Deep Learning-based Polyp Detection in Wireless Capsule Endoscopy Images

Doaa S. A. Elsersy, Mahmoud M. Selim, Amira S. Ashour
Electronics and Electrical Communication Engineering Department,Faculty of Engineering, Tanta University, Tanta, Egypt

Abstract- Gastrointestinal (GI) system diseases have increased significantly, where colon and rectum cancer is considered the second cause of death in 2020. Wireless Capsule Endoscopy (WCE) is a revolutionary procedure for detecting Colorectal lesions. It was automatically used to detect the polyps, multiple SB lesions, bleeding, and Ulcer. The acquired video by the WCE can be processed using a Computer-Aided Diagnosis (CAD) system. However, such videos suffer several problems, including burling, high illumination. and distortion. These effects obligate the development of image processing techniques of high accuracy in detection using deep learning-based segmentation. In this paper, a transfer learning-based U-Net was proposed to transfer the knowledge between the medical images in the training phase and the subsequent segmentation using transfer learning to achieve better results and high accuracy results compared to other related studies. The improvement is done by using an algorism written in python code The results showed average segmentation accuracy of 98.67%.

Keywords: Deep learning; machine learning; wireless capsule endoscopy; gastrointestinal track; u-net; polyps detection; transfer learning.

I. INTRODUCTION

The rate of colorectal cancer has risen dramatically in recent decades. Hence, the need for early diagnosis has been increased relatively. All abnormalities in the GI, especially the polyps, are abnormal tissue growths, which may appear wherever in the GI tract especially in the colorectal region [1, 2]. This indicates the presence of cancer and the possibility of adenoma existing or even an early cancer stages detection. The early detection of the Colorectal Cancer (CRC) at early stages reduces the mortality rate. Various segmentation approaches have been used starting from boundary detection-based approaches, while other approaches are region-based, such as the active contour. Nonetheless, most datasets of medical images have noise, intensity inhomogeneity, and weak boundaries. This requires hybrid approaches based on the boundary and region of interist (ROI) detection. Nevertheless, the accuracy of these methods are insufficient to match the ground truth of the dataset. Consequently, the diagnosis using endoscopy is essential, where the other traditional methods have disadvantages, such as pain treatment and inaccurate results due to the GI liquid [3].

Typically, the WCE capatures over 50,000 video frames [4], which can be divided as footages ready to be processed using CAD systems [5]. One of the most effective ways to avoid CRC is to remove the precancerous abnormalities. Also, the polyps detection in the colon is vital for diagnosis.

The WCE has been used essentially for patients with contraindications to general anesthesia or limited cooperation. It is a pill containing a charge-coupled devices, two batteries, and an RF (radio frequency) transmitter for the non-invasive detection of gastrointestinal abnormalities [6] without the requirement for hospitalization or anesthesia. It has been a breakthrough technique for diagnosing small intestinal diseases in the recent decade [7]. Many manufacturers, including Given Imaging, Intro Medic, and Olympus [8], have produced capsules enabling a comprehensive inspection of the gastrointestinal system. Different studies were concerned with developing automated systems for polyp detection. Podlasek et al. [9] designed a system for polyp detection with a real-time post-processing pipeline that runs on a variety of devices. The F1 score of CVC-ClinicDB ranged from 0.727 to 0.942. With a 3% FP rate, full examination films sample detected 94% of polyps [9].

Currently, deep learning (DL) techniques were constructed for polyp detection using unmodified colonoscopy footage with 96.7% sensitivity over 34 frame-per-second (fps) [10]. The most accurate systems that used convolutional neural networks had 89% identification accuracy when evaluated across 18,092 video frames [11]. A study using YOLOv2 on more than 8,000 colon polyp images achieved 93.4% accuracy [12 -13]. Other studies have considered DL in WCE for classifying gastrointestinal disorders. Ding et al. [14] used a CNN architecture (ResNet 152) with image resolution 480×480 pixels with achieved 99.88% sensitivity. Also, Tsuboi et al. [15] applied a CNN architecture (SSD) with the image resolution of 300×300 pixels for training on 2237 images from 141 patients and testing 10488 images from 28 patients with achieved 98.8% sensitivity. Leenhardt et al. [16] designed a CNN-based semantic segmentation architecture on 600 images achieved 96% sensitivity. Wang et al. [17] implemented a CNN (RetinaNet) with the image resolution of 480 × 480 pixels with achieved accuracy (AUC) 90% and sensitivity 89.71%.

