Received: March 13, 2024. Revised: May 1, 2024.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

GastroFPN: Advanced Deep Segmentation Model for Gastrointestinal Disease with Enhanced Feature Pyramid Network Decoder

Haithem Kareem Abass[1]* Mina H. Al-hashimi[1] Ammar S. Al-Zubaidi[2]
Mohammed Al-Mukhtar[2]

[1]*Computer Engineering Department, Al-Mansour University College, Baghdad, Iraq*
[2]*Computer Centre, University of Baghdad, Baghdad, Iraq*
*Corresponding author's Email: ammar.sabah@cc.uobaghdad.edu.iq

Abstract: The early identification of gastric cancer holds considerable importance within medicine due to its crucial role in mitigating fatality rates. Currently, artificial identification and annotations using gastroscopic images are the main methods of evaluation. Nevertheless, doctors have significant difficulties implementing these techniques due to the considerable heterogeneity in the visual characteristics of early cancer tumors. Weakness in segmentation remains the biggest obstacle to accurate detection and extraction of the main lesion from the tumor. In this paper, we propose a deep model combining two networks encoded: Unet++ and the feature pyramid network. The encoder backbone on first detection is based on ResNet34, which feeds the feature extraction to the next step. The second step is adding an enhanced feature pyramid network by merging blocks and final segmentation heads. The decoder improves the model's capacity to collect hierarchical characteristics at many levels, resulting in enhanced segmentation performance that adapts to changes in illness symptoms. The proposed model achieved a segmentation accuracy of 96.8%, a dice-score of 86.6%, and an F1-score of 85.3% when using the EDD2020 dataset. While the accuracy for DCSA-Unet achieved 92%, Unet++ 90%, 77% for FPN, and DeepLabv3+ 94.2%, We trained the proposed model on two different datasets, the CVC-ClinicDB and Kvasir-Seg datasets. For CVC-ClinicDB, the results metrics registered a Dice-Score 91.64%, an IoU of 84.63%, and an accuracy of 98.55%. For the Kvasir-Seg dataset, the Dice score is 92.54%, the IoU is 87.57%, and the accuracy is 96.62%.

Keywords: Neural network, Segmentation, Unet++, FPN, Gastroscopy.

1. Introduction

Gastrointestinal cancer is a malignant tumor that primarily affects the stomach mucosa and is the second leading cause of mortality among all malignancies, behind lung cancer [1, 2]. This illness exhibits a strong geographical distribution, with over 50% of cases concentrated in East Asia. Advancements in diagnostic and therapeutic methods for gastric cancer are continuously enhancing, and it has been demonstrated that early diagnosis significantly decreases death rates in individuals diagnosed with stomach cancer. Gastric inspection is conducted with the use of endoscopy and gastro fluoroscopy, employing barium as a contrast agent. Endoscopy is extensively used for stomach cancer screening and comprehensive diagnosis because of its exceptional sensitivity in detecting early-stage gastric cancer and its ability to facilitate tissue collection and treatment while being seen. Due to the continuous rise in the quantity of images captured by endoscopic equipment and the ongoing enhancement in image quality, doctors who heavily rely on visual examination for diagnosis are more burdened by weariness. Moreover, many physicians employ distinct cognitive shortcuts derived from their expertise. Therefore, due to personal prejudices and sometimes a lack of energy, misdiagnoses happen[3].

Nevertheless, the process is very intricate, necessitating the completion of several duties during the assessment. Therefore, there is a certain level of

apprehension that lesions could go unnoticed. Based on the survey findings, there is a 22.2% chance of not detecting a lesion during endoscopy[4]. The accuracy of a diagnosis relies heavily on the expertise and proficiency of the clinician.

Hence, there is a great need for technology that might aid in exams to alleviate the workload of physicians and enhance diagnostic precision. Artificial intelligence technologies have made significant advancements in recent years, with deep learning technology demonstrating exceptional performance in the domain of image identification[5]. Deep learning methodologies have been suggested for several applications, including numerous categories of medical images[6, 7].

Conventional approaches for identifying gastrointestinal stromal tumors often entail categorizing the image based on several characteristics, such as the size, form, and texture of the lesion. These methods rely on manually produced features. These approaches are very responsive to the quality of the images and produce inadequate diagnoses in many situations [8]. The division of the image into segments affects the categorization of the image. An inaccuracy in the segmentation process might lead to failure in the subsequent classification. The conventional approaches exhibit limited resilience due to their reliance on hand crafted features in ultrasound images[9]. A deep convolutional neural network (CNN) is capable of extracting features at various levels without the need for manually designed features. This eliminates the reliance on expert knowledge and reduces the processing time required [10]. Automatic segmentation is a cutting-edge technology that is being used in tumor or cancer detection. Extracting the region of interest enhances accuracy[9]. However, Gastrointestinal malignancies exhibit significant heterogeneity in terms of their anatomical site, structure, and histopathological features. Developing AI models that can precisely identify and separate tumors in all their complexity and diversity is a challenging endeavor that frequently necessitates the use of advanced algorithms capable of capturing subtle characteristics. Furthermore, acquiring different and precisely annotated information may be difficult due to disparities in imaging methods, patient demographics, and tumor attributes.

To mitigate this problem, we proposed model provides an improved automated segmentation mechanism utilizing Gastro FPN. Which is applied in three stages, the initial residual model is used to extract the main features. Accordingly, the whole information feed was expanded to Unet++, which forwarded the output results to the block of the feature pyramid network (FPN) decoder model.

UNet++ is an enhanced version of the original UNet architecture. It includes skip connections of varying lengths, allowing for the integration of features in a variety of sizes. This facilitates the acquisition of both local and global contextual information, resulting in enhanced segmentation performance, particularly for objects with diverse dimensions. It allows for the efficient transmission of high-resolution features across the network. This allows the model to accurately capture detailed features and enhance the accuracy of segmentation, especially in regions with complex architecture.

The FPN enhances the accuracy of object recognition by creating a hierarchical structure of features that include both high-level meaning and semantics. This allows for object detection at various sizes and scales. This solution tackles the difficulties caused by variations in image scale, enabling the model to more efficiently identify objects of different sizes.

The main contributions of this paper are as follows:

- We expanded the UNET++ model jointly with the residual model.
- We added the FPN merge block, which consisted of four segmentation blocks.
- Rescaling a UNet++ model sometimes entails making architectural changes to the model while preserving its fundamental structure and functioning to handle inputs of varying sizes.
- The main objective of FPN is to address the challenge of detecting objects of different sizes within an image. Objects of diverse nature might possess distinct dimensions, and the ability to identify objects at different.

