Deep Learning-based Multiple Bleeding Detection in Wireless Capsule Endoscopy

Prof. Ouiem Bchir, Ghaida Ali Alkhudhair, Lena Saleh Alotaibi,
Noura Abdulhakeem Almhizea, Sara Mohammed Almuhanna, Shouq Fahad Alzeer
Collage of Computer Science and Information, King Saud University, Riyadh, Kingdom of Saudi Arabia

***Abstract*—Wireless Capsule Endoscopy (WCE) is a diagnostic technology for gastrointestinal tract pathology detection. It has emerged as an alternative to conventional endoscopy which could be distressing to the patient. However, the diagnosis process requires to view and analyze hundreds of frames extracted from WCE video. This makes the diagnosis tedious. For this purpose, researches related to the automatic detection of signs of gastrointestinal diseases have been boosted. In this paper, we design a pattern recognition system for detecting Multiple Bleeding Spots (MBS) using WCE video. The proposed system relies on the Deep Learning approach to accurately recognize multiple bleeding spots in the gastrointestinal tract. Specifically, the You Only Look Once (YOLO) Deep Learning models are explored in this paper, namely, YOLOv3, YOLOv4, YOLOv5 and YOLOv7. The results of experiments showed that YOLOv7 is the most appropriate model for designing the proposed MBS detection system. Specifically, the proposed system achieved a mAP of 0.86, and an IoU of 0.8. Moreover, the results of the detection were enhanced by augmenting the training data to reach a mAP of 0.883.**

***Keywords*—Wireless Capsule Endoscopy (WCE); Multiple Bleeding Spots (MBS); Gastrointestinal (GI) disease; deep learning; pattern recognition**

I. Introduction

The digestive system disorders have been a concern for physicians over years. In fact, millions of people around the world suffer from gastrointestinal (GI) diseases. Specifically, among more than 73 thousand participants in a worldwide study, 40% of them have functional gastrointestinal disorders. In addition, disorders such as digestive system cancer are considered fatal and a major cause of mortality according to 2020 United States statistics. Several pathogens can affect the gastrointestinal tract such as inflammations, infections, cancers, benign tumors, ulcers, and hemorrhoids. Some of these pathogens have similar symptoms. Specifically, cancer, benign tumors, ulcers, and hemorrhoids may yield Multiple Bleeding Spots (MBS) in the gastrointestinal tract. The latter symptom consists of a loss of blood in the GI tract because of ruptured vessels indicating the presence of an abnormality [1]. These MBS appear as small dark red spots or as small light spots next to the red dark ones. Fortunately, with the emergence of new diagnostic techniques, it is possible for physicians to detect GI abnormalities. Endoscopy is the most common diagnostic technique for GI tract. Nevertheless, it is inconvenient and painful for the patient. In order to alleviate this inconvenience, Wireless Capsule Endoscopy (WCE) developed in 2000, emerged as a new diagnostic technique. The diagnosing process consists of the patient swallowing a capsule. The latter contains a camera to record the journey of the capsule internally to the GI tract. Then, the physician analyses the record to diagnose the patient by looking for abnormal spots. WCE generates an eight-hour video. In other words, 60,000 frames need to be visualized by the physician. However, due to the small size of the lesion region and the visual fatigue, the disease diagnosis may be missed at an early stage. In light of this, a diagnostic technology related to image processing and pattern recognition would help in the rapid and accurate detection of the disease. Nevertheless, due to the likeness of the MBS and other intestinal characteristics such as, bubbles, holes, or small food debris, etc. It is challenging to extract visual descriptors able to distinguish MBS pattern from the other ones. It is even more arduous due to the background clutter. In fact, MBS can occur in all parts of the GI tract exhibiting large variety of background in terms of color, and texture. One way to tackle this problem is through the use of Deep Learning (DL) models which learn automatically suitable features.

In this paper, we develop a multiple bleeding spot detection system for Wireless Capsule Endoscopy (WCE) videos. More specifically, we design a pattern recognition system based on deep learning models that are able to detect the bleeding spots through the GI tract. In particular, deep learning models adopted for pattern recognition were utilized. These models are designed to localize and categorize the object of interest. For this purpose, we employ the You Only Look Once (YOLO) deep learning approach [2]. In this regard, we propose to compare different versions of YOLO. These are YOLOv3 [3], YOLOv4 [4], YOLOv5 [5], and YOLOv7 [6].

II. Related Works

Recent researches have proposed aided-diagnosis systems for bleeding anomalies within the intestinal tract using WCE images. They can be categorized into classification-based approaches, and detection-based approaches. The former approaches classify the whole WCE frame as including bleeding spots or not including bleeding spots. Whereas, the detection approaches not only classify the frame but also localize the bleeding spots within the frame. Moreover, each of these two categories bifurcates into conventional and deep learning approaches according to the machine learning paradigm that have been adopted. More specifically, conventional approaches use "engineered" features (also referred to as hand crafted features). Alternatively, deep

learning approaches automatically extract the feature while training the deep learning model.