Moreover, Alaskar et al. [18] applied GoogLeNet; and AlexNet with the image resolution of 224 × 224 pixels; and 227 × 227 pixels, respectively. Majid et al.[19] used the architecture of the CNN with classical features fusion and selection with the image resolution of 224×224 pixels with accuracy 96.5%, and sensitivity 96.5%. Furthermore, Aoki et al. [20] used CNN (ResNet50) with the image resolution of 224×224 pixels with accuracy 99.89%, and sensitivity 96.63%. Saito et al. [21] applied CNN (ResNet50) with the image resolution of 300 × 300 pixels with 84.5% accuracy, and sensitivity 90.7%.

The U-Net is a fully convolutional network architected for segmentation and classification of the medical images. The contribution of the proposed system includes a trained model

using typical U-Net algorism to achieve better results by combining transfer learning models with an adaptable transfer learning. It is used to overcome the limited number of the sampled dataset images. The images were retained for three levels, producing more mean accuracy compared with previous research using the same dataset.

II. METHODOLOGY

Polyp images were used from 25 different video studies with their ground-truth [21]. The database is includes frames from 29 dissimilar sequences of polyp to produce 612 pictures of original resolution of 384 × 288 and a 612 of ground truth, a sample for both is shown in Fig. 1.

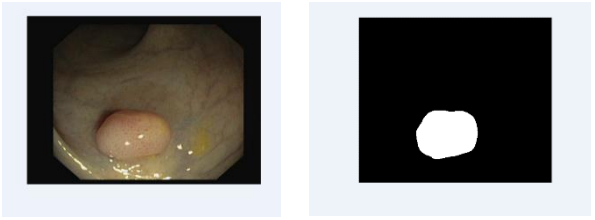

Figure1. Dataset image and its round truth

The images of this dataset required preprocessing to remove or reduce the of the artifacts. The steps are as follows, starting from removing the color mode leaving only the luminance of each pixel after converting them to a grayscale mode, preparing to next preprocessing, where a median filter applied is often used to remove noise. The images were resized to a smaller size of 128×96. After that, the typical U-Net is applied.

A. Traditional U-Net

The U-Net is one of the most important convolutional networks used in segmentation. It is a development of the Fully Convolutional Network (FCN) with a changable design for accurate segmentation. It has a U shape architecture consisting of two parts. The downsampling, which is the contracting path, which consists of many different operations starting with processing the input image. The data is propagated through the network and along all possible paths. Then, at the end, the accurate segmentation map comes out to show as a strip of the blue box corresponding to a multichannel feature map, and each feature is denoted to the top of the blue boxes. A linear activation function- rectified linear unit (ReLU) that returns 0 if it gets any negative input meaning follows the majority of the convolution operations. It returns the value 1 for every positive number x. As a result, it may be written as in equation (1) and (2).

$$f(x)=max(0,x). \quad (1)$$

$$b_{x,y,l}=\mathrm{ReLU}\left(\sum_{\substack{i\in\{-1,0,1\}\\ j\in\{-1,0,1\}\\ k\in\{-1,0,1\}}} \omega_{i,j,k,l}.a_{x+i,y+j,k+cl}\right) \quad (2)$$

The next U-Net operation is max pooling, which decreases the x-y size of the feature map. Each channel is expanded independently using the max-pooling procedure, which propagates the maximum activation from each 2x2 window to the next feature map. The number of the feature channels grew by 2 after each max pooling function. The output of the convolution and max pooling processes is a spatial construction and gradually increases the "what" and at the same time reduces the "where", that's it, the standard classification ends here, and all maps lead to show single output vector. Using a pooling layer with down sampling or pooling of the feature maps is a potted version of the detected features in the input. Due to the convolution pattern, the map is smaller than the input image, so the segmentation uses the input data of the bigger image. The energy function can be defined as [23]:

$$E=\sum w(x)\log(p_{k(x)}(x)) \quad (3)$$

where pk is the pixel-wise as defined in equation (4), as the SoftMax function was used with the final feature map, as [23]:

$$p_k=\exp(a_k(x))/\sum_{k'=1}^{k}\exp(a_{k'}(x)) \quad (4)$$

a_k signifies the activation in channel k. The trained output weights were saved from being applied to the final stage to take advantage of the transfer learning concept.

