Pattern Recognition Letters 131 (2020) 193–204

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Gastrointestinal diseases segmentation and classification based on duo-deep architectures[☆]

Mehshan Ahmed Khan[a], Muhammad Attique Khan[b], Fawad Ahmed[a], Mamta Mittal[c,*], Lalit Mohan Goyal[d], D. Jude Hemanth[e], Suresh Chandra Satapathy[f]

[a] Department of EE, HITEC University, Museum Road, Taxila, Pakistan
[b] Department of CS, HITEC University, Museum Road, Taxila, Pakistan
[c] Department of Computer Science and Engineering, G. B. Pant Govt. Engineering College, Okhla, New Delhi, 110020, India
[d] Department of CE, J. C. Bose University of Science & Technology, YMCA, Faridabad, India
[e] Department of ECE, Karunya University, Coimbatore, India
[f] School of Computer Engineering, Kalinga Institute of Industrial Technology (Deemed to be University), Bhubaneswar, Odisha, 751024, India

ARTICLE INFO

Article history:
Received 20 September 2019
Revised 13 December 2019
Accepted 31 December 2019
Available online 31 December 2019

Keywords:
Stomach diseases
Mask RCNN
CNN features
Feature selection

ABSTRACT

Nowadays, almost one million gastrointestinal patients are successfully treated by Wireless Capsule Endoscopy (WCE). It is the latest technology in the area of medical imaging for the diagnosis of gastrointestinal diseases such as ulcer, polyp, bleeding, etc. Manual diagnosis process is time-consuming and hard for doctors; therefore, researchers have proposed computerized techniques for detection and classification of these diseases. In this article, a deep learning-based method is presented for ulcer detection and gastrointestinal diseases (ulcer, polyp, bleeding) classification. Modified mask Recurrent Convolutional Neural Network (RCNN) based ulcer segmentation is proposed. The ulcer annotated images are utilized to train the Mask RCNN model to obtain output in the form of bounding box ulcer detected area and mask segmented region. In the classification phase, the ResNet101 pre-trained CNN model is fine-tuned through transfer learning to derive deep features. The acquired deep features are optimized through grasshopper optimization along with minimum distance fitness function. The best-selected features are finally supplied to a Multi-class Support Vector Machine (MSVM) of cubic kernel function for final classification. Experiments have been performed in two-steps; first, the ulcer segmentation results are computed through recall, precision, and Mean Overlap Coefficient (MOC). The ResNet50+FPN as backbone and training all the layers of Mask-RCNN gives best results in terms of MOC = 0.8807 and average precision = 1.0. Second, the best classification accuracy of 99.13% is achieved on the cubic SVM for $K = 10$. It is clearly perceived that the proposed method outperforms when compared and analyzed with the existing methods.

1. Introduction and background

During the last two decades, developments in the field of medical imaging has shown a lot of improvement for automatic diagnosis of human diseases in different organs of a human body, like stomach, brain [1,2], skin [3], etc. [4,5]. Among these, stomach disease is a most common type [6]. Gastrointestinal Stomach Infections (GSI) consists of ulcer, bleeding, and polyp. Since 2017, around 135, 430 GSI cases were diagnosed in the USA. Another survey conducted in 2017 shows that 765,000 deaths occurred due to stomach infections. Moreover, 21% of humans are diagnosed with GSI in the USA since 2017 [7]. Since 2019, an expected 27,510 new cases of GSI are diagnosed in the USA that involve 17,230 males and 10,230 females whereas, the estimated deaths are 11,140 including 6800 males and 4340 females [8].

The use of push gastroscopy tools for detection and analysis of gastrointestinal infections like polyps, ulcers and bleeding is not suitable for small bowls due to its complex structure [9]. This problem was solved in the year 2000 by introducing a new technology named as WCE [10]. According to an annual report of 2018, approximately one million patients are successfully treated using WCE [11,12]. In WCE, a physician examines the interior of Gastrointestinal Tract (GIT) for detecting the disease. During this process, a capsule containing a wireless camera, light emitting diodes, radio frequency emitter and a battery are swallowed by

[☆] Handled by Associate Editor: Sanniti di Baja.

[*] Corresponding author.

E-mail addresses: attique.khan440@gmail.com (M.A. Khan), mittalmamta79@gmail.com (M. Mittal).

https://doi.org/10.1016/j.patrec.2019.12.024
0167-8655/© 2019 Elsevier B.V. All rights reserved.

the patient. The system automatically moves in the GI tract and after transmitting real-time video, the capsule discharges through the anus. The physician examines the received video frames and decides about the disease [13].

WCE is mainly used for diagnosing serious diseases like ulcers malignancy, bleeding, and polyps in the digestive system [14].This technique has a wide visual range and is more convenient to use. Therefore, it helps to eliminate most of the discomfort and complications that a patient suffers during conventional endoscopy methods like enteroscopy and computed tomography enteroclysis. Diagnostic accuracy is improved for locating tumors and gastrointestinal bleeding especially in the small intestine [15].

On an average, the whole procedure takes more than two hours. A camera captures video frames of size 256 × 256 pixels with a speed of 2 frames per second. These frames are further compressed using JPEG [16]. Manual checking of diseases through the acquired video frames is very difficult as there are around 60, 000 images per person. Even for an experienced physician, it would take a considerable amount of time to review the whole data because the infected area may appear in only one or two frames. Though most of the frames contain futile information, however the physician needs to sequentially go through the entire video. Sometimes it causes misdiagnosis due to inexperience or negligence [17].

To solve this problem, researchers have been working to employ computerized methods. These methods follow few key steps like segmentation, feature visualization, and classification. Feature visualization is a main step in all computerized methods. Various features are computed in the literature like color-based, texture-based, and few others [18,19]. But, all computed features are not relevant which must be removed for high classification rate [20]. Therefore, the feature selection techniques are required [21]. Recently, deep neural networks based method was used for automated diagnosis of medical diseases [22,23]. A high-performance Convolutional Neural Network (CNN) is generally used to extracts automatic features from raw images and achieves best accuracy as compared to traditional methods. The architecture of a CNN includes several numbers of layers such as convolution, ReLu, pooling, normalization, fully connected, and an output layer. Meryem et al. [24] presented a multi-scale method for ulcer detection. They extracted texture patterns such as complete LBP and Laplacian pyramids and employed SVM at the end for classification. They tested their presented system on two WCE datasets and achieved an accuracy of 95.11% and 93.88%, respectively. Bchir et al. [25] introduced computerized approach for bleeding classification using supervised and unsupervised learning methods. An unsupervised method is used to reduce the computation cost while using supervised learning techniques, a model is built which is able to identify positive and negative multiple bleeding spots. Haya et al. [26] presented a CNN based ulcer detection method to handle the problems of low contrast and irregular lesions form endoscopy. The authors used AlexNet and GoogleNet pre-trained models to extract features. The extracted features were classified and validated through several parameters to achieve an improved accuracy. In [27], the authors presented a CNN based hookworm's detection. In the proposed architecture, two CNN models are used; first for edge detection and the second for the classification of hookworms. These CNNs are based on holistically-nested edge detection (HED) network and Inception, both of these networks were embedded using edge pooling in deep and shallow parts of the HED. Sharif et al. [28] introduced a fusion of deep CNN and geometric features-based method for classification of GI tract infections. The infected regions are initially extracted through color based approach and later VGG16 and VGG19 pre-trained CNN networks are utilized for feature extraction. Few geometric features are also fused along with CNN features and the best ones are selected for final classification through K-Nearest Neighbor (KNN). The presented method showed significant performance on the selected dataset.

1.1. Key challenges

The existing methods follow the well-known steps from prepre-processing to classification. But several challenges are exists in the following listed steps which cause in system efficiency degradation. In the pre-processing step, low contrast video frames make the segmentation process more challenging. Furthermore, few other challenges of this step are illumination and brightness of infected regions. Another major challenge is the accurate lesion segmentation because the well-known methods such as thresholding and clustering methods do not give a significant solution for GIT abnormalities detection. Third, the extraction of useful features is always a big challenge for accurate recognition of GIT abnormalities such as ulcers bleeding, and polyps. The number of redundant information in these features is also a major impediment to obtain good accuracy.