The remainder of the paper is organized as follows. Section 2 provides an overview of several related works. Section 3 describes the proposed model for Segmentation Model for Gastrointestinal with Enhanced FPN Decoder. The dataset, experimental results, and analysis are presented in Section 4, and the proposed study is concluded in Section 5

2. Literature review

To separate polyps from colonoscopy images, the authors introduce a multi-scale subtraction network[11]. First, Res2Net-50 is used as the primary framework to extract features at five different levels for enhancing features. Second, they intentionally provide the subtraction unit with

receptive fields of varying sizes at different levels, resulting in a diverse range of unique information at multiple scales. Finally, they have developed a network named "LossNet" that effectively manages polyp-aware features across all network layers, eliminating the need for training. This approach enables the model to acquire complex and structural data at the same time. However, Res2Net-50 is susceptible to overfitting, especially when trained on small datasets or datasets with substantial noise.

The authors suggested that the Deep Feature Aggregation Decoder is a technique that selectively combines just the most relevant features at different depths to extract important lesion characteristics. This approach helps to simplify the model and improve its efficiency[12]. In addition, they create a feature fusion module that uses multi-modal fusion methods to interact with independent features from different modalities. They use the linear Hadamard product to merge the feature information from both branches. Finally, for joint training, they evaluate the transformer loss, the U-Net loss, and the fused loss against the ground truth label. The test results show that when using the Kvasir dataset, the suggested method achieves an accuracy of 94.0%. Nevertheless, the multi-modal fusion method's scalability is a challenge, especially when dealing with large datasets.

This study introduces a framework for segmenting lesion areas in endoscopic images, known as a dual-guided network[13]. The network consists of two components: the bilateral attention branch and the border aggregation branch. the bilateral attention branch is used for developing a mask decoder known as the Progressive Partial Decoder and a module known as the Full-Context Bilateral Relation. This branch's main goal is to improve the correlation between foreground and background signals in the images to address the uncertain borders of lesion areas. The boundary aggregation branch consists of a boundary decoder, known as boundary-aware extraction, and a module, known as boundary-guided feature aggregation. This model achieved an accuracy of 96.48 when utilizing the dataset Kvasir. Nevertheless, due to the sequential decoding of the Progressive Partial decoder, there is a potential for loss of information in the first decoding phases. This has the potential to affect the overall accuracy of the decoded output, particularly when handling intricate or subtle data.

The authors introduced the SSFormer, a model for medical image segmentation that incorporates a pyramid transformer encoder to enhance the models' generalization capability[14]. They suggested customizing the Progressive Locality Decoder for the pyramid transformer backbone to highlight local characteristics and limit attention dispersion. The SSFormer model gets an accuracy of 96.02%, when using the Kvasir dataset. However, if the training data doesn't adequately represent these scales, the Pyramid Transformer may struggle to adjust to images with significant scale or aspect ratio differences.

Google designed DeepLabv3+, a convolutional neural network model, to segment images using deep learning techniques[15]. This is an expansion of the DeepLabv3 model, which was already an enhancement in earlier iterations. DeepLabv3+ uses a deep convolutional neural network to give semantic labels to every pixel in an image, essentially dividing the images into several classes or categories. However, DeepLabv3+ may have boundary artifacts, which occur when the segmentation boundaries do not coincide precisely with the borders of objects in the image. This can lead to errors in the segmentation results.

The authors suggested a design that includes an encoder structure that uses pretrained Mix Transformer encoders and an efficient stage-wise feature pyramid decoder structure[16]. The method provides physicians with a possible tool to accurately segment and detect lesions in real-time. When using the Kvasir dataset, the accuracy is 95.99 and the mIoU is 85.96. The dice score is 92.17. However, Mix Transformers rely on the dynamic selection of expert groups based on input characteristics. Creating efficient selection processes that attain a compromise between computing economy and model performance may be a difficult task, sometimes necessitating lengthy testing and fine-tuning.

For medical image segmentation, the authors proposed UNet++[17]. The design consists of an encoder-decoder network, with the encoder and decoder sub-networks coupled by a sequence of nested, dense skip routes. The redesigned skip paths have the objective of minimizing the semantic disparity between the feature maps of the encoder and decoder sub-networks. Kvasir-SEG serves as the training dataset. The results of our studies indicate that UNet++ with deep supervision produces an accuracy of 93.94. However, the model provides inadequate accuracy.

The author proposed the DCSAU-Net model using a primary feature conservation strategy and a compact split-attention block following the encoder-decoder architecture for segmenting medical images[18]. The results show that the proposed architecture achieves high scores.

Computer vision applications such as semantic segmentation and object detection frequently use FPNs as their architectural design. Their objective is to tackle the task of identifying items of varying sizes within an image[19]. FPNs aim to construct a feature pyramid by merging features from many convolutional layers with varied resolutions. This enables objects of varying sizes to be identified. However, complex structures or congested backgrounds may still restrict the effectiveness of FPN in capturing contextual information at multiple scales. This can result in inaccuracies in segmentation.

UNet is a specialized CNN structure created specifically for the purpose of segmenting biomedical images, with a particular focus on the field of medical imaging[20]. Nevertheless, UNet has boundary artifacts, in which the segmented borders may not precisely coincide with the boundaries of objects in the image. This can result in mistakes, particularly in areas where objects possess intricate forms or unclear borders.

3. Methodology

Our proposed model has three stages: ResNet-34 [21] is composed of 34 layers, which are applied to the encoder part. The composition consists of the remaining fundamental components that may be categorized into many phases. ResNet is constructed by concatenating numerous residual blocks in a sequential manner. The general structure entails the arrangement of residual blocks with skip connections in a stacked manner. We chose ResNet as the encoder process because it introduces the notion of bottleneck blocks for deeper networks. These blocks use 1x1, 3x3, and 1x1 convolutions to decrease computational costs. By including skip connections, the backpropagation process is facilitated, thereby reducing the issue of vanishing gradients and enabling the successful training of very complex networks. The process of block input and output can be described in Eq. (1)

$$out_{ConV} = Re\,L\,UB_i\left(B_{i-1}\left(B_2\left(B_1 BN(IN_{ConV})\right)\right)\right) + In\_ConV \quad (1)$$

Relu: Linear Unit activation function.

Bi: represent different convolutional layers.

BN: batch normalization operation.

IN_ConV represents the input to the convolutional layer.

It is an equation that typically appears in convolutional neural networks (CNNs), which involves convolutional layers, batch normalization, and ReLU activation functions. The deep network topology of ResNet makes the model's success highly dependent on the search for an appropriate learning rate during training. The optimal starting learning rate may be determined by simply using the learn_lr_find function. Once the initial learning rate has been determined, use the learning rate method known as the one-cycle policy to train the model and begin the training process. Basically, a cycle consists of two phases: an initial phase where the learning rate gradually increases from a lower value to a higher value, followed by a subsequent phase when the learning rate decreases back to the lowest value. To determine the maximum learning rate, one should gradually raise the learning rate from a minimal value to a substantial value and halt when the loss starts to become uncontrollable. However, after the convolutional layers, the network used a 3x3 max-pooling layer, an average pooling layer, and a fully connected layer. A ResNet-34 model, following conventional architecture, consists of 63.5 million parameters. Rectification nonlinearity (ReLU) activation and batch normalization (BN) are applied to the convolution layers in the "Basic Block" block. The final layer utilizes the softmax function. The input vector and the vector that is output via the convolutional layer may be added directly [22], and the result can then be output through the activation function known as the rectified linear unit (ReLU). To tackle the issue of the vanishing gradient problem that occurs during training, Eq. (2) is used.

$$AH(x) = F(x) + x \quad (2)$$

H(x): represents the outcome of a specific layer or group of layers.