B. Proposed transfer-based U-Net

Training the images from the dataset using the basic U-Net is challenging, including the few number of trained images. Subsequently, transfer learning is used by retraining for three stages and saving the output weights for each time. The resulting deformed image looks like the masked ground-truth. The second challenge was catching objects of the same class that had to be correctly speared. A background pixels' insertion occurs between all touching objects and assigns an individual loss weight for each pixel. This allows a substantial penalization of network extendedly close these gabs. In the 2D U-Net, the network, the encoder part, and a decoder are mainly filters of variable numbers, where we used 4 in the present work. The dataset was trained by 60% and 17.5% for testing, and 22.5% for validation. The output was trained again using the transfer learning technique by calling the model of EfficientnetB7 and save the weights. Then, this step is retrained for more than two levels by adding the best weights of the EfficientnetB5 each time to get the high achieved accuracy by measuring it with the evaluation matrix. The diagram of the proposed model is illustrated in Fig. 2.

C. Evaluation matrix

The evaluation matrix is used to measure the results of the parameters starts with True Positive (tp), True Negative (tn), False Negatives (fn), and False Positives (fp). The most important metric is accuracy, which is defined as the ratio of total accurately predicted observations to the total number of observations using the following formulas [24]:

$$\text{Accuracy}=\frac{tp+tn}{tp+tn+fp+fn} \quad (5)$$

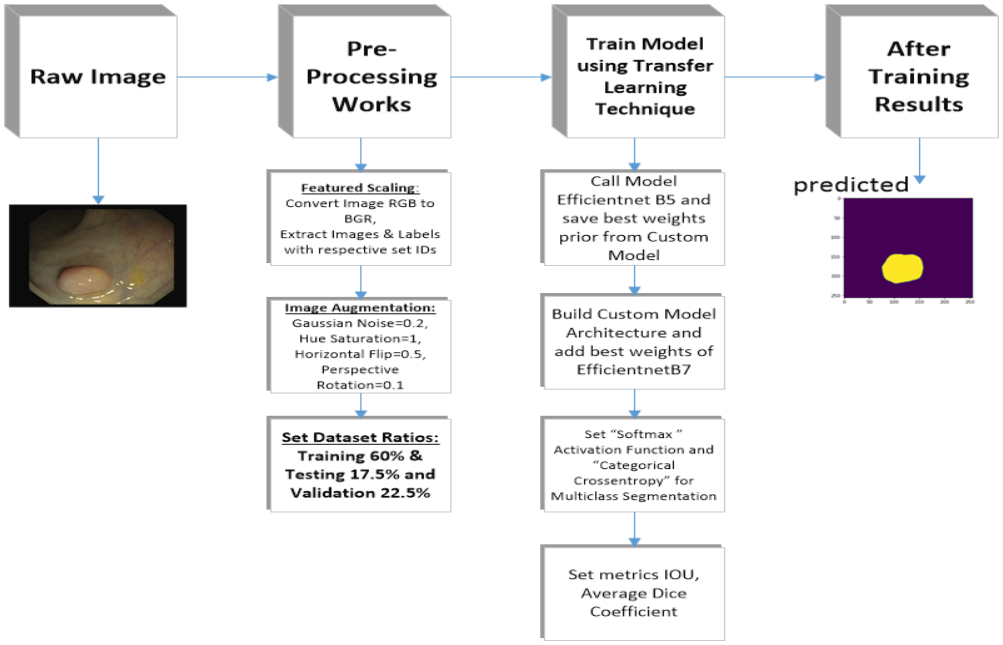

Figure 2. The overall diagram of the proposed method

$$\text{Precision} = \frac{\text{tp}}{\text{tp+fp}} \quad (6)$$

$$\text{Recall} = \frac{\text{tp}}{\text{tp+fn}} \quad (7)$$

$$\text{F1score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \quad (8)$$

The Intersection Over Union (IoU) is also measured for calculating the MAP using the following expression:

$$IoU = \frac{tp}{tp+fp+fn} \quad (9)$$

III. RESULTS

A. Result of Traditional U-Net

The proposed methodology uses an ensemble of traditional U-Net get input images of preprocessed by applying color mode change to grayscale and using a median filter. The results were measured using the evaluation matrix and its parameters as shown in Table 1. The predicted output of this stage was bad, and the loss was high as shown in Fig. 3.