A. WCE Frame Classification System

1) Conventional approaches: The work in [7] propose to classify WCE frames into "Bleeding" and "No Bleeding". For this purpose, it extracts a hand-crafted feature, namely, the color moment feature from WCE frames. Then, it is conveyed to a Support Vector Machine (SVM) [8] classifier. The choice of the visual feature to be adopted has been made through empirical experimentation. In fact, MPEG-7 visual descriptors, "color moment", "Discrete Wavelet Transform", "Edge Histogram Descriptor", "Gabor", and a combination of "Discrete Wavelet Transform" and "color moment". Similarly, the proposed system in [9] extracts hand crafted features. More specifically, MPEG-7 features [10] are considered. These are the "color moments", the "color histogram", the "local color moments", the "Gabor filter", the "Discrete Wavelet Transform" (DWT) and the "Local Binary Pattern" (LBP) features [10]. The extracted features are then conveyed to a machine learning approach to categorize the frames as "Bleeding" or "No Bleeding". This is performed by clustering each of the training "Bleeding" frames, and the training "No Bleeding" frames into similar groups using Fuzzy C-Means (FCM) [11]. As such, in the testing phase, the unknown frame is compared to the obtained cluster centroids from the training phase. It is then assigned to class of the closest centroid.

2) Deep learning approaches: The authors in [12] use a well-known CNN model that won of the ImageNet Large Scale Vision Recognition Competition (ILSVRC). Specifically, it exploits LeNet-5 [13] architecture. Alternatively, the work in [14] uses deep learning CNN models for feature extraction. In particular, VGG-19 [15], ResNet50 [16], and InceptionV3 [17] are adopted. Similarly, these are well known CNN models which won the ILSVRC competition. Nevertheless, inceptionV3 is an evolved version of InceptionV1 used in GoogleNet displays the architecture of inceptionV3. The obtained features from the three considered models are concatenated. Then, a feature selection is performed to select the most distinctive features. The selected features are conveyed to SVM classifier [8] to categorize the frames as "Bleeding" or "No Bleeding". The study in [18] proposed a system to diagnose the abnormalities in the GI. This study proposed a model which utilizes MobileNet [19]. The latter is a lightweight deep learning model. Specifically, it uses the independent convolutions for each depth dimension, then employs 1×1 pointwise convolution to recover the depth. The output of MobileNet [19] is fed to a custom built convolutional neural network model. It is constituted of 64 filters with a kernel size of 3×3. The resulting feature map is passed to a three fully connected layers for classification purpose. In [20] authors proposed to classify WCE frames as "Bleeding" and "No Bleeding". They employ a customized CNN model architecture. It consists of an eight-layer convolutional neural network that is composed of three convolutional layers (C1-C3), three pooling layers (MP1-MP3) and two fully connected layers (FC1, FC2). Moreover, Support Vector Machine (SVM) [8] classifier is utilized instead of the Softmax layer.

B. Bleeding Detection System

1) Conventional approaches: The study in [21] extracted color and texture features. These features are used to generate bag of words using K-means clustering algorithm. Next, the Expectation Maximization (EM) is employed on the "Bag-of-Visual-Words" for super-pixel segmentation. From the region of interest, geometric features like centroid, area, and eccentricity are extracted and fed to the SVM classifier [8]. The authors in [22] proposed an approach based on statistical color feature analysis. First, the frame is split into blocks. After that, dark or light blocks are excluded. Moreover, canny operator [23] is applied to discard the edges. Furthermore, Wavelet db2 with soft thresholding [24] is applied to reduce noise. The Red channel of the RGB color space is exploited to detect bleeding regions. More specifically, red ratio is computed for individual pixels. Finally, Support Vector Machine (SVM) is used to classify WCE frames into bleeding and non-bleeding classes. Alternatively, the system described in [25] performs semantic segmentation by classifying the pixels as a "Bleeding" or "No Bleeding" pixel. This results in detecting the bleeding pixel within the frame. More specifically, the proposed system in [26] extracts the Red-Green-Blue (RGB) color feature [26] and the Gray-Level Co-occurrence Matrix (GLCM) texture feature [26]. These two features are combined and fed to Random Tree (RT) [27], Random Forest (RF) [28], and Logistic Model Tree (LMT) [29] classifiers.