F(x): combines the extra information with the input x to produce the desired output of the residual block.

(x): denotes the input that is fed into the residual block.

However, when the number of F (x) and x channels is different in Eq. (2), the identity mapping cannot be connected to the next convolutional layer. A Rectified Linear Unit (ReLU) activation function and an extra batch normalization layer after each convolution layer. Figure 1 depicts that the last batch normalization step and the output is element-wise mapped using identity mapping. After a series of convolution processes, the resolution of the feature image is extremely low; thus, transpose

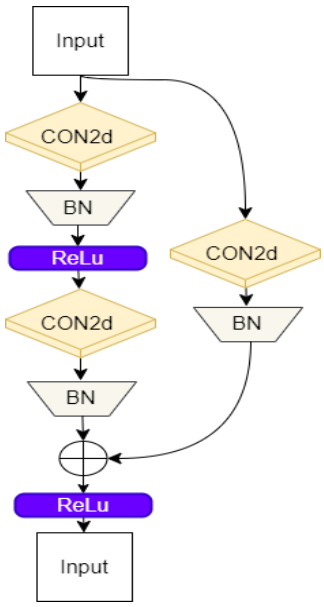

Figure. 1 3×3 convolution layers and an identity mapping, which is also called a shortcut connection, make up the residual block

convolution is used to expand the size of the feature map. This is the reason why the transpose convolution layer was added. The sampling of the feature matrix results in an increase in the dimension of the matrix. The forward propagation of the maximum pooling layer is designed to send the maximum value in the patch to the subsequent layer, while the values of the remaining pixels are immediately ignored. This is the reason why the maximum pooling layer is being added.

For transfer learning, convolutional neural networks (CNNs), particularly those with more layers, possess a significant quantity of parameters. To achieve good generalization to new, unknown data, it is usually necessary to have a large and varied dataset while training complicated models. Methods such as transfer learning and fine-tuning can benefit from annotated datasets.

Using the information gathered from the wider pre-training, smaller annotated datasets may be used to adjust pre-trained models on big datasets to objectives. Transfer learning is often used on a limited number of datasets to train the neural network, and it has been shown to be a very efficient technique.

Thus, the goal of implementing transfer learning is to improve the learning algorithms' performance via the transfer of knowledge. There are three primary ways in which transfer learning is advantageous. First, the pre-training knowledge; second, the training time needed to master the objective task; and finally, the performance after training [23]. In a neural network, residual connections are connections that enable information to skip one or more levels. A neural network, such as Unet, may have several layers. It is common practice to take advantage of these connections to solve the issue of disappearing gradients and to assist in the movement of information across the network. UNet++ [24] was developed with the primary goal of bridging the semantic gap that exists between the feature maps of the encoder and the decoder before the fusion was performed.

The encoder and decoder are the two main components of most segment models. The first step is to reduce the dimensionality of the supplied data while preserving its key characteristics. Second, it takes the compressed representation of the input data and uses it to rebuild the input data, producing an output that is ideally identical to the input. Most networks use "connectivity transformation," which implies a modification or restructuring in the way the various levels of the encoder and decoder are linked. One such approach is to modify the configuration of skip connections to optimize the transmission of information and strengthen the network's ability to learn.

The GastroFPN improves feature extraction when used in conjunction with ResNet model. The objective of this redesign is to revolutionize the transmission of information between the encoder and decoder sub-networks, with the potential to enhance the network's capacity to accurately capture and recreate characteristics in the data. These adjustments are often used to optimize the performance and training dynamics of neural networks for applications. They move through a small convolutional block, the number of layers of which depends on the pyramid level. The encoder's feature maps are semantically closer to the decoder's feature maps after passing through the dense convolution block. Let x be the input for Eq. (3) to a specific layer in the neural network.

$$\text{Output}(x) = H(x) + \text{Skip}(x) \tag{3}$$

Output (*x):* denotes the result produced by a certain layer or block inside the network.

H(*x*): denotes the result produced by the layer or block itself.

Skip(x): denotes the result of a preceding layer or block that is being circumvented.

The main role of GastroFPN is to work faster to collect the final segmentation map, which is chosen from just one of the segmentation branches being considered. This is related to several different semantic levels. the backbone UNet++ can produce

feature maps with high resolution. The hypercolumn concatenation technique is used by UNet++. This technique includes the simultaneous concatenation of feature maps from several layers of the network. Consequently, this aids in collecting information on several scales, which in turn makes the model more resistant to fluctuations in the sizes and forms of objects. To provide a more comprehensive and precise representation of features at different sizes, UNet++ incorporates layered skip connections and deep supervision into its architecture. Because of this, the model can tell the difference between the input image's fine and coarse details.

The generation of hypotheses about the locations of objects is accomplished by modern object detection networks via the use of area proposal algorithms. Propose the implementation of a Region Proposal Network (RPN) that can use convolutional features from the whole image, allowing for region proposals without significant additional computational expense[25]. RPN is a kind of neural network that performs fully convolutional operations. It can forecast both the boundaries of objects and the scores indicating the presence of objects at each point.Fast R-CNN [26] reaches near-real-time performance by using deep networks without considering the time used for region proposals. Presently, recommendations provide a substantial hindrance to the computational efficiency of state-of-the-art detection systems.

Our model with the hierarchical structure of a ConvNet is used to extract the best information, which contains semantic information ranging from low to high levels, to construct a feature pyramid that maintains high-level semantics consistently. The pyramid model involves two phases: a bottom-up pathway and a top-down pathway [27]. Firstly, the computation of the backbone using the feedforward methodology ConvNet entails the computation of a feature hierarchy that encompasses feature maps at many scales, where each scale is twice the size of the preceding one. There is a distinct pyramidal level that corresponds to each phase, which we define. We will choose the output of the last layer of each phase as our reference set of feature maps to create our pyramid. This will allow us to complete the construction of our pyramid. These maps are going to undergo additional development. Secondly, the top-down technique improves the clarity of characteristics by increasing the spatial resolution of feature maps at higher pyramid levels. Initially, these maps were less intricate but included more substantial data. Because of fewer subsampling operations, the bottom-up feature map can collect lower-level semantic information, and the activations that are captured are more correctly localized.