In this article, a new Computer Aided Diagnostic (CAD) system is proposed for GI diseases segmentation and classification using a duo-deep architecture. The Mask-RCNN is implemented for ulcer segmentation by employing backbone ResNet50+FPN on all layers. Later, best deep features are selected through grasshopper optimization along with minimum distance fitness function for final classification of GI tract diseases.

1.2. Manuscript organization

The rest of the manuscript is ordered as follows: The proposed methodology which includes Modified Mask RCNN, CNN features based classification steps are presented in Section 2. Detailed segmentation of infected regions and classification results are analyzed in Section 3. At the last, the conclusion of this work is described in Section 4.

2. Proposed methodology

The proposed framework of ulcer segmentation and gastrointestinal stomach infections classification system contains two steps; ulcer segmentation using Mask-RCNN and classification of gastrointestinal infections by using deep CNN feature optimization. The proposed methodology is shown in Fig. 1. In this Figure, segmentation of ulcer regions are done through modified Mask-RCNN whereas classification is achieved through the proposed deep features selection process. Experiments are performed on Private Dataset and CV Clinic DB which include Ulcer, Bleeding, Polyps, and Healthy images. The detail of each phase is given in the following sections.

2.1. Modified Mask-RCNN based ulcer segmentation

In this work, we have utilized Mask-RCNN [29,30] for ulcer segmentation using RGB images. Mask-RCNN is quicker than Faster RCNN for semantic segmentation tasks. Mask-RCNN works in different number of components such as backbone, region proposal network (RPN), ROI aligns, network head, and loss function, as shown in Fig. 2. Brief description of each component is given below:

Backbone - Standard CNN is used as a backbone in the Mask-RCNN to extract features from the image. In this work, we have utilized ResNet50 along with Feature Pyramid Network (FPN) [31] as a feature extractor. The low-level features are computed from early layers and later layers are utilized for high-level features. For this purpose, original images are converted into a dimension of 1024 × 1024 × 3 for a features map of size 32 × 32 × 2048.

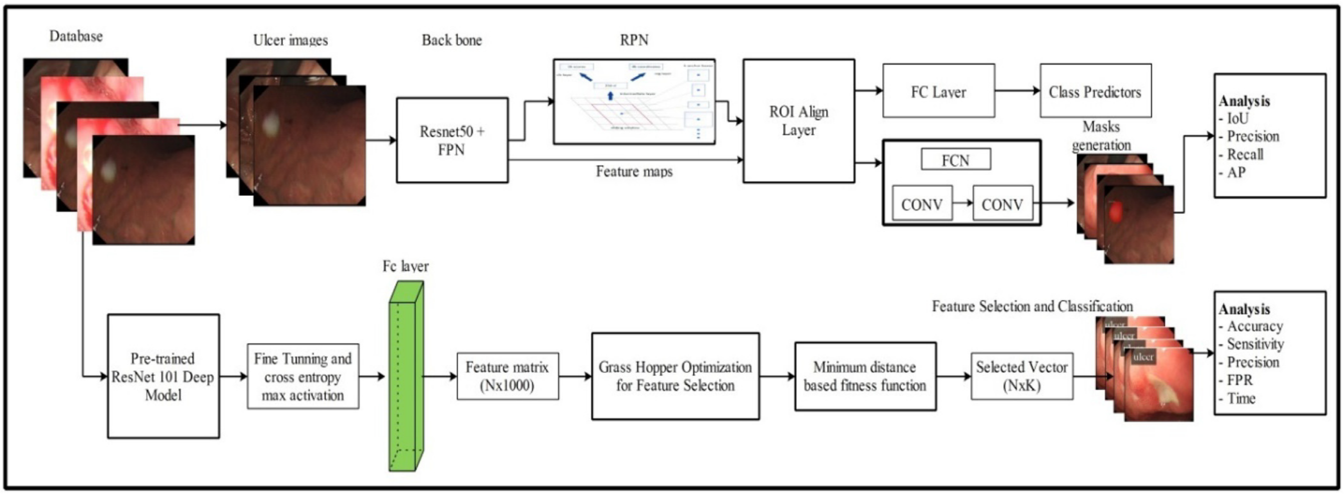


Fig. 1. Proposed design of ulcer segmentation and gastrointestinal stomach infections classification using CNN.

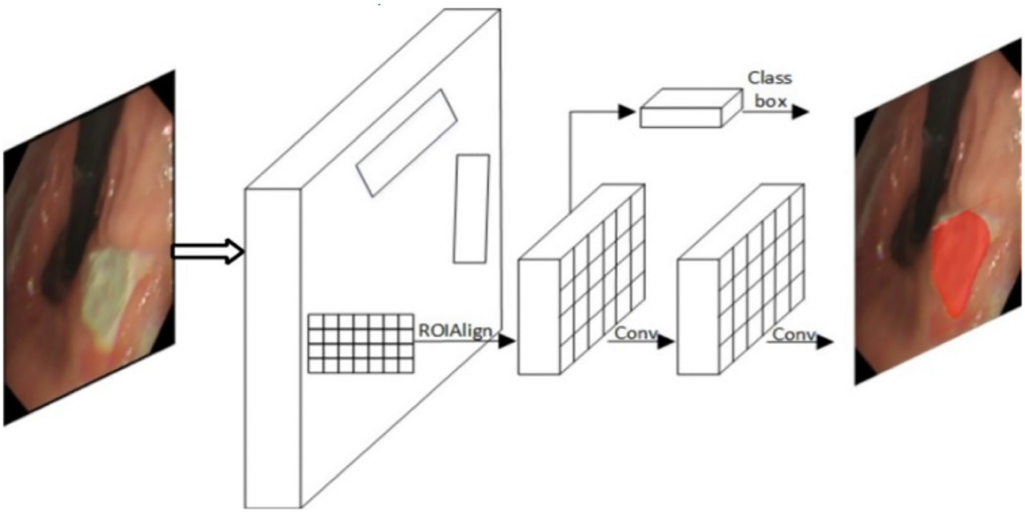


Fig. 2. General architecture of Mask-RCNN.

Region Proposal Network (RPN) - RPN is a lightweight CNN which was firstly introduced in the Faster RCNN. This network aims to replace the selective search [32], which is a slower mechanism for generating bounding boxes. The input of this network are the image features from the backbone to produce bounding boxes of the object. RPN scans regions of the image, called anchors which are 200 K in this work. Through these anchors, boxes spread across the entire image. To cover as many areas as possible, overlap anchors of different sizes and aspect ratios are produced.

The top N anchors based on the highest probability of RPN predictions are chosen. Since the proposals overlap with each other and reduce the redundancy, Non-maximum suppression (NMS) based on class scores are used. The loss function of RPN is defined as follows:

$$\lambda(b_i,\ B_i) = \frac{1}{N_c}\sum_i L_c\left(b_i, b_i{}^*\right) + \psi\frac{1}{N_{reg}}\sum b_i^* L_{reg}(B_i, B_i{}^*) \qquad (1)$$

Where, b_i is the predicted probability of an anchor point i, b_i^* denote the ground truth label which is 1 or 0. If the anchor point is positive then it is 1 otherwise 0. B_i is a vector which has four-point coordinates for drawing bounding box of the segmented region and $B_i{}^*$ denotes the ground truth bounding box.

ROI Align - To predict the precise pixel masks of the object, well-aligned Region of Interest (ROI) features are required. For this purpose, the ROI pooling layer in the Faster-RCNN is replaced with the ROI Align layer. Anchors produced by the RPN layer have different size and ratio. In ROI pooling, the size of these proposals are normalized. To compute the exact pixel of features for the ROI Align layer, bilinear interpolation (BI) [33] is employed. The average and max pooling operations are performed in four regularly sampled locations using current features for further refinement.