2) Deep learning approaches: The authors in [30] use AlexNet [31] CNN model to classify the frames as "Bleeding", or "No Bleeding". This is a well-known CNN model, which is one of the earliest models that won the ILSVRC run by ImageNet. Once the bleeding frames are separated, they are segmented using SegNet [32] in order to detect the "Bleeding" areas. It is a deep learning model designed for image segmentation. It is constituted of convolutional stacked auto-encoder. Similarly, the authors in [33] use U-Net deep learning segmentation approach to detect "Bleeding" regions in the small intestines. The model architecture has a "U" shape. The model down-samples the input image to a small feature map. Next, it up-samples it. The up-sampling process use skip connections to benefit from the down-sampling process. In fact, at each level, the down-sampled feature map is concatenated to the up-sampled one to generate the next up-sampled feature map. The work in [34] employs a Cascade Proposal network to generate region of interest proposals. These are regions susceptible to include bleeding pattern. The proposed regions are then fed to the Region Proposal Rejection (RPR). The latter is a small network consisting of one convolutional layer, one fully connected layer, and two output layers. It is used to rank the regions based on a score. Its output

is fed to a detection module which predicts the bounding box and the corresponding class. For the testing phase, the unseen image is provided to both a Salient Region Segmentation (SRS) and a Multiregional Region Combination (MRC). While SRS captures the exact location of the regions [34], and MRC that gains adequate coverage of the concerned region and apply the SRS to locate region of interest's positions. Moreover, object boundaries are refined using the Dense Region Fusion (DRF) approach by checking the density of a specific area [34].

C. Discussion

As it can be noticed, the related works in [7], [9], [12], [14], [18], and [20] classify the frames into "Bleeding" and "No Bleeding". While the earliest studies in [7] and [9] are based of extracting "hand crafted" features that are fed to a classifier, the works in [12], [14], [18], and [20] exploit deep learning models befitting therefore from the automatic learning of the features. In fact, using deep learning paradigm alleviates the problem of selecting the suitable features which is usually performed through empirical comparison of the features. Nevertheless, classification approaches do not localize the bleeding within the frame. Alternatively, the works in [21], [22], [25], [30], [33], and [34] perform bleeding detection. In particular, the studies in [21], [22], and [25] utilize "hand crafted" features. While the work in [21] and [22] splits the frame into blocks to transform the problem into a set of local problems and identifies in which block the bleeding occurs, the work in [25] perform semantic segmentation through pixelwise classification. The deep learning detection-based approaches in [30] and [33] are segmentation approaches. In fact, they exploit well known deep learning segmentation approaches SegNet and U-Net. Nevertheless, these two approaches are known to be very slow and not suitable for real world applications [35]. On the other hand, the work in [34] is not employing segmentation. It learns a bounding box to localize the anomaly. Specifically, it is based on a customized CNN. Thus, the adopted model could be fit the considered datasets. Moreover, it includes several modules, namely, SRS, MRC, RPR, and detection modules. This is advantageous when compared to end-to-end model. In fact, the error inducted by one of these modules affects all other modules. Moreover, the error of the different modules gets accumulated.

III. Proposed Approach

Computer aided-diagnosis can lessen the visualization task and help detecting automatically the MBS. As shown in the related works investigation, MBS aided diagnosis systems are based on image processing and machine learning techniques. In particular, most of the reported works related to detecting MBS employ segmentation techniques. As a result, "hand crafted" features for the segmentation task and for the classification task are required. This can be alleviated by the use of deep learning approaches designed for object detection. Nonetheless, to the best of our knowledge, deep learning models have not been explored for MBS detection. In particular, the end-to-end state of the art YOLO models were not investigated.

YOLO deep learning detection model outperformed the other object detection approaches in many pattern recognition applications [36], [37]. Moreover, the success of YOLO model and its applicability to real world applications, yield the evolution of the model and the publication of different versions. However, a throughout comparisons of these versions in terms of performance and efficiency needs to be performed. In this regard, YOLO model, specifically, its latest versions YOLOv3 [3], YOLOv4 [4], YOLOv5 [5], and YOLOv7 [6] are investigated for detecting MBS in the GI tract. In the following, we describe the four considered models.

A. YOLOv3 Architecture

YOLO version 3 (YOLOv3) [3] is an improved version of YOLO which seeks to enhance the performance through the use of residual blocks and different scale feature maps. Inspired by Residual Networks [38] YOLOv3 employs alternatively 3×3 and 1×1 convolutional layers to form a residual unit. This unit aims at avoiding the vanishing gradient problem faced by very deep network. YOLOv3 is composed of five residual block which incorporate a number of residual units. Since a stride of 2 is used at each residual block, the input is down-sampled five times. In particular, the last three down-sampled feature maps are used for the prediction task. Specifically, after the third residual block, the feature map is down-sampled by factor 8. It is exploited for small object prediction. On the other hand, the output of the fourth residual block is down-sampled by a factor of 16, and it is utilized to generate scale 2 feature map. The latter is employed for medium object prediction. Alternatively, big objects, referred to as scale 1 objects, are predicted using the last residual block for which the feature is down-sampled by a factor of 32. Furthermore, YOLOv3 performs feature fusion to benefit from the feature maps at the different scales. As such, it up-samples scale 1 feature map and concatenate it with scale 2 feature map. The obtained feature map is then up-sampled, and concatenate with scale 3 feature map [38].