The basis of our top-down feature maps is formed by this building component. To enhance the spatial resolution of a feature map with a lower resolution, we use a two-fold increase (using nearest neighbor-up sampling for simplicity). Along with the bottom-up map that corresponds to the up-sampled map. We add the merge block, which consists of four segmentation blocks for each block input and output (2,128,64,64), and the final layer is dropout2d. After that, all the extracted features went forward to the segmentation head. This block contains a conv2d layer, an upsampling bilinear layer, and finally an activation layer of 256 x 256.

An image pyramid refers to a multi-scale depiction of an image. Iterative subsampling or scaling of the image produces this kind of representation, which results in a series of images of various sizes. Within the pyramid, each tier displays the image at a distinct resolution or size compared to the preceding tier. By using shared classifiers across all levels, the architectural design is streamlined, particularly when the feature dimension remains constant. It allows for the use of the same parameters (weights and biases) to process features of different sizes, leading to enhanced generalization and reduced computational complexity. Our solution entails using a single-scale image of any dimension as the input and producing feature maps at different levels that are suitably scaled. This is accomplished by a whole convolutional procedure. The impressive results obtained via parameter sharing indicate that all strata of our pyramid exhibit similar degrees of semantic understanding. This advantage is comparable to that of using a featured image pyramid, in which a common head classifier may be used for features computed at any image scale.

In the FPN decoder model, the backbone network Unet++ is applied, as shown in Figure. 2, which consists of the first four decoder blocks at 128x128 and ends at 16x16. The second part includes three decoder blocks, the first of which is 128x128 and the last of which is 32x32. The third row illustrates the two decoder blocks at 128x128 and 64x64.

The final block will be 128x128 and forward to the decoder block at 256x256. The last segmentation block is 256 x 256, including the activation map.

Rescaling a UNet++ model sometimes entails making architectural changes to the model while preserving its fundamental structure and functioning to handle inputs of varying sizes. Semantic segmentation tasks are often assigned to basic UNet,

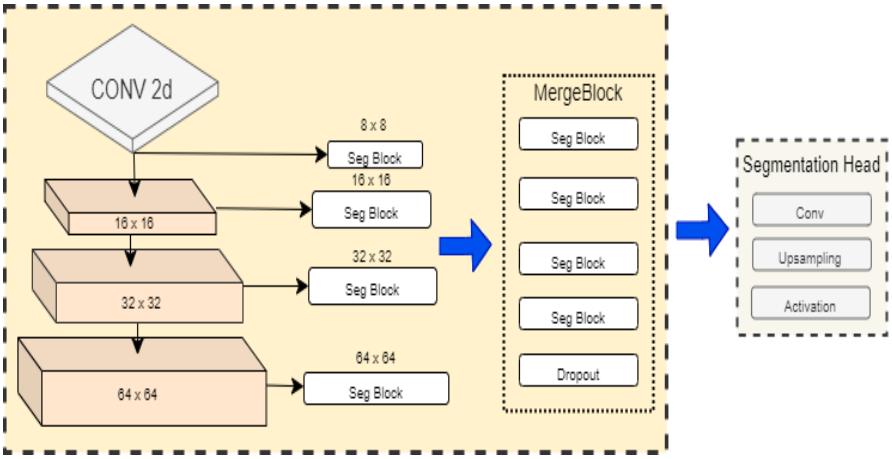

Figure. 2 Merge Decoded Layers of our proposed model (GastroFPN)

a prominent convolutional neural network architecture, in the context of medical image processing.

Adjustments to the input and output layers of the UNet++ model are made to generate images with the specified dimensions. This process entails modifying the dimensions of the input layer to correspond with the size of the input picture and ensuring that the output layer generates

segmentation maps that have the same dimensions as the input image.

However, to ensure that the size of the input is properly accounted for, it is recommended to modify the number of feature mappings in the encoder and decoder layers accordingly. To do this, it may be necessary to modify the dimensions of the feature maps in each layer or vary the quantity of convolutional filters. The most effective step is making an update to the Pooling and Up sampling Layers. Ensure that they are configured to accommodate input images of varying dimensions. This will include altering the size of the pooling window or the stride of the pooling layers, as well as tweaking the scaling factor of the up-sampling layers. Our model recomputes skip connections to account for the changes in feature map sizes resulting from the rescaling of the UNet++ architecture.

Eq. (5) represents the flattening prediction probabilities of the ground truths of the image. The skip pathway represent as $x^{i;j}$ for Where "i" represents the index of the down-sampling layer in the encoder, and "j" represents the index of the convolution layer in the dense block along the skip-connection.

$$X^{i,j} = \begin{cases} H\left(x^{i-1,j}\right), j = 0 \\ H\left(\left[x^{i,k}\right]_{k=0}^{j-1}, U\left(x^{i+1,j-1}\right)\right), j > 0 \end{cases} \quad (4)$$

X^{ij}: represent skip pathway connection.

H(.): for a convolutional process with an activation function

U(.): represents the upsampling layer and concatenation layer.

When j is 0, that means it will receive only one input from the encoder. While j equals 1, the node will feed two inputs, first from the encoder and second from the sub-skip connection. While the batch size indicated by N

$$P(Y, \hat{Y}) = -\frac{1}{N}\sum_{b=1}^{N}\left(\frac{1}{2}.Y_b.log\hat{Y}_b + \frac{2.Y_b.\hat{Y}_b}{Y_b+\hat{Y}_b}\right) \quad (5)$$

N: represents the overall quantity of samples or occurrences.

Y_b: denotes the true label for the b-th sample.

$\hat{Y}_b$: denotes the predicted label for the *b*-th sample.

After rescaling the UNet++ architecture, it's important to reevaluate the regularization techniques (such as dropout or batch normalization) and optimization parameters.

(such as learning rate and batch size) to ensure optimal performance during training. To increase the depth of our proposed model, we feed the output of the UNET++ to the FPN Net model with a novel structure, as shown in Figure 3.

The core principle of UNet++ is the integration of characteristics acquired from different levels of resolution inside the network. This allows for the integration of both low-level and high-level data, leading to improved accuracy in segmentation. The second part of our model is related to FPN, as we revisited and added merge-decoded layers and heads of segmentation.

The advantage of our model clarifies that the main objective of FPN is to address the challenge of detecting objects of different sizes within an image.

