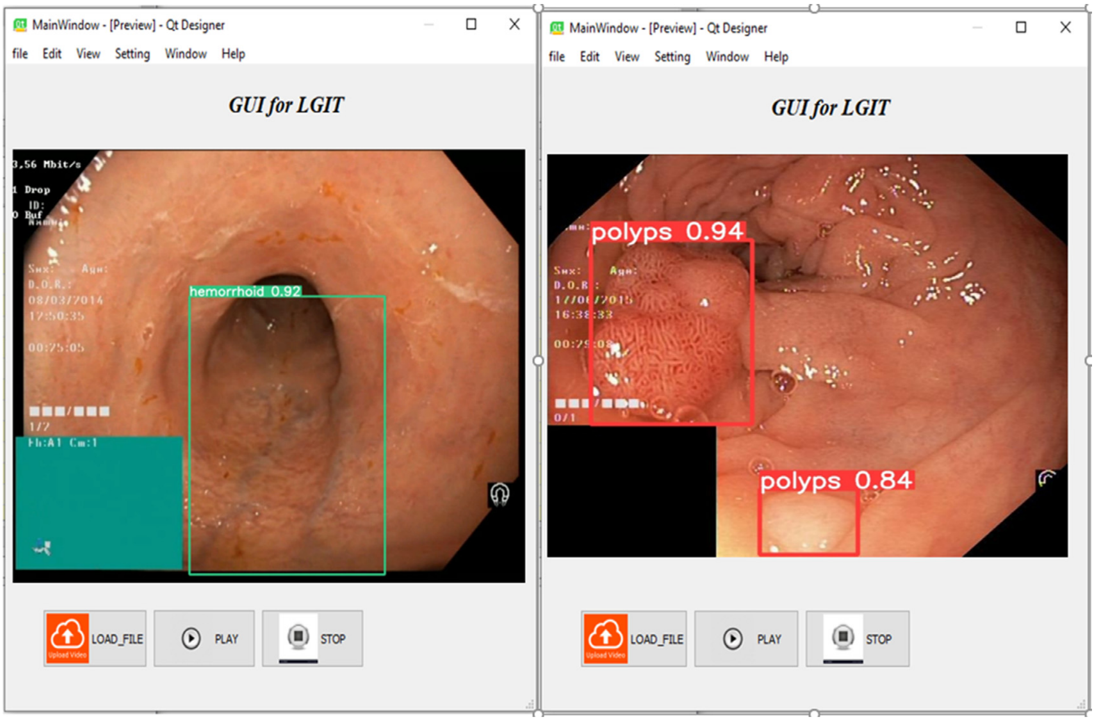

MainWindow - [Preview] - Qt Designer
file Edit View Setting Window Help
GUI for LGIT
hemorrhoid 0.92
LOAD_FILE
Upload Video
PLAY
STOP
MainWindow - [Preview] - Qt Designer
file Edit View Setting Window Help
GUI for LGIT
polyps 0.94
polyps 0.84
LOAD_FILE
Upload Video
PLAY
STOP

Fig. 15. Objects detected on GUI from video.

Table 4
Comparison of developed models based on precision, recall, and mAP.

Models	Precision (%)	Recall (%)	Mean average precision(mAP) (%)
SSD	90.42	91.32	87.5
YOLOv4	93.02	92.83	92.7
YOLOv5	**99.071**	**98.064**	**98.8**

inadequate, 99.5% for normal, 99.3% for impaction, 99.5% for retroflexion, 99.4% for UC, 99.5% for hemorrhoid AP, and 98.8% mAP for all classes at 0.5 IOU threshold.

As shown in Table 3, the lowest error during training and validation was achieved by the YOLOv5 model.

Table 4 below shows the ideal model from developed models by comparing them by precision, recall, and mAP. The model misunderstanding between the classes is very minimal, and the class match with the target class scores shown on the confusion matrix for polyps (90%), UC (97%), hemorrhoids (100%), adequate (99%), inadequate (99%), impacted stool (99%), retroflexed rectum (97%), and cecum (97%) are all higher than 90%. Moreover, all 64 test objects in the videos are successfully detected and localized by the model.