B. YOLOv4 Architecture

YOLOv4 [4] is the fourth version of the YOLO model family. YOLOv4 model architecture is composed of multiple sections. Namely, they are the Input, the Backbone, the Neck, and the Head (dense prediction, and the sparse prediction). The backbone and the neck sections are responsible for feature extraction and aggregation, respectively. In particular, the CNN deep learning model, CSPDarkNet53 [39], is used as a feature extractor in the backbone section. Alternatively, Spatial Pyramid Pooling (SPP) and Path Aggregation Network (PANet) were utilized in the neck section to fuse the features using Bag of Specials (BoS). Finally, the head which is responsible for both localizing the object in the image and classifying it, amounts to YOLOv3 models. It consists of two stage detectors. The first one is the one stage object detector and the second one is the one is the two-stage object detector [4]. Compared to the previous versions of YOLO, YOLOv4 mainly introduced two additional concepts. Bag of Freebies (BoF) and Bag of Specials (BoS). Bag of freebies are a set of techniques that alters the training framework or perform data augmentation. Many techniques can be incorporated for the purpose of enhancing the model performance without affecting on the inference cost [10]. Alternatively, BoS are strategies such as enlarging the receptive field, integrating features, incorporating attention modules, or post-processing.

These strategies aim at significantly enhancing the performance of accuracy at the expense of increasing the inference cost [4].

C. YOLOv5 Architecture

YOLOv5 [5] is implemented using PyTorch which allows faster training [40]. As such, YOLOv5 allows rapid detection with the same accuracy as YOLOv4. Specifically, YOLOv5 has been proved to have higher performance than YOLOv4 under certain circumstances and partly gained confidence in the computer vision community besides YOLOv4. YOLOv5 model architecture is similar to YOLOv4 architecture. It employs CSPDarknet53 [40] for the backbone section as feature extractor. The latter aims at addressing the gradient in deep networks and decreases the inference time through the use of cross-layer connections between the network's front and back layers. Moreover, it seeks improving the accuracy and utilizing lightweight model. Furthermore, the SPP module referring to the Spatial Pyramid Pooling module, performs maximum pooling with several kernel sizes and then fuses the features by concatenating them together. Additionally, YOLOv5 exploits Path Aggregation Network (PANet) in the neck section as feature aggregator to increase the flow of information and to enhance the object localization. Besides, PANet incorporates a Feature Pyramid Network (FPN) [41]. On the other hand, the head is designed in the same way as YOLOv3 and YOLOv4. Specifically, it produces three different scale feature maps. The CSP network in the backbone is made up from one or more residual units, whereas the CSP network in the neck is made up of new module called CBL modules that replace the residual units. The CBL module consist of Convolution layers, Batch normalization layers, and Leaky ReLU activation function modules [42]. YOLOv5 introduces a new layer referred to as Focus layer [43]. It takes the place of the first three layers of YOLOv3. Therefore, it reduces the GPU requirement and decreases the number of layers.

D. YOLOv7 Architecture

The most recent YOLO architecture, YOLOv7 [44], is based on YOLOv4 version. The main modifications consist of (i) the introduction of the Extended Efficient Layer Aggregation Network (E-ELAN), (ii) the incorporation of model scaling component, (iii) the use of planned re-parameterized convolution, (iv) the employment of auxiliary head, and (v) the exploration of label assigner mechanism. E-ELAN is a computational component in YOLOv7 backbone part. It enhances the prediction performance continuously by employing “expand, shuffle, merge cardinality”. Alternatively, the model scaling optimizes the number of layers, the number of channels, the number of stages in the feature pyramid, and the resolution of the input image in order to meet the requirements of various problems. Nevertheless, YOLOv7 introduces a new model scaling paradigm which optimizes the scaling factors jointly, not independently one from the other. Similarly, YOLOv7 modifies RepConv by discarding the identity connection. In fact, it uses RepConvN in order to prevent the presence of identity connection for re-parametrized convolution. Moreover, YOLOv7 exploits the Deep Supervision training technique. More specifically, YOLOv7 uses an auxiliary head in the intermediate layers to guide the training. The head responsible for the final prediction is referred to as lead head. Additionally, to further enhance the training, YOLOv7 outputs soft labels instead of hard one referring to the ground truth.

We propose to compare the performance between different YOLO approaches which are YOLOv3 [3], YOLOv4 [4], YOLOv5 [5], and YOLOv7 [6] in recognizing in recognizing “Bleeding” spots. For this purpose, the considered models need to be trained. Therefore, each YOLO model is fed with images indicating the bleeding areas, if any. Specifically, the coordinates of the bounding boxes surrounding the MBS patterns are provided as input along with the “Bleeding” images. They consist of the upper left corner coordinates (X, Y), the width, and the height of each box. Concerning the “Non-Bleeding” images, no boundary box is specified. To determine the best version of YOLO, the considered YOLO models are evaluated using the test set. Specifically, the different models are tested in terms of the inference time, MBS localization and classification. The best performing model is adopted to build the required system.