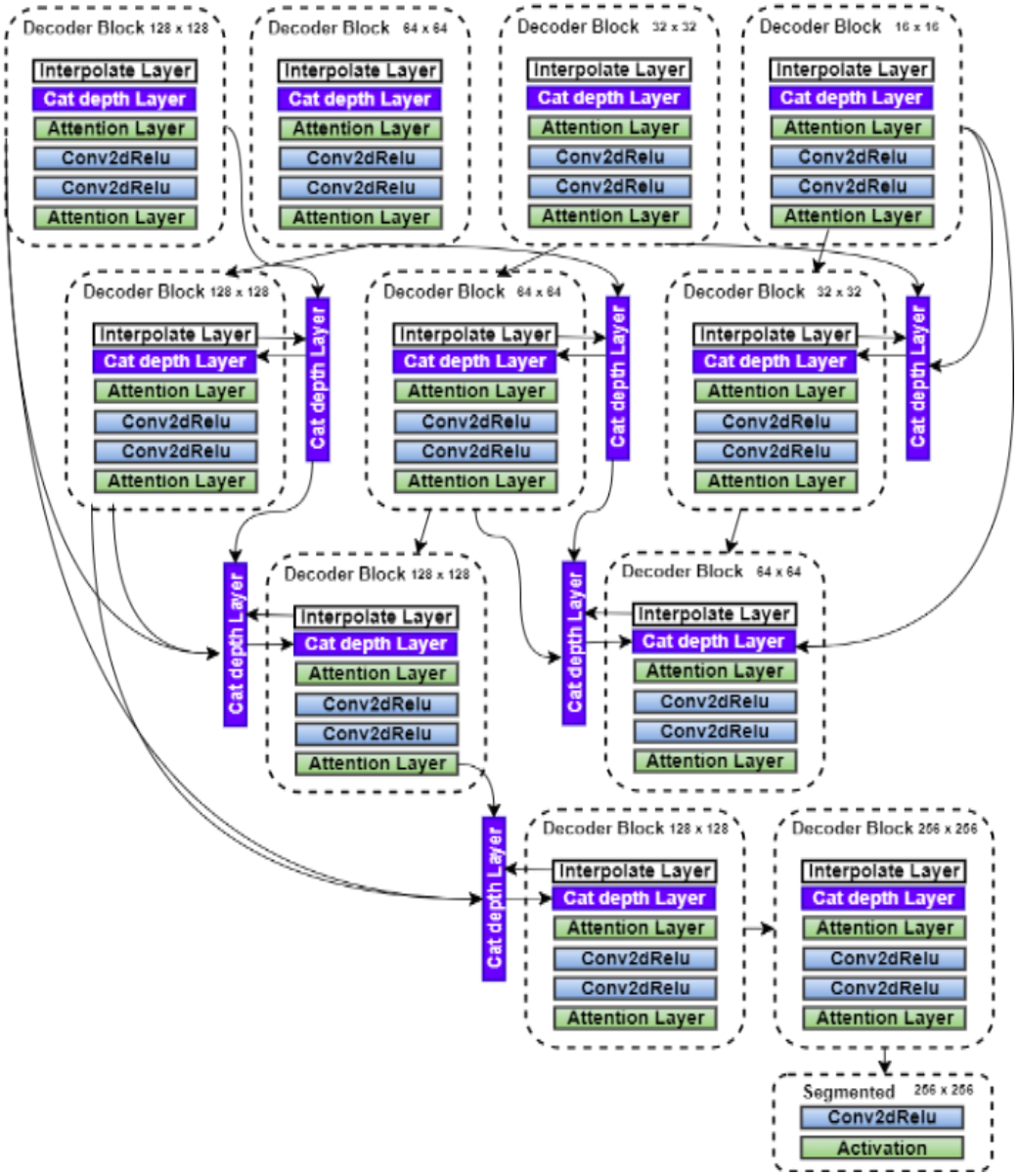

Figure. 3 combined model using UNET++ and merge decoded layers (FPN) aiming to capture more contextual information and improve segmentation performance

The GastroFPN contributes to an improvement in the feature extraction process. This rescale is intended to revolutionize the communication of information between the encoder and decoder sub-networks, with the potential to boost the network's ability to precisely capture and reproduce features in the data.

Objects of diverse nature might possess distinct dimensions, and the ability to identify objects at different scales requires the examination of characteristics at different levels of abstraction. FPN does this by constructing a feature pyramid via the use of a single convolutional neural network (CNN) backbone. The most important part of FPN networks is the pyramid pooling module, which takes features from different levels of the feature pyramid and puts them all together to make a single, consistent representation for each ROI. This aids in guaranteeing that the object detection method is resilient to fluctuations in object size.

3.1 Dataset

The EDD2020 dataset [28] comprises 386 static images obtained from various films. The training set consists of 160 hand-label annotations for nondysplastic Barrett's, 88 instances for suspected precancerous lesions, 74 for high-grade dysplasia, 53 for cancer, and 127 polyp masks. In all, there are 503 ground truth annotations. The Kvasir-SEG is a freely available dataset [29] consisting of images of gastrointestinal polyps and their related segmentation masks. A medical practitioner

manually annotated the images, and an expert gastroenterologist subsequently confirmed them. The dataset consists of 1000 photos of polyps, together with their related ground truth data from the Kvasir Dataset v2. The photos in Kvasir-SEG have resolutions ranging from 332x487 to 1920x1072 pixels. The CVC-ClinicDB dataset serves as a repository for frames captured during colonoscopy procedures [30]. These frames contain a large number of polyps. In addition to the frames, we provide accurate and reliable data for the polyps.

4. Result and discussion

We have used many quantitative evaluation measures to compare our models. We conducted an evaluation of our models using a separate test set, without informing the models about its existence, to accurately measure their performance. The comparison studies were carried out to achieve Objective, which included evaluating the performance without depending on official split-based testing. During the studies, the performance

of the model segmentation was evaluated using five different measures. These metrics were Intersection over Union (IoU), Dice Similarity Coefficient (DSC), recall, precision, and accuracy.

The GastroFPN model of Unet++ backbone uses first decoder block in 128 × 128, second decoder block 64 × 64, 32× 32, and 16 × 16 decoder blocks respectively. Feeding to FPN Merge Decoded Layers block. Our experiments for robustness analysis were carried out to show the best impact of our model, and they were divided into two parts: (i) experiments for external validation, and (ii) experiments for comparing the performance of model segmentation on the three different datasets.

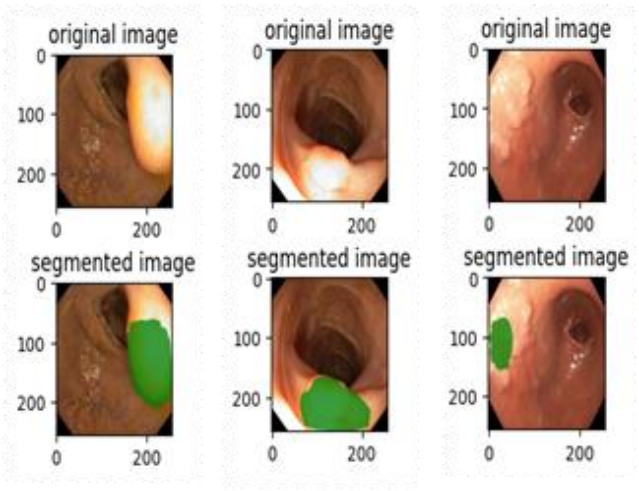

Figure. 5 FPN net segmentation model results

In order to carry out the experiments, a Windows 10 operating system was used, and PyTorch version 2.0.1 and Python version 3.9.16 were utilized. There was a computer that had a 13th generation Intel®CoreTM i7-13600KF central processing unit, 32 GB of random-access memory (RAM), and an NVIDIA GeForce RTX 3060 graphics processing unit (GPU) that had 16 GB of RAM. Model training lasted for more than one hundred epochs.