Since there is more GI tract disease, different researchers proposed machine learning and deep learning-based diagnosis systems for specific types of diseases [23,26,27,35–38]. Our study focuses to detects and localizing more prevalent and CRC-causing pathology in the lower GI tract from colonoscopy image and video. Moreover, it also includes the detection of the two most important factors in detecting pathology in a colonoscopy image or video which are the quality of the mucosal view and the identification of anatomical landmarks in the lower GI tract. Previous works are focused only on classification. Even though the evaluation matrix, the dataset, used, and the types of diseases considered were slightly different, the current work achieved significantly improved overall based on the common evaluation matrix compared to [23,26,27,35–38] by the recall and precision evaluation matrix. Furthermore, developed GUI for our best model enables non-expert users to identify lower GI tract disorders. The developed system has the potential to be used as a decision support system for physicians, general practitioners, and patients.

Conclusions

Clinicians may use a mixture of clinical symptoms, laboratory indicators, radiation monitoring, endoscopy, and histological analysis of tissue samples to assess disease occurrence and make treatment decisions in lower GI tract problems. Even though there are more technologies for the diagnosis of the lower GI tract the process is time-consuming, tedious, and complex. More than that those technologies are operator-dependent so the shortage of experience of a gastroenterologist, tiredness of the operator, and lack of concentration during examination leads to missing and misclassifying of the lesions in the tract. The proposed study enables real-time detection of eight colonoscopy findings from video in addition to detection of the same objects from the image. The developed diagnostic system assists gastroenterologists in decision-making and helps general practitioners, essential for time-consuming and easy work. Based on our plan, in this research, we tried to identify the most cancerous lesions in the lower GI tract and prevalent lower GI tract conditions in our country and throughout the world and develop an automated real-time detection of lower GI tract finding diagnosis system with GUI. However, the dataset we collected was small for experimentation. The study was limited to detecting only eight colonoscopy findings from the lower GI tract. Therefore, including more colonoscopy findings and upper GI tract findings is very important for real-time diagnosis of the GI tract. Finally, we recommend developing a combined model and knowledge base system that diagnoses diseases using patient information or other input, in addition to diagnosing only from images and video which can further boost the performance.

Funding Statement

This work was supported by Jimma University, Jimma Institute of Technology, School of Biomedical Engineering.

Ethics Statements

This research did not involve direct human, animal, or another subject contact. According to Jimma University's Institutional Review Board (JUIRB), no formal ethics approval was required in this particular case. Moreover, we have carried out our work according to The Code of Ethics of the World Medical Association (Declaration of Helsinki) and Uniform Requirements for manuscripts submitted to Biomedical Journals. A relevant informed consent was obtained from those subjects both for study participation and publication of identifying information/images in an online open-access publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Salih Aliyi: Conceptualization, Methodology, Software, Data curation, Writing – original draft. **Kokeb Dese:** Data curation, Writing – original draft, Visualization, Investigation, Supervision, Validation, Writing – review & editing. **Hakkins Raj:** Visualization, Investigation, Supervision, Validation, Writing – review & editing.

Acknowledgments

We would like to thank the Jimma Institute of Technology, School of Biomedical Engineering for research support. Moreover, we would like to express our gratitude to my clinical collaborators, Dr. Shamil Nuri (gastroenterologist), Dr. Zarihun Tadele, Sr. Kalkidan Mamushet, Mr. Solomoon Dajane and Mr. Amir Chaka for data collection and annotation, including data collected from the hyper kvasir dataset online.

Data availability

The data that support the findings of this study are available from School of Biomedical Engineering but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of School of Biomedical Engineering.

References

[1] T. Greuter, S. Vavricka, A.O. König, L. Beaugerie, M. Scharl, Malignancies in Inflammatory Bowel Disease, Digestion 101 (Suppl1) (2020) 136–145, doi:10.1159/000509544.