IV. EXPERIMENT

Kvasir-Capsule dataset [45] is considered in this project. It is a dataset of WCE videos collected from clinical examinations performed at the Department of Medicine, Bærum Hospital, and Vestre Viken Hospital Trust in Norway. It consists of 406 “Bleeding” images representing bleeding spots of different size, color, and texture. In addition, it includes 34338 “Non-Bleeding” images representing normal GI tract frames (without bleeding). According to [46], it is not recommended to add images without region of interest (“non-bleeding” images) to the training set. More specifically, “non-bleeding” images should not exceed more than 10% of the total number of images in the training set. As such, only 328 non-bleeding images are first considered. This results in the distribution reported in Table I, where the images are divided into 60% for training, 20% for validation, and 20% testing sets. Nevertheless, in order to get a glimpse of the models’ performance on the real-world, 6000 non-bleeding images are used in the test set. More specifically, both test sets which are the test set after omitting most of non-bleeding images (Test 1) and the test set that containing 6000 background images (Test 2) are assessed.

The available Ground Truth consists of labeling the whole image as including bleeding or not. Nevertheless, in order to train YOLO, a different ground truth should be provided. In fact, the coordinates of the bounding boxes surrounding the bleeding spots should be fed to model to be trained. As such, the dataset is labeled using labeling software tool [47]. As a result, 960 bleeding regions are considered.

TABLE I. DATASET DISTRIBUTION

	Training set	***Validation set***	***Testing set without additional non-bleeding (Test 1)***	***Testing set with additional non-bleeding (Test 2)***
Bleeding	231	75	100	100
Non-bleeding	328	108	86	6000

Two performance measures are considered to evaluate the performance of YOLOv3 [3], YOLOv4 [4], YOLOv5 [5], and YOLOv7 [6] in terms of recognizing MBS. Specifically, we considered Intersection over Union (IoU) [48] and mean Average Precision (mAP) [49], since the localization and the categorization of the object of interest are assessed using these performance measures. Moreover, Floating Point Operations per second (FLOP) [50] is also considered to compare the time efficiency of the considered YOLO models. Fig. 1 shows a comparison between the performances of the considered YOLO models on Test 2 in terms of both mAP and IoU.

As illustrated in Fig. 1, YOLOv3 performs better than YOLOv4 and YOLOv5 in terms of recognition with mAP equal to 0.828. This is an expected outcome since the architecture of YOLOv3 consists of residual blocks. One of them is exploited specifically for small object detections which concord with the small pattern of the bleeding spots. Moreover, in terms of IoU, YOLOv4 achieves an IoU of 0.736 which is better than 0.589 for YOLOv3 and 0.727 for YOLOv5. In fact, YOLOv4 is better in localizing bleeding spots since it incorporates two stage detectors. The first one is called the one stage object detector and the second one is the two-stage object detector. Nevertheless, YOLOv5 exploits path aggregation network that enhances the model localization ability. Alternatively, YOLOv7 achieved the highest IoU and mAP equal to 0.8 and 0.86 respectively. This makes YOLOv7 the most appropriate model to design the proposed approach.

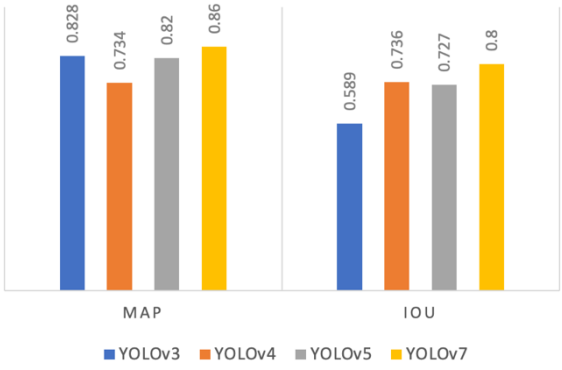

Fig. 1. Performance comparison of YOLOv3, YOLOv4, YOLOv5, and YOLOv7 in terms of mAP and IoU.

Moreover, data augmentation is employed to increase the size of the training data set conveyed to the best performance model, namely YOLOv7 [6] . This is achieved by adding more images to train the model. These images were created by flipping and rotating existing training images. The augmented dataset contains "1056" images. The performance of YOLOv7 without using the augmented data is compared with its performance when training the model with additional data. Table II depicts the performance of YOLOv7 when including and excluding data augmentation. As it can be seen, the augmented dataset improved YOLOv7 performance in terms of mAP.