The results of our experiment output depict in Figure.4 that is describe the original image and segmented image.

During the analysis of segmentation performance using official split-based testing, the accuracy measure was used as the benchmark. Precision is not the optimal criterion for assessment when the primary concern is minimizing the occurrence of life-threatening ailments. Curiously, the runs that had the greatest macro averages for the F1-score also resulted in the best accuracy. Figure 4.

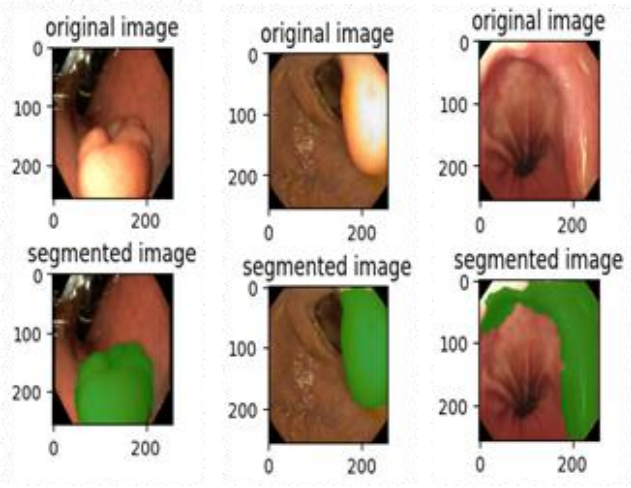

Figure. 4 Proposed model segmentation of our proposed model GastroFPN

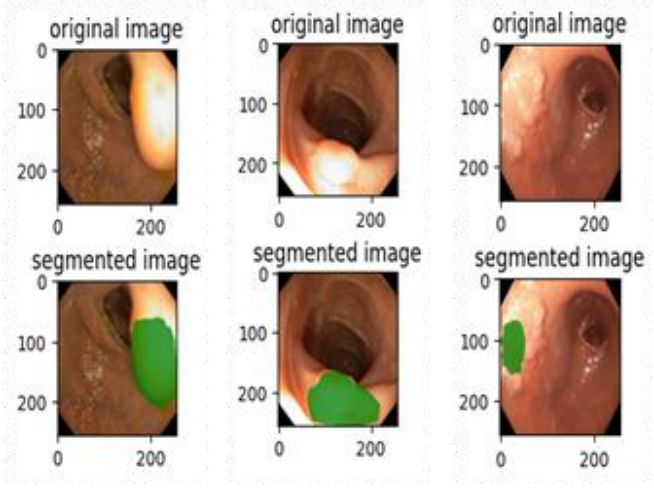

Figure. 6 DeeplabV3+ model segmentation results

Table 1. A comparison of the suggested model's segmentation performance with those of models that are considered to be state-of-the-art

Model	Loss IoU	F1	REC	ACC	DSC
UNet-ResNet34	56.6	63.8	86.8	93.8	63.7
DeepLabv3	54.3	72.4	87.8	94.2	72.4
DeepLabv3+	43.5	74.3	88	91	63
FPN	42.2	72.3	58.4	77	44.1
U-Net++	45.9	80.8	90.8	90	76.5
U-Net	41.5	77	84.5	87	43
DCSA-Unet	43.2	78	85.7	92	79.5
Proposed model	**38.7**	**86**	**87.8**	**96.8**	**86.6**

Illustrate the results of the GastroFPN model, which is surrounded by green, for the best detection of lesions. And comparing other models that train, our results show that segmentation is more accurate.

Figure 5 depicts the results of the FPN model. The model was trained separately without being combined with UNET++ to show the difference between them. Thus, the FPN is a feature extractor that, by taking only one-scale, arbitrary-sized images as an input, produces correspondingly proportioned maps at different levels, all of which are fully convolutional.

Fig. 6. shows the results for the DeeplabV3 model with enhancement, which is more accurate than the FPN model. One of the problems to overcome is a different way of segmenting objects according to different scales. So, modules are made that use atrous convolution in cascade or parallel, which can pick up on multiple scales by using different atrous rates. On top of these, the model used to add the Atrous Spatial Pyramid Pooling from DeepLabv2, which is also responsible for the global context and the deep learning feature encoding, which were again included, and they pushed the performance to the top. Table 1. describes the differences between various models of EDD2020 dataset. Our proposed model achieves the best results, as shown for 0.38 of the IoU, F1 at 0.85, and REC at 0.87. while accuracy of segment registered at 0.968 and Dice score at 0.866. Our study trained seven models to report the best results for segmentation lesions.

The suggested model surpasses all other models according to four out of five metrics, including IoU and DSC, which are the two metrics that are considered to be the most significant and frequently used in the area of medical picture segmentation. As can be observed, the proposed model exhibits superior performance.

Table 2 applies different models to Kvasir datasets to register the best network segment. The results focus on three metrics. DICE, IOU, and ACC.

Table 2 apply different model on Kvasir and CVC-ClinicDB datasets to show best network of segment. the results focus of three metrics DICE, IOU and ACC. again, the suggested model surpasses all other models in terms of four of the metrics (including IoU and DSC, which are the two metrics that are considered to be the most significant and commonly utilized in the area of medical image segmentation). A comparison was made between the suggested model and other models that have been offered in recent times, based on the outcomes that these models attained on the datasets, as published in the literature sources that corresponded to the proposed model.

Figure 7. shows the training model of Unet++ has an archive dice score near 0.76 at 70 epochs.