[2] K. ReFaey, et al., Cancer Mortality Rates Increasing vs Cardiovascular Disease Mortality Decreasing in the World: Future Implications, Mayo Clin. Proc. Innov. Qual. Outcomes 5 (3) (2021) 645–653, doi:10.1016/j.mayocpiqo.2021.05.005.

[3] M. AmeliMojarad, Jian Wang, M.A. Mojarad, The function of novel small non-coding RNAs (piRNAs, tRFs) and PIWI protein in colorectal cancer, Cancer Treat. Res. Commun. 31 (2022) 100542, doi:10.1016/j.ctarc.2022.100542.

[4] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA. Cancer J. Clin. 68 (6) (2018) 394–424, doi:10.3322/caac.21492.

[5] G. Obonna, A. Arowolo, A. Agbakwuru, Experience with colonoscopy in the riverine southwestern Nigeria, J. West African Coll. Surg. 2 (2) (2012) 80–90.

[6] H.T. Ayele, B. Redae, H. Tekesilassie, EXPERIANCE OF COLONOSCOPY AT A TERTIARY HOSPITAL, ADISS ABABA, ETHIOPIA, Ethiop. Med. J. 58 (2020).

[7] F.G. Gudissa, B. Alemu, S. Gebremedhin, E.K. Gudina, H. Desalegn, Colonoscopy at a tertiary teaching hospital in Ethiopia: a five-year retrospective review, PAMJ Clin. Med. 5 (2021), doi:10.11604/pamj-cm.2021.5.37.26398.

[8] L. Yang, E. Levi, J.H. Du, H.H. Zhou, R. Miller, A.P.N. Majumdar, Associations between markers of colorectal cancer stem cells, mutation, microRNA and the clinical features of ulcerative colitis, Colorectal Dis 18 (6) (2016) O185–O193, doi:10.1111/codi.13371.

[9] E.F. de Campos Silva, et al., Risk factors for ulcerative colitis-associated colorectal cancer: A retrospective cohort study, Medicine (Baltimore) 99 (32) (2020) e21686, doi:10.1097/MD.0000000000021686.

[10] M. Mascarenhas, et al., Deep learning and colon capsule endoscopy: automatic detection of blood and colonic mucosal lesions using a convolutional neural network, Endosc. Int. open 10 (2) (2022) E171–E177 Feb., doi:10.1055/a-1675-1941.

[11] C. Li, W. Yu, P. Wu, X.D. Chen, Current in vitro digestion systems for understanding food digestion in human upper gastrointestinal tract, Trends Food Sci. Technol. 96 (December 2019) (2020) 114–126, doi:10.1016/j.tifs.2019.12.015.

[12] M.J. Kutyla, et al., Improving the Quality of Bowel Preparation: Rewarding Patients for Success or Intensive Patient Education? Dig. Dis. 39 (2) (2021) 113–118, doi:10.1159/000510461.

[13] A.F. Peery, et al., Risk Factors for Hemorrhoids on Screening Colonoscopy, PLoS One 10 (9) (Sep. 2015) e0139100.

[14] A. Chowdhury, J. Yao, R.L. VanUitert, M.G. Linguraru, R. Summers, Detection of Anatomical Landmarks in Human Colon from Computed Tomographic Colonography Images, 2009, doi:10.1109/ICPR.2008.4760969.

[15] A.S. Lundervold, A. Lundervold, An overview of deep learning in medical imaging focusing on MRI, Z. Med. Phys. 29 (2) (2019) 102–127, doi:10.1016/j.zemedi.2018.11.002.

[16] V. Jayasekeran, B. Holt, M. Bourke, Normal Adult Colonic Anatomy in Colonoscopy, Video J. Encycl. GI Endosc. 1 (Oct. 2013) 390–392, doi:10.1016/S2212-0971(13)70173-0.

[17] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (1) (2019) 44–56, doi:10.1038/s41591-018-0300-7.

[18] O. Catalano, A. Kilcoyne, A. Signore, U. Mahmood, B. Rosen, Lower Gastrointestinal Tract Applications of PET/Computed Tomography and PET/MR Imaging, Radiol. Clin. North Am. 56 (5) (2018) 821–834, doi:10.1016/j.rcl.2018.05.001.