Network Head - Features extracted by the ROI Align layer are passed to network head to predict class labels, bounding boxes and mask generation. Fully connected (FC) layer predicts the class labels and bounding boxes based on the ROI Align features. The FCN predict a $m \times m$ mask for each ROI. The loss function utilized is defined as:

$$\lambda = \lambda_{class} + \lambda_{box} + \lambda_{mask} \qquad (2)$$

Where, λ_{class} is the loss of classes, λ_{box} is the loss of bounding boxes and λ_{mask} is the loss of masses generated by the Mask-RCNN. The overall loss, λ is sum of all these losses.

Parameters - Several parameters are utilized during the implementation of Mask RCNN such as validation steps, layer size,

Table 1
Configuration parameters of Mask-RCNN.

Per epoch training steps	500
Validation steps	100
Mini batch Size	128
Backbone	ResNet50
Backbone stride	[4, 8, 16, 32, 64]
FC layer size	1024
RPN anchor scale	[32, 64, 128, 56, 512]
RPN stride	1
Threshold	0.6
Per Image ROI training	200
Positive ratio for mask head	0.40
ROI pool size/max size	7/14
Mask shape	[28,28]
Learning rate	0.01
Momentum	0.6

Table 2
Precision and Recall by using backbone ResNet50+FPN(Training network heads only).

$I_{\circ}U$ Threshold Value	Precision	Recall
0.5	0.6666	0.6666
0.75	0.6445	0.6556
0.9	0.1556	0.4111

Table 3
Precision and Recall by using backbone ResNet50+FPN (Training all layers).

$I_{\circ}U$ Threshold Value	Precision	Recall
0.5	0.6666	0.6666
0.75	0. 6666	0. 6666
0.9	0.2000	0.4333

Table 4
Precision and Recall by using backbone ResNet101+FPN (Training network heads only).

$I_{\circ}U$ Threshold Value	Precision	Recall
0.5	0.6666	0.6666
0.75	0.6112	0.6388
0.9	0.2334	0.4500

threshold, momentum, as listed in Table 1. In Fig. 3, output of the proposed technique is illustrated by training the modified Mask-RCNN model using the parameters mentioned in Table 1. The trained model is able to correctly segment the detected ulcer area by drawing bounding boxes at the appropriate spatial locations along with the predicted labels and further masking the detected ulcer area.

2.2. Gastrointestinal diseases classification

In the area of medical imaging, classification of multiple diseases using computerized methods are helpful for doctors. The increase in patient's data has made computerized methods more complex. More recently, deep convolutional neural network (DCNN) has shown huge performance improvement for large datasets with minimum time as compared to classical methods. DCNN includes low, middle, and high-level features which are integrated in a single matrix for better performance.

In this work, we have used the ResNet101 pre-trained CNN model [34] for features extraction. This model comprises a total of 344 layers; 104 convolution, 100 ReLu, 104 batch normalization, one average pool, one FC, 100 addition layers, and a softmax classifier.

Transfer learning is performed for features mapping on the endoscopy data using the previous trained ImageNet dataset. These features are mapped through an activation function. In this work, cross entropy activation function is used. Mathematically, it is defined as:

$$H_c(f_i, \mathbb{C}) = -\sum_{m}^{M} P_{(V,\mathbb{C})} \log\left(P_{(V,\mathbb{C})}\right) \tag{3}$$

Where, M denotes the total number of classification classes, $\mathbb{C}$ is the class labels, and P is probability for V observations over class m. The cross-entropy activation function is applied on both training and testing data. The selected training/testing ratio is 70, 30 and two feature vectors are return as an output. Later both the vectors are optimized through feature selection algorithm which is discussed in the next section. An optimized training vector is used to train the Support Vector Machine (SVM) model for final classification.

2.2.1. ResNet architecture and implementation for WCE

The ResNet101 architecture accept input of dimension 224×224. Brief description of each convolution layer is provided below in the form of output, pooling size and stride. The mini batch size is selected as 128. The learning rate and momentum is set as 0.1 and 0.6, respectively.

Name	Output size	Layers structure
Convolution1	112×112	7×7, 64, stride=2
Convolution 2-x	56×56	Max pooling filter= 3×3 $1 \times 1,\ 64$ Stride=2 $[\ 3 \times 3,\ 64\ ] \times 3$ $1 \times 1,\ 256$
Convolution 3-x	28×28	$1 \times 1,\ 128$ $[3 \times 3,\ 128] \times 4$ $1 \times 1,\ 512$
Convolution 4-x	14×14	$1 \times 1,\ 256$ $[\ 3 \times 3,\ 256\ ] \times 23$ $1 \times 1,\ 1024$
Convolution 5-x	7×7	$1 \times 1,\ 512$ $[\ 3 \times 3,\ 512\ ] \times 3$ $1 \times 1,\ 2048$
	1×1	Avg Pool=2048 Fc layer= 1000-d Classifier=Softmax
Number of Floating Points Operations=7.6×10^9		

2.2.2. Features selection

The extraction of useful learning from the extracted set of features is an active research area nowadays. The originally extracted features may include the number of irrelevant and redundant features which needs to be removed before final learning. The removal of irrelevant information is important to maintain consistency of the proposed model. In the literature, several metaheuristic methods are proposed for features selection, for example, GA, PSO, Firefly, etc. [24]. These methods are used for different applications such as medical, agriculture, etc. In this work, we implement an Improved Grasshopper Optimization Algorithm (IGOA) along with minimum distance-based fitness function for feature selection. Through this method, we select the best features from a fused feature vector. Initially, features are utilized as a GrassHopper (GH) and based on a fitness function, a target GH is selected for next iterations. This process is continuous (or continued?) for all features and at the end; the best vector is obtained as an output for final classification.

Let P be the number of features, where each feature represents a GH. Based on minimum Euclidean distance fitness function, each GH is evaluated and the best one is known selected) as the target. The target GH starts interaction to other GHs' who move towards the target GH. Mathematically, position of P_{th} GH is defined as:

$$\Delta_p = S_p + G_p + A_p \tag{4}$$

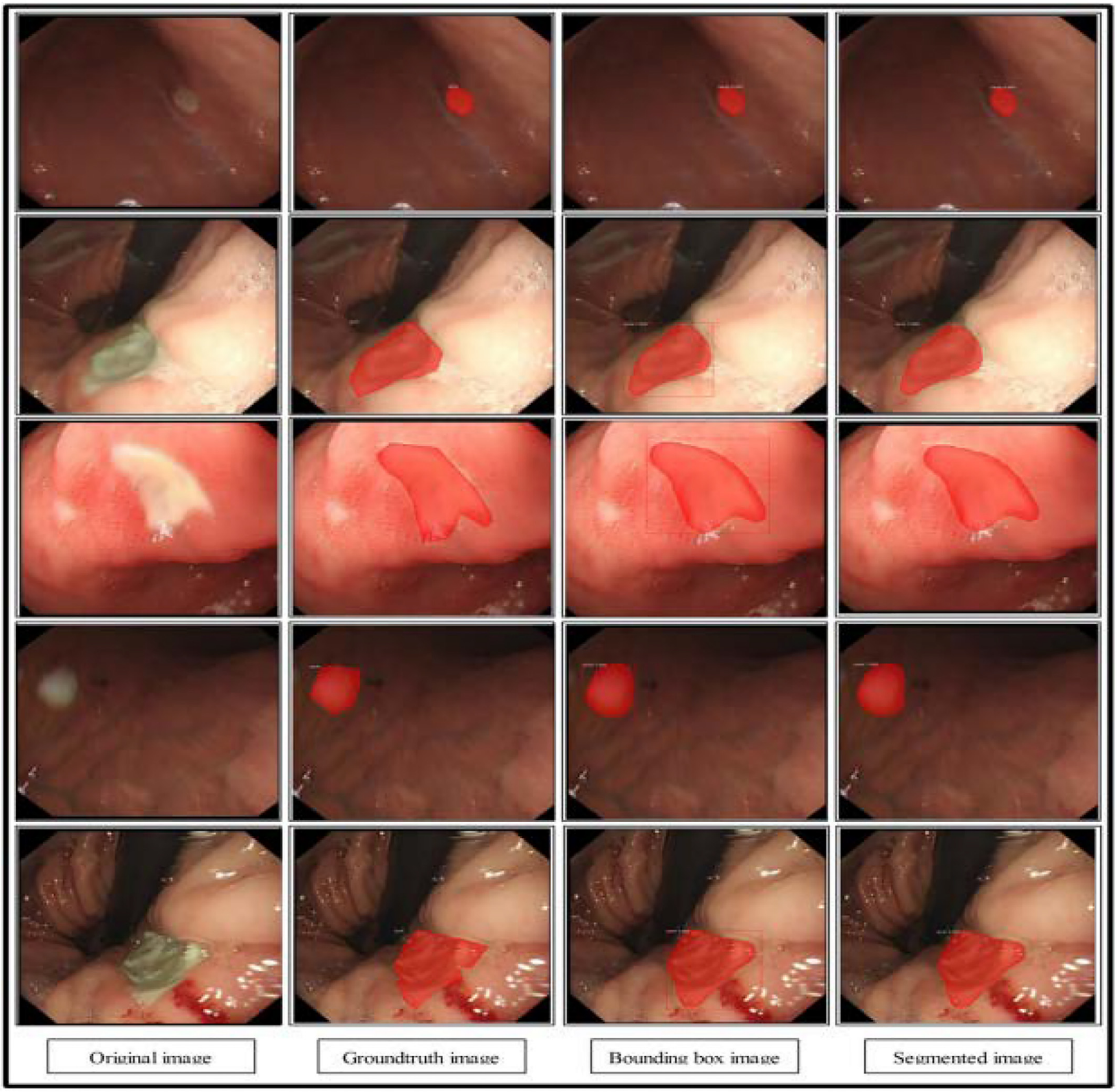


Fig. 3. Proposed Mask-RCNN bases ulcer segmentation.

Where $S_p$denotes the social interaction among GHs'. It is a core component during the process and computes the Euclidean Distance (ED) among pth and kth GHs' as:

$$S_p = \sum_{k=1,\, k\neq p}^{N} s(\varphi_{pk})\,\overbrace{\varphi_{pk}} \tag{5}$$

The symbol φ_{pk} denotes ED and $\overbrace{\varphi_{pk}}$ is a unit vector of *pth* GH to *kth* GH and computed through the following mathematical expression:

$$\varphi_{pk} = |f_k - f_p| \tag{6}$$

$$\overbrace{\varphi_{pk}} = \frac{|f_k - f_p|}{\varphi_{pk}} \tag{7}$$

The parameter $\mathbf{G}_p$ denotes the gravity force on pth GH and is mathematically computed as follows:

$$\mathbf{G}_p = -C \times \widehat{e_c} \tag{8}$$

Where, C denotes the gravitational cost and $\widehat{e_c}$ is a vertical unit vector. $\mathbf{A}_p$ denotes the wind advection and is computed as:

$$\mathbf{A}_p = u \times \widehat{e_w} \tag{9}$$

Here, u is a constant drift and $\widehat{e_w}$ is a unity vector according to wind pressure. Therefore, $\mathbf{S}_p$, $\mathbf{G}_p$, and $\mathbf{A}_p$ are utilized to update the position of each GH. In each iteration, the best GH is selected as a target and the remaining GHs' move towards the target for the next iteration. This process continues until all features P are updated. Mathematically, portion of each GH is updated as follows:

$$\Delta_f - \alpha\left(\sum_{k=p,\ k\neq p}^{N} \alpha \frac{ub_\varphi - lb_\varphi}{2} s\left(\left|f_k{}^{\varphi} - f_p{}^{\varphi}\right|\right) \frac{f_k - f_p}{\varphi_{pk}}\right) \tag{10}$$

The sigmoidal transfer function is utilized as a search agent, denoted by $\Delta_{\boldsymbol{sg}}$ which is later utilized in a probability function for position updating of GHs'. The probability function denoted by $\Delta_{q+1}^{k}(q+1)$ is mathematically defined as follows:

$$\Delta_{q+1}^{k}(q+1) = \begin{cases} 1 & if \quad r\,and < \Delta_{\boldsymbol{sg}+1} \\ 0 & if \quad r\,and \geq \Delta_{\boldsymbol{sg}+1} \end{cases} \tag{11}$$

The ED is utilized as a fitness function in this work which considers only those features whose distance from other features is minimum (less <) and achieves higher accuracy on a Multi SVM (MSVM) along Cubic kernel function. The final selected features are given to the MSVM to obtain results in the form of labeled and numerical values.

3. Experimental results and discussion

The proposed method is validated on a Private Dataset using different number of classifiers. The results are presented in two different steps; a) analysis of segmentation results using Mask-RCNN; b) analysis of the proposed classification results on the selected features.

3.1. Performance measures

The following performance measures are utilized to evaluate the

Intersection over union (I ∘ U) - For a given set of images, $I \circ U$ is used to measure the similarity between the region predicted by the model and its ground truth. It is given by the following equation.

$$I \circ U = \frac{A_g \cap A_p}{A_g \cup A_p} \tag{12}$$

This equation can be applied for the evaluation of bounding boxes and segmentation. The parameters A_p and A_g are the predicted and ground truth bounding box for segmentation.

Precision and Recall - Precision is the ratio between the correctly predicted observations with the total observations whereas Recall is the ratio between correctly predicted observations and all observations in the relevant class. Mathematically, Precision and Recall are defined as:

$$Precision = \frac{T_p}{T_p + F_p} \tag{13}$$

$$Recall = \frac{T_p}{T_p + F_n} \tag{14}$$

Where T_p denotes the number of cases which are true and detected correctly. F_p denotes those cases which are false but detected correctly. Similarly, F_n denotes the number of cases that are false but detected negative.

Average Precision ($\mathrm{A_p}$) - It is defined as the correctly predicted pixels divided by the number of pixels of that class in the ground truth. Mathematically, it is defined as follows:

$$Ap = \frac{T_p}{T_p + F_p + F_n} \tag{15}$$

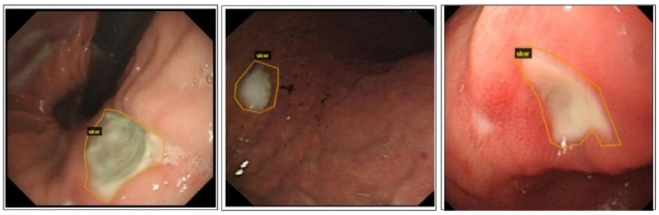


Fig. 4. Sample bounding box annotation provided by the physician.

3.2. Experimental setup

We have used the Mask-RCNN implemented by Matterport Inc. [29,30] using open source libraries. It is released under MIT License. For training we have used the pre-trained model which was trained on the MS-COCO dataset [23] by using transfer learning. It is used to solve the problem of small dataset because general features were already extracted. The model is implemented on Core i7 4770 CPU, 16GBRAM having NVIDIA GTX 1070 GPU. For implementation of the proposed system, we tested the ResNet50 + FPN and ResNet101 + FPN as backbone of the network during the segmentation process. Four cases are considered during the training process of Mask-RCNN- i) ResNet50+FPN as backbone and training of network heads only; ii) ResNet101+FPN as backbone and training of network heads only; iii) ResNet50+FPN as backbone and training of all the layers of Mask-RCNN; iv) ResNet101+FPN as backbone and training of all the layers of Mask-RCNN. In the first and second case, we train the network heads only. When only network heads are trained, weights of other layers will be same as that of the pre-trained model trained on the COCO dataset. To fine tune the model, the learning rate of 0.01 and 40 iterations were used. After 40 iterations, the loss is almost constant. It takes approximately 3–4 h to train the model in each case. In the classification stage, we utilized ResNet101 and trained the MSVM on 70% images from the selected dataset. All images are selected randomly and the remaining 30% was used for testing. The MSVM is utilized along cubic kernel function for classification.