Furthermore, we compare YOLOv3 [3], YOLOv4 [4], YOLOv5 [5] and YOLOv7 [6] in terms of space complexity. It refers to the space needed to store and train the model. Table III shows the space memory for each model. As depicted, YOLOv5 requires less space memory due to its optimized implementation, while YOLOv4 needs more space memory.

TABLE II. PERFORMANCE COMPARISON OF YOLOV7 [6] WHEN USING DATA AUGMENTATION AND WITHOUT USING IT

	mAP	***IoU***	***FLOPs***
Test results using data augmentation	0.883	0.81	188.9G
Test results without data augmentation	0.86	0.8	188.9G

TABLE III. PERFORMANCE ANALYSIS IN TERMS OF SPACE COMPLEXITY

Model	Space
YOLOv3 Redmon and Farhadi, "YOLOv3."	123.5 MB
YOLOv4 Bochkovskiy, Wang, and Liao, "YOLOv4."	491.6 MB
YOLOv5 "Releases • Ultralytics/Yolov5."	14.4 MB
YOLOv7 Wang, Bochkovskiy, and Liao, "YOLOv7."	142 MB

As illustrated in in Fig. 1, YOLOv7 exceeds the other models in terms of test result, yet there is no significant increase in term of time complexity. It is noticeable that YOLOv4 consumed more time when training the model. On the other hand, when training YOLOv5 it took the least time, and that is predictable since YOLOv5 uses less floating-point operations. Regarding the time considered to train all four models, Table IV reports the training and testing times per image when using Google Collaboratory to train all models.

TABLE IV. PERFORMANCE ANALYSIS IN TERMS OF TRAINING AND TESTING TIME COMPLEXITY

Model	Training Time (s)	Testing Time (ms)
YOLOv3 Redmon and Farhadi, "YOLOv3."	6.5295	0.00026
YOLOv4 Bochkovskiy, Wang, and Liao, "YOLOv4."	23.07	0.00837
YOLOv5 "Releases • Ultralytics/Yolov5."	3.8103	0.00031
YOLOv7 Wang, Bochkovskiy, and Liao, "YOLOv7."	11.0554	0.00124

V. CONCLUSION AND FUTURE WORKS

The arduousness of MBS diagnosis through the burdensome visualization of an eight-hour WCE video of the GI tract has led to the development of aided-diagnosis system. They are based on pattern recognition techniques to detect MBS. In this paper, we proposed to design an aided- diagnosis for MBS detection from WCE video. It is based on deep learning pattern recognition model. In particular, different versions of YOLO model are investigated. Four YOLO models are trained and tested. The comparison and the analysis of the obtained results yielded the selection of the most suitable YOLO model for MBS recognition from WCE videos of the GI tract. Namely, YOLOv7 outperformed the other models.

As future works, the proposed system can be implemented to an applicable and more convenient user-friendly system that

can be used by physicians. Additionally, the performance of the proposed system can be further enhanced by collecting more WCE data to train the model.

REFERENCES

[1] T. Wilkins, B. Wheeler, and M. Carpenter, "Upper Gastrointestinal Bleeding in Adults: Evaluation and Management," Am. Fam. Physician, vol. 101, no. 5, pp. 294–300, Mar. 2020.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[3] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement." arXiv, Apr. 08, 2018. Accessed: Oct. 16, 2022. [Online]. Available: http://arxiv.org/abs/1804.02767

[4] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection." arXiv, Apr. 22, 2020. Accessed: Oct. 16, 2022. [Online]. Available: http://arxiv.org/abs/2004.10934

[5] "Releases • ultralytics/yolov5," GitHub. https://github.com/ultralytics/yolov5/releases (accessed Oct. 17, 2022).

[6] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." arXiv, Jul. 06, 2022. Accessed: Oct. 08, 2022. [Online]. Available: http://arxiv.org/abs/2207.02696

[7] S. Alotaibi, S. Qasim, O. Bchir, and M. M. Ben Ismail, "Empirical Comparison of Visual Descriptors for Multiple Bleeding Spots Recognition in Wireless Capsule Endoscopy Video," in Computer Analysis of Images and Patterns, R. Wilson, E. Hancock, A. Bors, and W. Smith, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 402–407. doi: 10.1007/978-3-642-40246-3_50.

[8] Y. Liu and Y. F. Zheng, "One-against-all multi-class SVM classification using reliability measures," in Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., Jul. 2005, pp. 849–854 vol. 2. doi: 10.1109/IJCNN.2005.1555963.

[9] O. Bchir, M. M. Ben Ismail, and N. AlZahrani, "Multiple bleeding detection in wireless capsule endoscopy," Signal Image Video Process., vol. 13, no. 1, pp. 121–126, Feb. 2019, doi: 10.1007/s11760-018-1336-3.

[10] M. Verma, B. Raman, and S. Murala, "Multi-resolution Local extrema patterns using discrete wavelet transform," in 2014 Seventh International Conference on Contemporary Computing (IC3), Aug. 2014, pp. 577–582. doi: 10.1109/IC3.2014.6897237.