Table 2. Segmentation performance comparison of the proposed model for Kvasir SEG and CVC-ClinicDB dataset

Dataset	Methods	Accuracy	Dice-Score	IoU
Kvasir SEG	101	95.28	90.80	83.86
	102	95.87	92.30	85.90
	103	96.1	92.35	86.27
	Unet++ 104	93.54	82.44	75.35
	100	96.50	92.67	86.08
	Msnet	96.45	90.74	86.27
	GastroFPN	**96.62**	**92.54**	**87.57**
CVC-ClinicDB	102	98.32	91.20	90.89
	103	98.27	90.99	90.78
	104	93.47	63.88	71.43
	MSnet	98.30	86.25	80.46
	GastroFPN	**98.55**	**91.64**	**89.63**

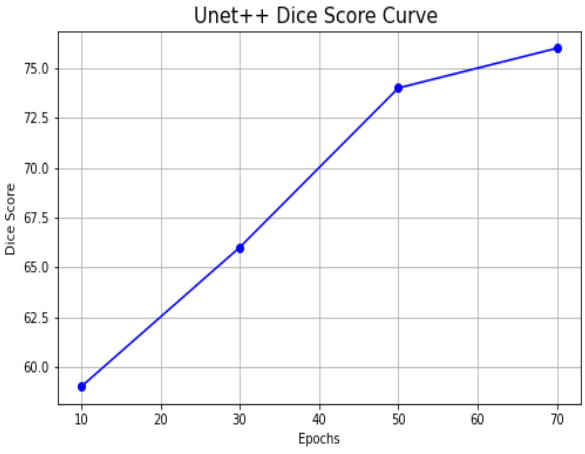

Figure. 7 training model of Unet++

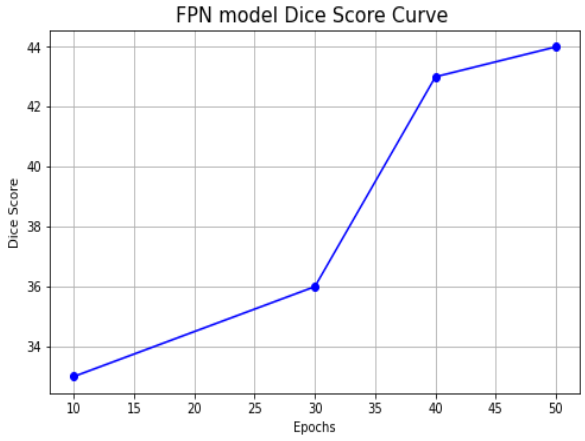

Figure. 8 training model of FPN

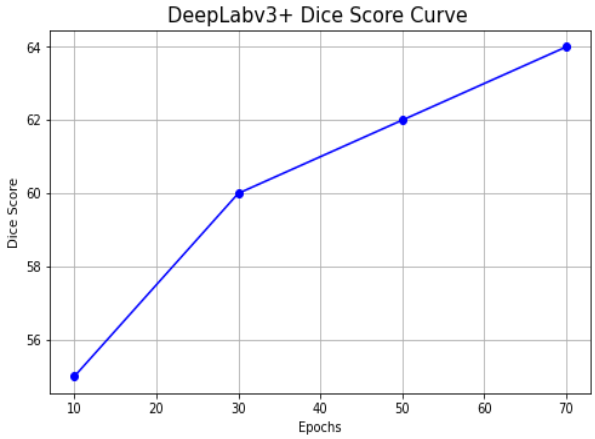

Figure. 9 The training model of deeplabv3+

After epoch 50, the model has archived more accuracy after the learning rate was reduced to 0.001.

Figure 8 depicts the low range of the dice score, which is near 0.44 at epoch 50.

Figure 9 illustrates the dice curve for best reaching 0.65 with epoch 70 for deeplabv3+.

5. Discussion and future work

The advantage of our model is that the main work of Unet++ is to be fast enough to decide the definite classification map, chosen from one of the segmentation branches under consideration. It is this that makes with several different types of semantic levels and can also produce feature maps with very high resolution. The deconvolutional work of the input feature UNet++ is used by UNet++. These refer to the multilayer of a neural network's inputs as an evenly laid tensor product. Our model passes through three stages of segmentation with FPN segment merge-decoded layers.

Figure 10 illustrate the comparative dice score of three comparative model applied on Kvasir-SEG dataset. It represents our proposed model GastroFPN and Unet++ and MSNet model. the

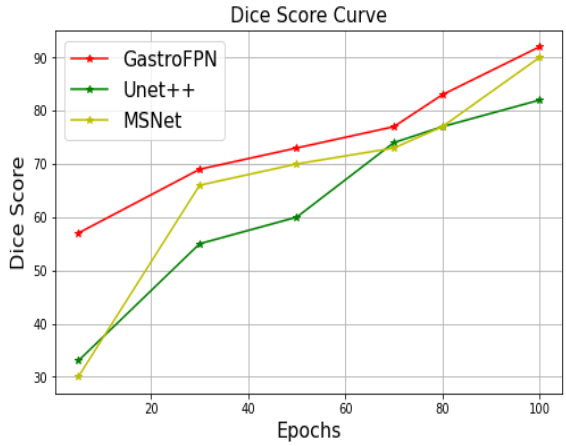

Figure. 10 the comparative Dice score of three model of the Kvasir-SEG dataset

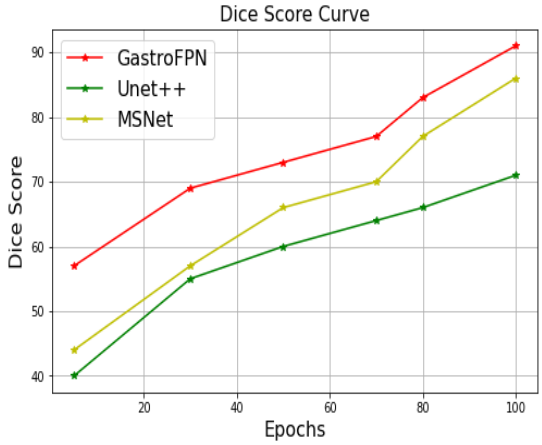

Figure. 11 the comparative Dice score of three model of the CVC-Clinic DB dataset

epochs at 100 is the best results of each model for GastroFPN 92.54, Unet++ 82.44 and MSnet 90.74.

Figure 11 shows the comparative dice score of three comparative model applied on CVC-ClinicDB dataset. It represents our proposed model GastroFPN and Unet++ and MSNet model. the training epochs at 100 is the best results of each model for GastroFPN 91.64, Unet++ 71.43 and MSNet 86.25.

6. Conclusion

This paper makes significant progress in the field of identifying gastrointestinal disorders by presenting an enhanced feature pyramid network decoder, which is an excellent segmentation model that utilizes advanced deep learning techniques. The suggested technique shows potential for improving clinical processes by facilitating fast and accurate detection of gastrointestinal issues, ultimately leading to improved patient outcomes and reduced healthcare expenses. In order to address the

challenges arising from insufficient annotated data, the proposed model employs transfer learning and data augmentation techniques to leverage pre-trained CNN models and enhance its ability to generalize. Rigorous testing on a varied dataset showcases that the suggested model surpasses existing segmentation methods in terms of both accuracy and computational efficiency. It attains cutting-edge outcomes. The results demonstrate the efficacy of our approach. The proposed model, when applied to the EDD2020 dataset, produced a segmentation accuracy of 96.8%, a dice-score of 86.6%, and an F1-score of 85.3%. We trained the suggested model using two separate datasets, specifically the CVC-ClinicDB and Kvasir-Seg databases. The CVC-ClinicDB obtained a Dice-Score of 91.64%, an Intersection over Union (IoU) of 84.63%, and an accuracy of 98.55%, as shown by the results metrics. The Kvasir-Seg dataset achieved a Dice score of 92.54%, an IoU score of 87.57%, and an accuracy score of 96.62%. In the future, we will investigate further to apply our proposed work to detecting other diseases.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

For this research work all authors have equally contributed to Conceptualization, methodology, validation, resources, writing—original draft preparation, writing—review and editing.