3.3. Datasets

A private dataset [28] was used in this work for validation of segmentation and classification. This dataset consists of total 30 WCE videos of different patients. These videos are originally collected from the POF hospital, Wah Cantt, Pakistan. A total of 10 videos belong to ulcer patients and from them, we separate 500 images with the help of an expert who identify ulcer in the relevant images. The physician provided the bounding boxes of ulcer in the form of polygon. The VGG Image Annotator (VIA) [35] saves these bounding masks around the ulcer region in a JSON file as shown in Fig. 4. This JSON file is then used as ground truth of these images during the training process. The other 20 videos include 10 healthy patients and 10 bleeding patients. Moreover, we have also collected the polyp images from CVC–Clinic DB database. From this dataset, 612 polyp images are selected from different video sequences of colonoscopy. At the end, we perform data augmentation and increased images up to 1836 for each class (ulcer, bleeding, polyp, and healthy).

3.4. Segmentation results and analysis

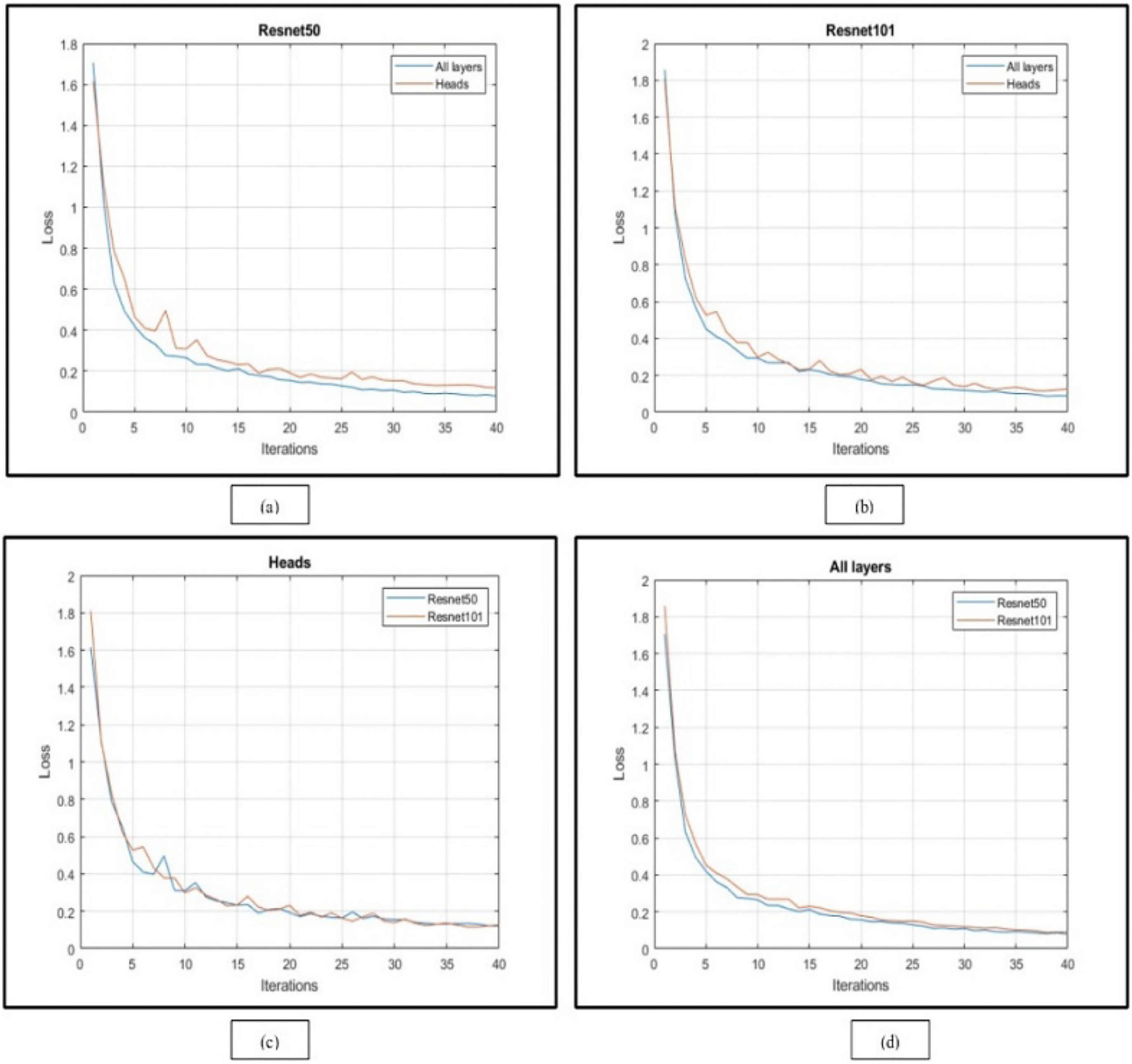
To evaluate the segmentation model, the precision and recall values are calculated at *IoU* along different threshold values of 0.5, 0.75, and 0.9. Results of case one and case three are presented in Tables 2 and 3. It is observed that values of precision and recall at lower threshold are better than higher threshold value. It is due to the fact that at low threshold value, the area which is not infected is falsely classified as ulcerous. By increasing the threshold, false identification is removed. Therefore, training of all the layers of network results are better than by training network heads only.

Similar behavior is observed when the same experiment was repeated by considering case two and case four; the results are given in Tables 4 and 5. The computed precision and recall values by training network heads only and all layers is 0.6666 for threshold0.5. Up to threshold0.75, case three works better than other cases but when the threshold value is increased to 0.9 , case two and case four gives better results. In Table 5, the values of precision and recall rate are0.2666 and 0.4666 which are higher than all other cases.

Table 5
Precision and Recall by using backbone ResNet101+FPN (Training all layers).

$I_{o}U$ Threshold Value	Precision	Recall
0.50	0.6666	0.6666
0.75	0. 6222	0. 6445
0.90	0.2666	0.4666

For further evaluation of each case, the Average precision (Ap) at threshold values of0.5, 0.75, and 0.9 along with the Mean Overlap Coefficient (MOC) are computed and shown in Tables 6 and 7. When all the layers of Mask-RCNN are trained by using ResNet50 as backbone, then the Ap value is reached to 1 for threshold values



(a)

(b)

(c)

(d)

Fig. 5. Training Loss of all cases (a, b, c, and d).