Table 1. Results using typical U-Net Only

Image label	Loss	IoU-score	F1-score
1	0.3581	0.6723	0.7627
2	-0.8750	0.7679	0.8643
3	-0.8975	0.7668	0.8631
4	-0.8903	0.8056	0.8897
5	-0.9117	0.8179	0.8963
6	-0.9211	0.8158	0.8957
7	-0.9081	0.7762	0.8696
8	-0.9289	0.8109	0.8924
9	-0.9352	0.8288	0.9044
10	-0.9235	0.8043	0.8890
Mean		0.87718	0.93243

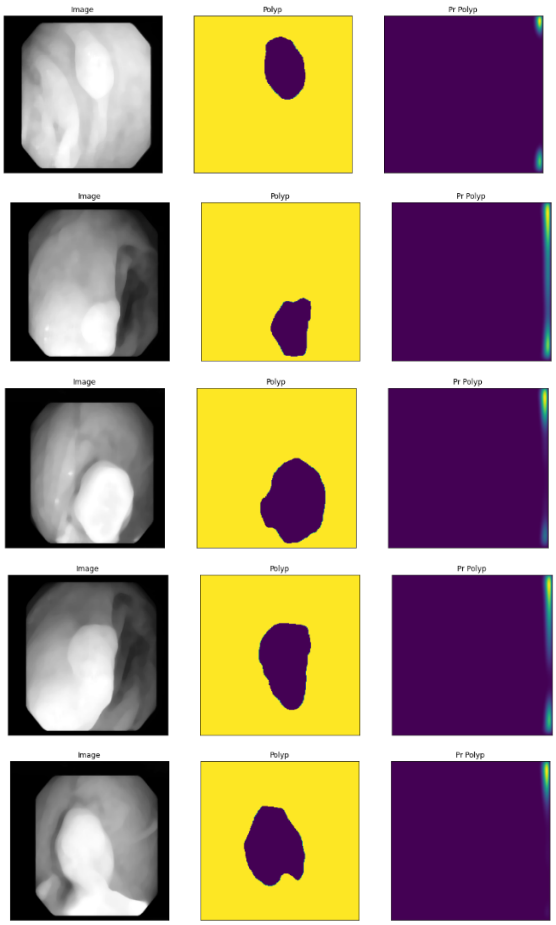

Figure 3. Results images and predicted typical U-Net Only

As shown in Table 1, most values of the loss function are negative. Some modification in position and changes in

orientation needed to be made to score more accurate results.

B. Result of U-Net after augmentation

An augmentation and change in orientation of the images are applied by horizontal flipping and perspective rotation of 0.1 after applying a Gaussian noise=0.2 and hue saturation then using U-Net approach used again the results listed as in Table 3. The input grayscale images and their ground truth and the predicted output.

Table 3. Results using typical U-Net after augmentation

Image label	Loss	IoU-score	F1-score
1	0.3254	0.7945	0.8808
2	0.2684	0.8443	0.9149
3	0.2687	0.8401	0.9125
4	0.2435	0.8498	0.9183
5	0.2228	0.8651	0.9272
6	0.2088	0.8603	0.9244
7	0.1974	0.8734	0.9320
8	0.1949	0.8751	0.9331
9	0.1816	0.8730	0.9319
10	0.1867	0.8753	0.9331
Mean		0.89165	0.93942

The result shows an improvement with 0.014 in IoU-score and 0.007 in F1-score, but this is still not much good. Another additional trials done by retraining the system to overcome the limited number of the used available dataset and a change in trained images percentages comparing with images used for testing and validation.