[11] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," Comput. Geosci., vol. 10, no. 2, pp. 191–203, Jan. 1984, doi: 10.1016/0098-3004(84)90020-7.

[12] R. Shahril, A. Saito, A. Shimizu, and S. Baharun, "Bleeding Classification of Enhanced Wireless Capsule Endoscopy Images using Deep Convolutional Neural Network," p. 18.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Ha, "Gradient-Based Learning Applied to Document Recognition," p. 46, 1998.

[14] A. Caroppo, A. Leone, and P. Siciliano, "Deep transfer learning approaches for bleeding detection in endoscopy images," Comput. Med. Imaging Graph., vol. 88, p. 101852, Mar. 2021, doi: 10.1016/j.compmedimag.2020.101852.

[15] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv, Apr. 10, 2015. Accessed: Oct. 17, 2022. [Online]. Available: http://arxiv.org/abs/1409.1556

[16] S. Mukherjee, "The Annotated ResNet-50," Medium, Aug. 18, 2022. https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758 (accessed Oct. 21, 2022).

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision." arXiv, Dec. 11, 2015. doi: 10.48550/arXiv.1512.00567.

[18] F. Rustam et al., "Wireless Capsule Endoscopy Bleeding Images Classification Using CNN Based Model," IEEE Access, vol. PP, pp. 1–1, Feb. 2021, doi: 10.1109/ACCESS.2021.3061592.

[19] A. Pujara, "Image Classification With MobileNet," Analytics Vidhya, Jul. 15, 2020. https://medium.com/analytics-vidhya/image-classification-with-mobilenet-cc6fbb2cd470 (accessed Oct. 21, 2022).

[20] X. Jia and M. Q.-H. Meng, "A deep convolutional neural network for bleeding detection in Wireless Capsule Endoscopy images," in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA: IEEE, Aug. 2016, pp. 639–642. doi: 10.1109/EMBC.2016.7590783.

[21] P. Sivakumar and B. M. Kumar, "A novel method to detect bleeding frame and region in wireless capsule endoscopy video," Clust. Comput., vol. 22, no. S5, pp. 12219–12225, Sep. 2019, doi: 10.1007/s10586-017-1584-y.

[22] S. Suman et al., Detection and Classification of Bleeding Region in WCE Images using Color Feature. 2017. doi: 10.1145/3095713.3095731.

[23] F. Wu, C. Zhu, J. Xu, M. W. Bhatt, and A. Sharma, "Research on image text recognition based on canny edge detection algorithm and k-means algorithm," Int. J. Syst. Assur. Eng. Manag., vol. 13, no. S1, pp. 72–80, Mar. 2022, doi: 10.1007/s13198-021-01262-0.

[24] P. V. V. Kishore, A. S. C. S. Sastry, A. Kartheek, and Sk. H. Mahatha, "Block based thresholding in wavelet domain for denoising ultrasound medical images," in 2015 International Conference on Signal Processing and Communication Engineering Systems, Guntur, India: IEEE, Jan. 2015, pp. 265–269. doi: 10.1109/SPACES.2015.7058262.

[25] K. Pogorelov et al., "Bleeding detection in wireless capsule endoscopy videos — Color versus texture features," J. Appl. Clin. Med. Phys., vol. 20, no. 8, pp. 141–154, 2019, doi: 10.1002/acm2.12662.

[26] C. Sri Kusuma Aditya, M. Hani'ah, R. R. Bintana, and N. Suciati, "Batik classification using neural network with gray level co-occurence matrix and statistical color feature extraction," in 2015 International Conference on Information & Communication Technology and Systems (ICTS), Surabaya: IEEE, Sep. 2015, pp. 163–168. doi: 10.1109/ICTS.2015.7379892.

[27] S. Kalmegh, "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News," vol. 2, no. 2, p. 9.

[28] E. K. Sahin, I. Colkesen, and T. Kavzoglu, "A comparative assessment of canonical correlation forest, random forest, rotation forest and logistic regression methods for landslide susceptibility mapping," Geocarto Int., vol. 35, no. 4, pp. 341–363, Mar. 2020, doi: 10.1080/10106049.2018.1516248.

[29] M. Abedini, B. Ghasemian, A. Shirzadi, and D. T. Bui, "A comparative study of support vector machine and logistic model tree classifiers for shallow landslide susceptibility modeling," Environ. Earth Sci., vol. 78, no. 18, p. 560, Sep. 2019, doi: 10.1007/s12665-019-8562-z.