[19] F.-Z. CUI, et al., Synthesis of PEGylated BaGdF5 Nanoparticles as Efficient CT/MRI Dual-modal Contrast Agents for Gastrointestinal Tract Imaging, Chinese J. Anal. Chem. 48 (8) (2020) 1004–1011, doi:10.1016/S1872-2040(20)60039-1.

[20] V.L. Thambawita, et al., The Medico-Task 2018: Disease Detection in the Gastrointestinal Tract Using Global Features and Deep Learning, ArXiv (2018) abs/1810.1.

[21] V. Thambawita, et al., An Extensive Study on Cross-Dataset Bias and Evaluation Metrics Interpretation for Machine Learning Applied to Gastrointestinal Tract Abnormality Classification, ACM Trans. Comput. Healthc. 1 (3) (2020), doi:10.1145/3386295.

[22] S. Petscharnig, K. Schöffmann, and M. Lux, “An Inception-like CNN Architecture for GI Disease and Anatomical Landmark Classification,” 2017.

[23] H. Borgli, et al., HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy, Sci. Data 7 (1) (2020) 1–14, doi:10.1038/s41597-020-00622-y.

[24] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc, 2015, pp. 1–14.

[25] P. Zhang, Y. Zhong, X. Li, SlimYOLOv3: Narrower, Faster and Better for Real-Time UAV Applications, 2019.

[26] K. Pogorelov, et al., Efficient disease detection in gastrointestinal videos – global features versus neural networks, Multimed. Tools Appl. 76 (21) (2017) 22493–22525, doi:10.1007/s11042-017-4989-y.

[27] V. Thambawita, et al., The Medico-Task 2018: Disease detection in the gastrointestinal tract using global features and deep learning, in: CEUR Workshop Proc, 2283, 2018.

[28] K. Pogorelov, et al., Nerthus: A bowel preparation quality video dataset, in: Proc. 8th ACM Multimed. Syst. Conf. MMSys 2017, 2017, pp. 170–174, doi:10.1145/3083187.3083216.

[29] M. Riegler, et al., Multimedia for medicine: The medico task at mediaeval 2017, CEUR Workshop Proc 1984 (2017) 7–9.

[30] S. Hicks, et al., Medico multimedia task at mediaeval 2019, in: CEUR Workshop Proc, 2670, 2019, pp. 7–9.

[31] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” 2020, [Online]. Available: http://arxiv.org/abs/2004.10934

[32] R. Padilla, S.L. Netto, E.A.B. da Silva, A Survey on Performance Metrics for Object-Detection Algorithms, in: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), 2020, pp. 237–242, doi:10.1109/IWSSIP48289.2020.9145130.

[33] L. Alzubaidi, et al., Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, 8, Springer International Publishing, 2021, doi:10.1186/s40537-021-00444-8.

[34] Y. Sun, et al., Correlation between lower gastrointestinal tract symptoms and quality of life in patients with stable chronic obstructive pulmonary disease, J. Tradit. Chinese Med. 33 (5) (2013) 608–614, doi:10.1016/s0254-6272(14)60029-7.

[35] M. A. Riegler, “An Extensive Study on Cross-Dataset Bias and Evaluation Metrics Interpretation for Machine Learning Applied to Gastrointestinal,” vol. 1, no. 3, 2020.

[36] T. Cogan, M. Cogan, L. Tamil, MAPGI: Accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning, Comput. Biol. Med. 111 (2019) 103351, doi:10.1016/j.compbiomed.2019.103351.

[37] S. Petscharnig, K. Schoffmann, M. Lux, An inception-like CNN architecture for GI disease and anatomical landmark classification, CEUR Workshop Proc 1984 (2017) 0–2.

[38] K. Pogorelov, et al., Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection, in: Proc. 8th ACM Multimed. Syst. Conf. MMSys 2017, 2017, pp. 164–169, doi:10.1145/3083187.3083212.