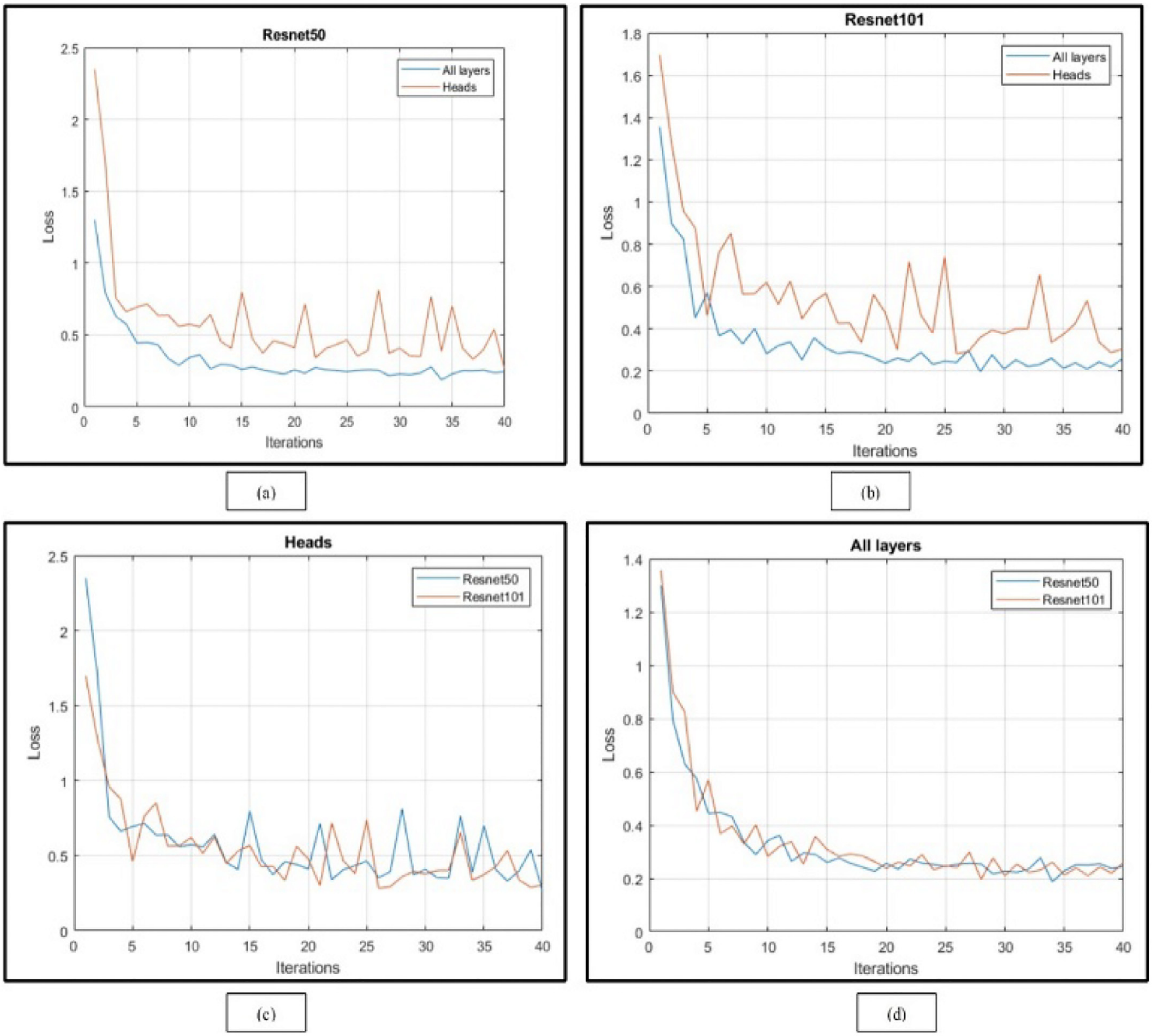


Fig. 6. Testing loss of Mask-RCNN.

Table 6
Training on network heads.

Backbone	$Ap_{0.5}$	Ap0.75	Ap0.9	Mean Overlap Coefficient
ResNet50+FPN	1.0	0.9667	0.2334	0.8576
ResNet101+FPN	1.0	0.9166	0.3500	0.8639

Table 7
Training all layers.

Backbone	$Ap_{0.5}$	Ap 0.75	Ap 0.9	Mean Overlap Coefficient
ResNet50+FPN	1.0	1.00	0.3	0.8807
ResNet101+FPN	1.0	0.93	0.4	0.8738

0.5and0.75. Moreover, the value of overlap coefficient is 0.8807as shown in Table 7, which is higher than all other cases. This ensures that the model with ResNet50+FPN as backbone and by training all layers (Case 3) works better than all other cases.

In the end, the loss is computed by employing loss function and is plotted in Fig. 5. As mentioned earlier, we perform four cases for validation of segmentation process. In Fig. 5, (a)–(d) loss is plotted for all cases. This results show that when all the layers of the network are trained, then the network consumes minimum time for training. Moreover, we also show that the loss of all the layers is comparatively less than training the network heads only.

In Fig. 6, the testing loss is presented. In Fig. 6(a) and (b), it is shown that when only heads of the network are trained, then it has more variation as compared to training loss of (c) and (d). On the other hand, when all the layers of the network are trained, the loss is smoothly reduced to a lower value. Similarly, on the same backbone, loss is almost similar like training. From all these discussions, it is clear that Case 3 (ResNet50+FPN as backbone and

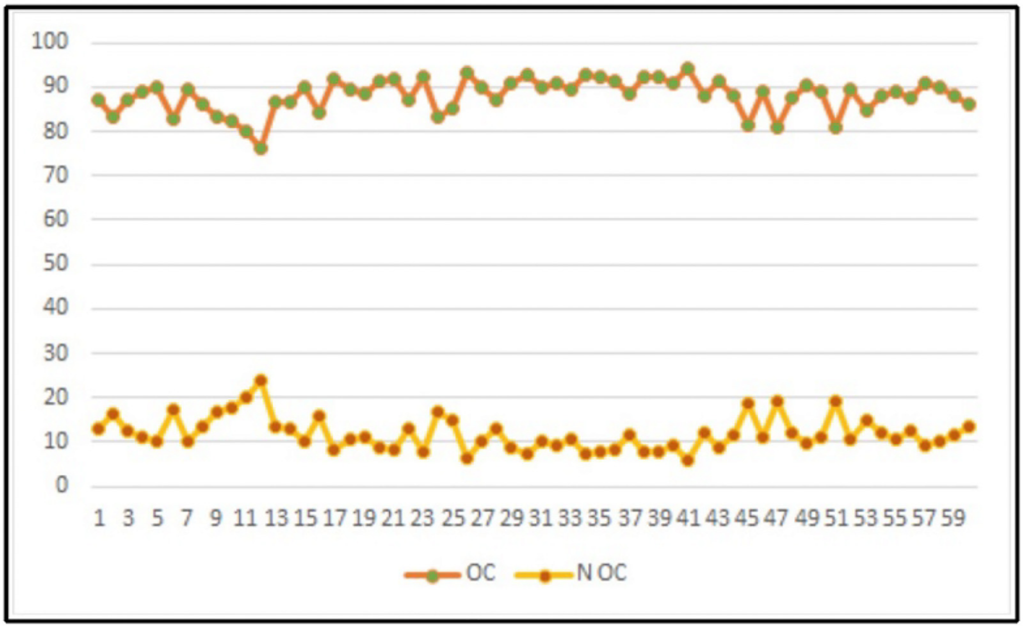


Fig. 7. Image wise demonstration of OC and false N OC.

Table 8
Proposed classification results using different K-folds.

Method	K-Fold validations				Performance Measures		
	$K = 5$	$K = 10$	$K = 15$	$K = 20$	Accuracy (%)	Error rate (%)	Time (sec)
Cubic SVM	√				99.92	0.08	25.75
		√			99.13	0.87	36.54
			√		99.32	0.68	49.18
				√	99.29	0.71	104.54
ESDA	√				97.82	2.18	100.65
		√			97.69	2.31	130.66
			√		96.99	3.01	46.04
				√	96.92	3.08	186.05
WKNN	√				96.42	3.58	127.56
		√			96.16	3.84	146.04
			√		96.09	3.91	216.41
				√	96.02	3.98	205.73
Decision Trees	√				96.48	3.52	172.24
		√			96.16	3.84	218.17
			√		95.92	4.08	240.51
				√	95.90	4.1	289.51

training of all the layers of Mask-RCNN) gives better results than other cases. By using this model, the segmentation overlapping coefficient for selected images is shown in Fig. 7 which depicts that the Average Overlapping Coefficient (AOC) is more than 80% and false Negative Overlapping Coefficient (NOC) is under 20%. A few sample Mask-RCNN based segmentation results are shown in Fig. 8.

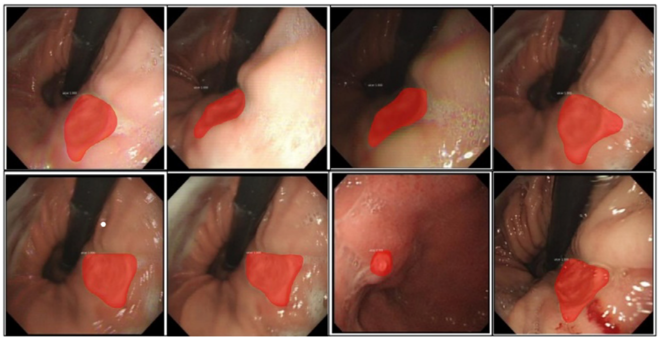


Fig. 8. Mask-RCNN based ulcer segmentation effects.