[30] T. Ghosh and J. Chakareski, "Deep Transfer Learning for Automated Intestinal Bleeding Detection in Capsule Endoscopy Imaging," J. Digit. Imaging, vol. 34, no. 2, pp. 404–417, Apr. 2021, doi: 10.1007/s10278-021-00428-3.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[32] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation." arXiv, Oct. 10, 2016. Accessed: Oct. 21, 2022. [Online]. Available: http://arxiv.org/abs/1511.00561

[33] P. Coelho, A. Pereira, A. Leite, M. Salgado, and A. Cunha, "A Deep Learning Approach for Red Lesions Detection in Video Capsule Endoscopies," in Image Analysis and Recognition, A. Campilho, F. Karray, and B. ter Haar Romeny, Eds., in Lecture Notes in Computer Science, vol. 10882. Cham: Springer International Publishing, 2018, pp. 553–561. doi: 10.1007/978-3-319-93000-8_63.

[34] L. Lan, C. Ye, C. Wang, and S. Zhou, "Deep Convolutional Neural Networks for WCE Abnormality Detection: CNN Architecture, Region Proposal and Transfer Learning," IEEE Access, vol. 7, pp. 30017–30032, 2019, doi: 10.1109/ACCESS.2019.2901568.

[35] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari, "Recurrent residual U-Net for medical image segmentation," J. Med. Imaging, vol. 6, no. 01, p. 1, Mar. 2019, doi: 10.1117/1.JMI.6.1.014006.

[36] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: challenges, architectural successors, datasets and applications," Multimed. Tools Appl., Aug. 2022, doi: 10.1007/s11042-022-13644-y.

[37] "(9) (PDF) KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection." https://www.researchgate.net/publication/316215961_KVASIR_A_Multi-Class_Image_Dataset_for_Computer_Aided_Gastrointestinal_Disease_Detection (accessed Oct. 31, 2022).

[38] Ju, Luo, Wang, Hui, and Chang, "The Application of Improved YOLO V3 in Multi-Scale Target Detection," Appl. Sci., vol. 9, no. 18, p. 3775, Sep. 2019, doi: 10.3390/app9183775.

[39] N. Kwak and D. Kim, "Object detection technology trend and development direction using deep learning," Int. J. Adv. Cult. Technol., vol. 8, no. 4, pp. 119–128, Dec. 2020, doi: 10.17703/IJACT.2020.8.4.119.

[40] M. Sozzi, S. Cantalamessa, A. Cogato, A. Kayad, and F. Marinello, "Automatic Bunch Detection in White Grape Varieties Using YOLOv3, YOLOv4, and YOLOv5 Deep Learning Algorithms," Agronomy, vol. 12, no. 2, Art. no. 2, Feb. 2022, doi: 10.3390/agronomy12020319.

[41] T.-K. Nguyen, L. Vu, V. Vu, T.-D. Hoang, S.-H. Liang, and M.-Q. Tran, "Analysis of Object Detection Models on Duckietown Robot Based on YOLOv5 Architectures," vol. 4, pp. 17–12, Mar. 2022.

[42] X. Xu, X. Zhang, and T. Zhang, "Lite-YOLOv5: A Lightweight Deep Learning Detector for On-Board Ship Detection in Large-Scene Sentinel-1 SAR Images," Remote Sens., vol. 14, no. 4, p. 1018, Feb. 2022, doi: 10.3390/rs14041018.

[43] K. Patel, C. Bhatt, and P. L. Mazzeo, "Deep Learning-Based Automatic Detection of Ships: An Experimental Study Using Satellite Images," J. Imaging, vol. 8, no. 7, p. 182, Jun. 2022, doi: 10.3390/jimaging8070182.

[44] G. Boesch, "YOLOv7: The Most Powerful Object Detection Algorithm (2022 Guide)," viso.ai, Aug. 11, 2022. https://viso.ai/deep-learning/yolov7-guide/ (accessed Oct. 08, 2022).

[45] P. H. Smedsrud et al., "Kvasir-Capsule, a video capsule endoscopy dataset," Sci. Data, vol. 8, no. 1, Art. no. 1, May 2021, doi: 10.1038/s41597-021-00920-z.

[46] "how to use Background images in training? • Issue #2844 • ultralytics/yolov5." https://github.com/ultralytics/yolov5/issues/2844 (accessed Feb. 07, 2023).

[47] "heartexlabs/labelImg." Heartex, Oct. 30, 2022. Accessed: Oct. 30, 2022. [Online]. Available: https://github.com/heartexlabs/labelImg

[48] Naoki, "Object Detection: Intersection over Union (IoU)," Medium, Oct. 08, 2022. https://naokishibuya.medium.com/object-detection-intersection-over-union-iou-f7b91555eb5f (accessed Oct. 26, 2022).

[49] B. Wang, "A Parallel Implementation of Computing Mean Average Precision." arXiv, Jun. 19, 2022. Accessed: Oct. 22, 2022. [Online]. Available: http://arxiv.org/abs/2206.09504

[50] "Floating-Point Operation - an overview | ScienceDirect Topics." https://www.sciencedirect.com/topics/computer-science/floating-point-operation (accessed Oct. 30, 2022).