C. Result of the Proposed U-Net with the Transfer learning

The trained images from U-Net with augmentation was retrained for three more stages as shown in Tables 4-6 after saving the weights to gain the advantages of transfer learning, which was used to overcome the limited number of dataset images. The SoftMax was used as an activation function and the categorical crossentropy for multiclass segmentation to help inaccurate prediction of the polyp. The samples of output predicted images are illustrated in Fig. 4.

After comparing with other models, especially that using the same dataset appear to have a higher F1-score which indicates the accuracy of the predicted results the mean IoU-score reaches 97.4% and the mean F1-score equals 98.7% for all of 612 dataset images.The algorithm built on the concept of transfer learning makes changes on two sides. The first one is to save the weights from the trained images and start it good and near result and use it to start another level of training to tell the results to reach an accurate one; the second is to increase the percentage of testing images and validation to improve the mean results.

Table 4. Results of 1st stage of training

Image	Loss	IoU-score	F1-score
1	0.2318	0.8297	0.9030
2	0.1946	0.8612	0.9242
3	0.1742	0.8782	0.9340
4	0.1537	0.8892	0.9403
5	0.1589	0.8860	0.9385
6	0.1469	0.8973	0.9448
7	0.1412	0.9010	0.9470
8	0.1429	0.8969	0.9447
9	0.1267	0.9083	0.9512
10	0.1180	0.9119	0.9532

Table 5. Results of 2nd stage of training

Image	Loss	IoU-score	F1-score
1	0.0411	0.9660	0.9825
2	0.0415	0.9658	0.9825
3	0.0413	0.9660	0.9826
4	0.0430	0.9655	0.9823
5	0.0407	0.9666	0.9829
6	0.0429	0.9652	0.9821
7	0.0411	0.9668	0.9829
8	0.0419	0.9656	0.9823
9	0.0418	0.9660	0.9826
10	0.0410	0.9662	0.9827

Table 6. Results of 3rd stage of training

Image	Loss	IoU-score	F1-score
1/150	0.0689	0.9477	0.9727
2/150	0.0713	0.9462	0.9719
3	0.0754	0.9424	0.9698
4	0.0645	0.9487	0.9733
5	0.0687	0.9465	0.9720
6	0.0687	0.9465	0.9720
7	0.0731	0.9449	0.9711
8	0.0658	0.9478	0.9728
9	0.0705	0.9442	0.9707
10	0.0658	0.9485	0.9731
Mean		0.97395	0.98674