3.5. Classification results and analysis

In the classification phase, three stomach infections and one healthy class are considered. The infection classes are ulcer, bleeding, and polyps which have an average of 2000 images for each class. As mentioned above, 70% of the images are selected to train the model and the remaining for testing. The testing results are computed through different K folds, where$K = 5, 10, 15,$ and 20. Testing results are presented in Table 8 which shows that CSVM achieve the best performance as compared to other listed classifiers. The CSVM achieve accuracy of 99.92%on the proposed selected features for $K = 5$ which is confirmed by Fig. 9(a). By usingK $= 10,$ the best accuracy is 99.13%which is given by Fig. 9(b). The third best accuracy of CSVM is 99.32% on $K = 15$ which is verified by Fig. 9(c). We also compute classification accuracy using $K = 20$ which is 99.29% using CSVM. The maximum achieved accuracy on the ESDA classifier using the proposed selected features is 97.82% for$K = 5$. Similarly, for WKNN and decision trees,

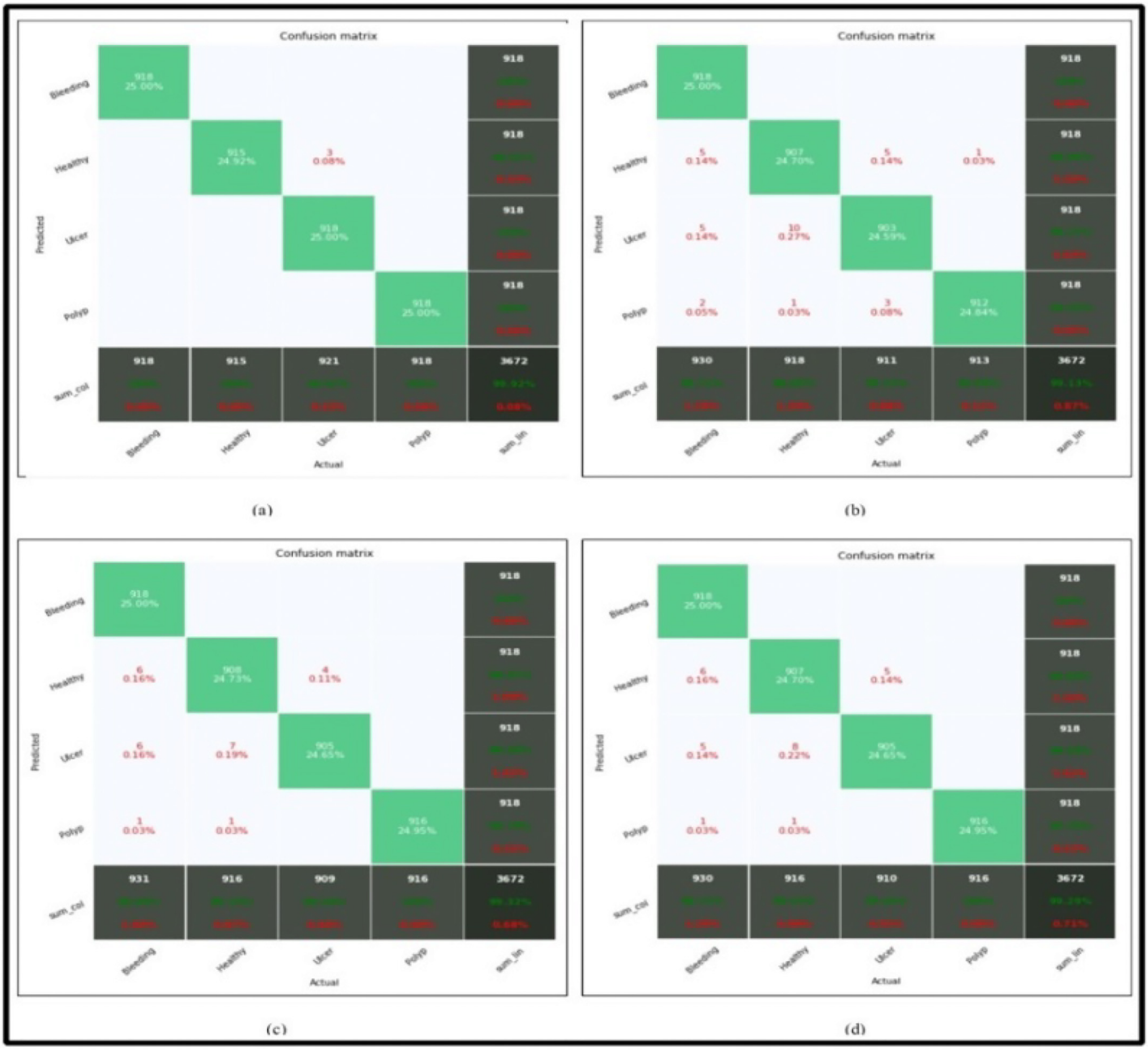


Fig. 9. Confusion matrix of Cubic SVM for different K-folds. (a) 5-Fold, (b) 10-Fold, (c) 15-Fold, and (d) 20-Fold.

the best accuracy achieved is 96.42%and96.48%, respectively when $K = 5$. From Tae 8 and Fig. 9, it is clearly shown that the proposed selected features outperformed on CSVM for $K = 5$ whereas the best average performance is reached when K = 10. Besides, we also compute the classification time of each classifier on each*K Fold*. The time is given in Table 8 which shows that the minimum testing time is 25.75 sec for $K = 5$, for $K = 10$,the best time is 36.54sec, for $K = 15$ the best noted time is 104.54 s, and 46.04 s when $K = 20$. Overall, the CSVM performs fast with better accuracy.

3.6. Analysis and comparison-

The proposed segmentation and best features selection-based classification results are analyzed in this section. First the segmentation results are presented using Mask-RCNN based method. The Mask-RCNN based segmentation results are evaluated on ulcer images based on four different experiments. As shown in Table 2, the backbone ResNet 50+FPN using network head only achieves best precision and recall rates of 0.6666 and 0.6666, respectively on 0.5 IoU threshold value.

On the same backbone using all layers, the best precision and recall rates are obtained on 0.5 and 0.75 IoU threshold values, given in Table 3. Similarly, the best recall and precision rates are 0.6666 computed on backbone ResNet101+FPN using heads only and all network layers as given in Tables 4 and 5. In Tables 6 and 7, the average precision rates are computed for both backbones and achieves best MOC= 0.8807. Moreover, the image wise Mask-RCNN based segmentation OC and NOC are plotted in Fig. 8 which shows the authenticity of segmentation process. The classification results are given in Table 8 which shows that $K = 10$gives overall better accuracy on the proposed selected features.

Furthermore, a detailed statistical analysis was also performed to analyze the consistency of proposed classification method, results can be seen in Table 9. In this table, the following parame-

Table 9
Statistical analysis of proposed classification results based on standard deviation (SD) and Confidence Interval (CI).

Method	K-Folds		Parameters				
	$K = 5$	$K = 10$	Min (%)	Avg (%)	Max (%)	SD	CI
Cubic	√		98.25	99.08	99.92	0.682	0.396
SVM		√	97.69	98.41	99.13	0.587	0.339

Table 10
Comparison with existing techniques.

Method	Year	Accuracy (%)
[12]	2018	98.49
[28]	2019	99.54
[36]	2017	97.89
Proposed Classification	2019	99.92
Proposed Segmentation	2019	88.08 OC

ters are calculated as: Maximum (Max), Minimum (Min), Average (Avg). Standard Deviation (SD), and Confidence Interval (CI). For results computation, 500 times iterations were performed for K-Fold (5 and 10). From the results, it is show that the proposed classification accuracy is still consistent and a very little change is occurred during initialized number of iterations.

In addition, the CSVM accuracy was compared with few other well-known methods for a fair comparison. Also, a comparison with the existing methods is given in Table 10. In this table, the achieved accuracies of few recent articles are added and compared with proposed results. From the results, it is show that the proposed accuracy is best in terms of classification accuracy and segmentation MOC [12] [28] [36].