References

[1] S. A. Lee, H. C. Cho, and H. C. Cho, "A Novel Approach for Increased Convolutional Neural Network Performance in Gastric-Cancer Classification Using Endoscopic Images", *IEEE Access*, Vol. 9, pp. 51847–51854, 2021.

[2] W. Hu., C. Li, X. Li, "GasHisSDB: A new gastric histopathology image dataset for computer aided diagnosis of gastric cancer", *Comput Biol Med*, Vol. 142, p. 105207, 2022.

[3] M. Pennazio, R. Santucci, E. Rondonotti, "Outcome of Patients with Obscure Gastrointestinal Bleeding after Capsule Endoscopy: Report of 100 Consecutive Cases", *Gastroenterology*, Vol. 126, No. 3, pp. 643–653, 2004.

[4] A. Teramoto, T. Shibata, H. Yamada, Y. Hirooka, K. Saito, and H. Fujita, "Detection and Characterization of Gastric Cancer Using Cascade Deep Learning Model in Endoscopic Images", *Diagnostics*, Vol. 12, No. 8, 2022.

[5] M. AL-Mukhtar, A. S. Al-Zubaidi, and M. N. Albadri, "Predicting COVID-19 in Iraq using Frequent Weighting for Polynomial Regression in Optimization Curve Fitting", *Iraqi Journal of Science*, pp. 455–467, 2024.

[6] A. M. Dhayea, N. K. El Abbadi, and Z. G. Abdul Hasan, "Human Skin Detection and Segmentation Based on Convolutional Neural Networks", *Iraqi Journal of Science*, pp. 1102–1116, 2024.

[7] M. A. Kadhim and A. M. Radhi, "Machine Learning Prediction of Brain Stroke at an Early Stage", *Iraqi Journal of Science*, Vol. 64, No. 12, pp. 6596–6610, 2023.

[8] N. Sharma, S. Qupta, and D. Koundal, "U-Net Model with Transfer Learning Model as a Backbone for Segmentation of Gastrointestinal Tract", *Bioengineering*, Vol. 10, No. 1, 2023.

[9] S. Tuladhar, A. Alsadoon, P. W. C. Prasad, A. E. Ali, and A. Alrubaie, "A novel solution of deep learning for endoscopic ultrasound image segmentation: enhanced computer aided diagnosis of gastrointestinal stromal tumor", *Multimed Tools Appl*, Vol. 81, No. 17, pp. 23845–23865, 2022.

[10] A. Tabari, S. M. Chan, O. M. F. Omar, S. I. Iqbal, M. S. Gee, and D. Daye, "Role of Machine Learning in Precision Oncology: Applications in Gastrointestinal Cancers", *Cancers*, Vol. 15, No. 1. MDPI, 01, 2023.

[11] X. Zhao, L. Zhang, and H. Lu, "Automatic Polyp Segmentation via Multi-scale Subtraction Network", *Medical Image Computing and Computer Assisted Intervention – MICCA*, Vol 12901, 120–130, 2021.

[12] D. He, Y. Zhang, H. Huang, Y. Si, Z. Wang, and Y. Li, "Dual-branch hybrid network for lesion segmentation in gastric cancer images", *Sci Rep*, Vol. 13, No. 1, 2023.

[13] D. He, Y.Li, L. Chen, and X. Xiao, "Dual-guided network for endoscopic image segmentation with region and boundary cues", *Biomed Signal Process Control*, Vol. 91, 2024.

[14] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song, "Stepwise Feature Fusion: Local

Guides Global", *Medical Image Computing and Computer Assisted Intervention – MICCAI*, Vol 13433, 2022.

[15] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Attention Deeplabv3+: Multi-level Context Attention Mechanism for Skin Lesion Segmentation", In: *Proc. of Computer Vision – ECCV 2020 Workshops*, pp. 251–266, 2020.

[16] Q. Chang, D. Ahmad, J. Toth, R. Bascom, and W. E. Higgins, "ESFPNet: efficient deep learning architecture for real-time lesion segmentation in autofluorescence bronchoscopic video", In: *Proc. of SPIE*, p. 1246803, 2023.

[17] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 3–11, 2018.

[18] Q. Xu, Z. Ma, N. HE, and W. Duan, "DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation", *Comput Biol Med*, Vol. 154, p. 106626, 2023.

[19] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, 'Feature Pyramid Networks for Object Detection', In: *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117-2125, 2017.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 234–241, 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", In: *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[22] M. Gao, J. Chen, H. Mu, and D. Qi, "A Transfer Residual Neural Network Based on ResNet-34 for Detection of Wood Knot Defects", *Forests*, Vol. 12, No. 2, p. 212, 2021.

[23] E. S. Olivas, J. D. M. Guerrero, M. Martinez Sober, J. R. Magdalena Benedito, and A. J. Serrano López, "Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques", *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pp. 1–703, 2009.

[24] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 11045 LNCS, pp. 3–11, 2018.

[25] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp. 1137-1149, 1 2017.

[26] E. D. Cherpanath, P. R. Fathima Nasreen, K. Pradeep, M. Menon and V. S. Jayanthi, "Food Image Recognition and Calorie Prediction Using Faster R-CNN and Mask R-CNN", In: *Proc. of 2023 9th International Conference on Smart Computing and Communications (ICSCC)*, Kochi, Kerala, India, pp. 83-89, 2023.

[27] T. W. Cenggoro, A. H. Aslamiah, and A. Yunanto, "Feature Pyramid Networks for Crowd Counting", *Procedia Comput Sci*, Vol. 157, pp. 175–182, 2019.

[28] S. Ali, B. Braden, D. Lamarque, S. Realdon, A. Bailey, R. Cannizzaro, N. Ghatwary, J. Rittscher, C. Daul, J. East, "Endoscopy Disease Detection and Segmentation (EDD2020)", *IEEE Dataport*, Vol. 15, 2020.

[29] D. Jha, P. Smedsrud, M. Riegler, P. Halvorsen and T. Lange, "Kvasir-SEG: A Segmented Polyp Dataset", *arXiv:1911.07069*, 2019.

[30] K. Fitzgerald, J. Bernal, A. Histace, and B. J. Matuszewski, "Polyp Segmentation With the FCB-SwinV2 Transformer", *IEEE Access*, Vol. 12, pp. 38927-38943, 2024,