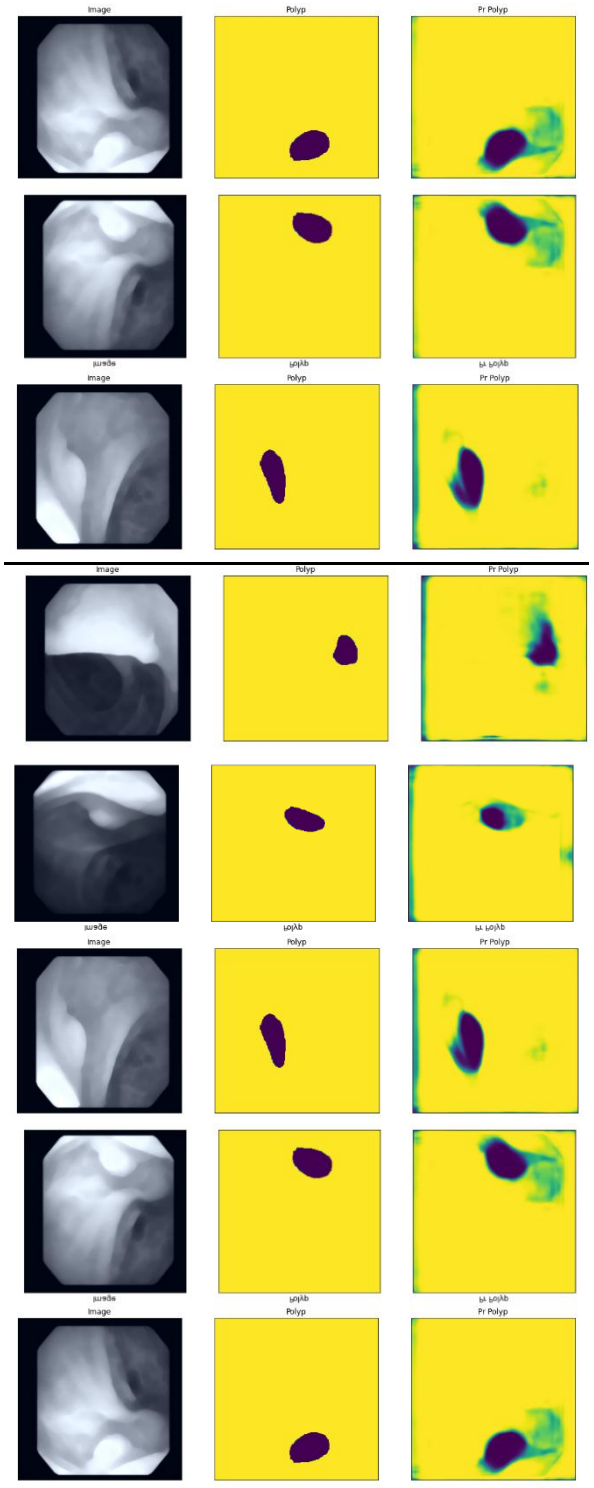

Figure 4. Original ground truth and predicted images after three levels of training

IV. DISCUSSION

The prediction results after applying the proposed methodology provided accurate results by saving the weights of the trained process which was done 3 times to increase the accuracy of results as shown in Table 7. The IoU value for the proposed algorithm shows a good intersection between the predicted image and ground truth which is 97.4.

Table 7. Comparison of methods

Method	Results	Notes
singUnet-VGG	IoU_score: 96.95% F1-score: N/A	The previous related work
Unet without augmentation	mean IoU_score: 0.87718 mean F1-score: 0.93243	Proposed Algorithm
Unet with augmentation	mean IoU_score: 0.89165 mean F1-score: 0.93942	Proposed Algorithm
Unet with transfer lear ning (BACKBONE : EfficientnetB5)	mean IoU_score: 0.97395 mean F1-score: 0.98674	Proposed Algorithm

V. CONCLUSION

For WCE image analysis, the importance of applying deep learning models establishes high accuracy. Nonetheless, the present deep learning models are found to be supervised learning and utilize a limited quantity of marked data for training, resulting in low training data quality. The proposed model resolves that issue, thus provide improvement at the overall classification accuracy by combining the transfer learning advantages, then uses its concept to gain the weights out of the trained. The proposed approach have a short-time for training with 98.7% accuracy compared to the other supervised models. The GPU of google co-lab was used to compensate the fast iteration needed to run the trained models. In the future, transfer learning may be used much broader and with an advanced procedure for improving the detection and prediction of medical images in general.

REFERENCES

[1] D. Jha et al., "ResUNet++: An Advanced Architecture for Medical Image Segmentation," in Proceedings of IEEE International Symposium on Multimedia (ISM), 2019, pp. 225–2255.

[2] R. G. Holzheimer and J. A. Mannick, Surgical treatment: evidence-based and problem-oriented, 2001

[3] J. Asplund, J. H. Kauppila, F. Mattsson, and J. Lagergren, "Survival trends in gastric adenocarcinoma: a population-based study in Sweden," Ann. Surgi. Oncol., vol. 25, no. 9, pp. 693–2702, 2018

[4]Prabhananthakumar Muruganantham, Senthil Murugan Balakrishnan, A survey on deep learning models for wireless capsule endoscopy image analysis, International Journal of Cognitive Computing in Engineering, Volume 2,2021, Pages 83-92.

[5]Atsawarungruangkit A, Elfanagely Y, Asombang AW, Rupawala A, Rich HG. Understanding deep learning in capsule endoscopy: Can artificial intelligence enhance clinical practice? Artif Intell Gastrointest Endosc 2020; 1(2): 33-43.

[6] A. Moglia, A. Menciassi, A. Dario, and A. Cuschieri, "Capsule endoscopy: progress update and challenges ahead," Nature Reviews. Gastroenterology & hepatology, no. 6, pp. 352–362, June 2009. 16

[7] C. Spada, C. Hassan, M. Munoz-Navas, *et al.*, "Second-generation colon capsule endoscopy compared with colonoscopy," Gastrointestinal Endoscopy, vol. 74, no. 3, pp. 581 – 589, 2011.