4. Conclusion

A new duo-deep architecture-based system is proposed for gastrointestinal diseases segmentation and classification using WCE. The proposed method performance is evaluated in two different phases; first, the analysis of Mask-RCNN based ulcer segmentation performance through recall, average precision, and MOC. Second, the classification accuracy is computed which shows better performance as compared to existing techniques. The overall results show that the proposed duo-deep architecture works robustly for both ulcer segmentation and classification of selected gastrointestinal diseases. The Private Dataset is utilized for validation and achieves MOC of 88.08% for ulcer segmentation and 99.92% classification accuracy. From the results, we conclude that the modified Mask-RCNN based ulcer segmentation would yield an increase in the MOC rate if we have more number of training data. Moreover, we also conclude from the classification step that the best selected features helps to improve the performance of the overall system.

A few limitations of this work are: (i) Not correctly segment the ulcer regions on less training data; (ii) rely on the correctness of groundtruth data; (iii) fail for the segmentation of polyp and bleeding regions. In the future, we will consider these limitations and will design a system which can segment ulcer, polyps and bleeding regions from WCE images.

Declaration of Competing Interest

The authors declare that they do not have any conflict of interests. This research did not involve any human or animal participation. All authors have checked and agreed the submission.

References

[1] M. Mittal, L.M. Goyal, S. Kaur, I. Kaur, A. Verma, D.J. Hemanth, Deep learning based enhanced tumor segmentation approach for MR brain images, Appl. Soft Comput. 78 (2019) 346–354.

[2] M.I. Sharif, J.P. Li, M.A. Khan, and M.A. Saleem, "Active deep neural network features selection for segmentation and recognition of brain tumors using MRI images," Pattern Recognit. Lett., 2019.

[3] M.A. Khan, M. Sharif, T. Akram, S.A.C. Bukhari, and R.S. Nayak, "Developed Newton-Raphson based deep features selection framework for skin lesion recognition," Pattern Recognit. Lett., 2019.

[4] D.J. Hemanth, J. Anitha, M. Mittal, Diabetic retinopathy diagnosis from retinal images using modified hopfield neural network, J. Med. Syst. 42 (2018) 247.

[5] M.A. Khan, S. Rubab, A. Kashif, M.I. Sharif, N. Muhammad, J.H. Shah, et al., "Lungs cancer classification from ct images: an integrated design of contrast based classical features fusion and selection," Pattern Recognit Lett, 2019.

[6] M.A. Khan, M. Sharif, T. Akram, M. Yasmin, R.S. Nayak, Stomach deformities recognition using rank-based deep features selection, J Med Syst 43 (2019) 329.

[7] M.A. Khan, M. Rashid, M. Sharif, K. Javed, T. Akram, Classification of gastrointestinal diseases of stomach from WCE using improved saliency-based method and discriminant features selection, Multimed. Tools Appl. (2019) 1–28.

[8] W. Street, Cancer Facts & Figures 2019, American Cancer Society: Atlanta, GA, USA, 2019.

[9] L. Lan, C. Ye, C. Wang, S. Zhou, Deep convolutional neural networks for WCE abnormality detection: cnn architecture, region proposal and transfer learning, IEEE Access 7 (2019) 30017–30032.

[10] G. Iddan, G. Meron, A. Glukhovsky, Wireless capsule endoscopy. nature, vol 405 (2000) 25.

[11] B. Li, M.Q.-H. Meng, Tumor recognition in wireless capsule endoscopy images using textural features and SVM-based feature selection, IEEE Transactions on Information Technology in Biomedicine 16 (2012) 323–329.

[12] A. Liaqat, M.A. Khan, J.H. Shah, M. Sharif, M. Yasmin, S.L. Fernandes, Automated ulcer and bleeding classification from WCE images using multiple features fusion and selection, J. Mech. Med. Biol. 18 (2018) 1850038.

[13] Q. Wang, N. Pan, W. Xiong, H. Lu, N. Li, X. Zou, Reduction of bubble-like frames using a RSS filter in wireless capsule endoscopy video, Optics Laser Technol. 110 (2019) 152–157.

[14] Q. Al-Shebani, P. Premaratne, D.J. McAndrew, P.J. Vial, S. Abey, A frame reduction system based on a color structural similarity (CSS) method and Bayer images analysis for capsule endoscopy, Artif. Intell. Med. 94 (2019) 18–27.

[15] H.-G. Lee, M.-K. Choi, B.-S. Shin, S.-C. Lee, Reducing redundancy in wireless capsule endoscopy videos, Comput. Biol. Med. 43 (2013) 670–682.

[16] R. Sharma, R. Bhadu, S.K. Soni, and N. Varma, "Reduction of redundant frames in active wireless capsule endoscopy," in *Proceeding of the Second International Conference on Microelectronics, Computing & Communication Systems (MCCS 2017)*, 2019, pp. 1–7.

[17] T. Aoki, A. Yamada, K. Aoyama, H. Saito, A. Tsuboi, A. Nakada, et al., Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network, Gastrointest. Endosc. 89 (2019) 357–363 e2.

[18] T. Saba, M.A. Khan, A. Rehman, S.L. Marie-Sainte, Region extraction and classification of skin cancer: a heterogeneous framework of deep CNN features fusion and reduction, J. Med. Syst. 43 (2019) 289.

[19] F. Afza, M.A. Khan, M. Sharif, and A. Rehman, "Microscopic skin laceration segmentation and classification: a framework of statistical normal distribution and optimal feature selection," Microsc. Res. Tech., 2019.

[20] M.A. Khan, T. Akram, M. Sharif, T. Saba, K. Javed, I.U. Lali, et al., Construction of saliency map and hybrid set of features for efficient segmentation and classification of skin lesion, Microsc. Res. Tech. 82 (2019) 741–763.

[21] J.K. Sethi, M. Mittal, A new feature selection method based on machine learning technique for air quality dataset, J. Stat. Manage. Syst. 22 (2019) 697–705.

[22] S. Fan, L. Xu, Y. Fan, K. Wei, L. Li, Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images, Phys. Med. Biol. 63 (2018) 165001.

[23] M.A. Khan, T. Akram, M. Sharif, K. Javed, M. Rashid, S.A.C. Bukhari, An integrated framework of skin lesion detection and recognition through saliency method and optimal deep neural network features selection, Neural Comput. Appl. (2019) 1–20.

[24] M. Souaidi, A.A. Abdelouahed, M. El Ansari, Multi-scale completed local binary patterns for ulcer detection in wireless capsule endoscopy images, Multimed. Tools Appl. 78 (2019) 13091–13108.

[25] O. Bchir, M.M.B. Ismail, N. AlZahrani, Multiple bleeding detection in wireless capsule endoscopy, Signal Image Video Process. 13 (2019) 121–126.

[26] H. Alaskar, A. Hussain, N. Al-Aseem, P. Liatsis, D. Al-Jumeily, Application of convolutional neural networks for automated ulcer detection in wireless capsule endoscopy images, Sensors 19 (2019) 1265.

[27] J.-.Y. He, X. Wu, Y.-.G. Jiang, Q. Peng, R. Jain, Hookworm detection in wireless capsule endoscopy images with deep learning, IEEE Trans. Image Process. 27 (2018) 2379–2392.

[28] M. Sharif, M. Attique Khan, M. Rashid, M. Yasmin, F. Afza, U.J. Tanik, Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images, J. Exp. Theoret. Artif. Intell. (2019) 1–23.

[29] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference On Computer Vision*, 2017, pp. 2961–2969.

[30] W. Abdulla, "Mask R-CNN for object detection and instance segmentation on Keras and tensorflow; 2017," ed.

[31] T.-.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[32] J.R. Uijlings, K.E. Van De Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, Int. J. Comput. Vis. 104 (2013) 154–171.

[33] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[35] A. Dutta and A. Zisserman, "The vgg image annotator (VIA)," arXiv preprint arXiv:1904.10699, 2019.

[36] S. Suman, F. Hussin, A. Malik, S. Ho, I. Hilmi, A. Leow, et al., Feature selection and classification of ulcerated lesions using statistical analysis for wce images, Appl. Sci. 7 (2017) 1097.