[8] A. Bergwerk, D. Fleischer, and J. Gerber, "A capsule endoscopy guide for the practising clinician: technology and troubleshooting," Medline, vol. 66, no. 6, pp. 1188–1195, Dec. 2007.

[9] Jeremi Podlasek, Mateusz Heesch, Robert Podlasek Wojciech Kilisiński, Rafał Filip: Real-time deep learning-based colorectal polyp localization on clinical video footage achievable with a wide array of hardware configuration. Endosc Int Open 2021; 09: E741–E748.

[10] Ji Young Lee, Jinhoon Jeong, Eun Mi Song, Chunae Ha, Hyo Jeong Lee, Ja Eun Koo, Dong-HoonYang, Namkug Kim & Jeong-Sik Byeon: Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets. Scientific Reports | (2020) 10:8379.

[11] Bernal, J. et al. Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. IEEE Trans. Med. Imaging 36, 1231–1249 (2017)

[12] Misawa, M. et al. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. Gastroenterology 154(2027- 2029),

e2023 (2018).
[13] Urban, G. et al. Deep learning localizes and identifies polyps in real-time with 96% accuracy in screening colonoscopy. Gastroenterology 155(1069-1078), e1068 (2018).
[14] Ding Z, Shi H, Zhang H, *et al*., Gastroenterologist-Level Identification of Small-Bowel Diseases and Normal Variants by Capsule Endoscopy Using a Deep-Learning Model. Gastroenterology 2019; 157: 1044-1054. e5
[15] Tsuboi A, Oka S, Aoyama K, Saito H, Aoki T, Yamada A, Matsuda T, Fujishiro M, Ishihara S, Nakahori M, Koike K, Tanaka S, Tada T. Artificial intelligence using a convolutional neural network for automatic detection of small-bowel angioectasia in capsule endoscopy images. Dig Endosc 2020; 32: 382-390
[16] Leenhardt R, Vasseur P, Li C, Saurin JC, Rahmi G, Cholet F, Becq A, Marteau P, Histace A, Dray X; CADCAP Database Working Group. A neural network algorithm for detection of GI angiectasia during small bowel capsule endoscopy. Gastrointest Endosc 2019; 89: 189-194
[17] Wang S, Xing Y, Zhang L, Gao H, Zhang H. A systematic evaluation and optimization of automatic detection of ulcers in wireless capsule endoscopy on a large dataset using deep convolutional neural networks. Phys Med Biol 2019; 64: 235014
[18] Alaskar H, Hussain A, Al-Aseem N, Liatsis P, Al-Jumeily D. Application of Convolutional Neural Networks for Automated Ulcer Detection in Wireless Capsule Endoscopy Images. Sensors (Basel) 2019; 19
[19] Majid A, Khan MA, Yasmin M, Rehman A, Yousafzai A, Tariq U. Classification of stomach infections: A paradigm of convolutional neural network along with classical features fusion and selection. Microsc Res Tech 2020; 83: 562-576
[20] Aoki T, Yamada A, Kato Y, Saito H, Tsuboi A, Nakada A, Niikura R, Fujishiro M, Oka S, Ishihara S, Matsuda T, Nakahori M, Tanaka S, Koike K, Tada T. Automatic detection of blood content in capsule endoscopy images based on a deep convolutional neural network. J Gastroenterol Hepatol 2020; 35: 1196- 1200
[21] Saito H, Aoki T, Aoyama K, Kato Y, Tsuboi A, Yamada A, Fujishiro M, Oka S, Ishihara S, Matsuda T, Nakahori M, Tanaka S, Koike K, Tada T. Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network. Gastrointest Endosc 2020; 92: 144-151.
[22] Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., & Vilariño, F. (2015). WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics, 43, 99-111.
[23] N. Siddique, S. Paheding, C. P. Elkin and V. Devabhaktuni, "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications," in IEEE Access, vol. 9, pp. 82031-82057, 2021.
[24] Saloni Kumari, Deepika Kumar, Mamta Mittal, An ensemble approach for classification and prediction of diabetes mellitus using a soft voting classifier, International Journal of Cognitive Computing in Engineering, Volume 2,2021, Pages 40-46.