Gastro-Intestinal Tract Image Segmentation using Edge U-Net and U-Net VGG19

Sashank Talakola
Department of Computer Science and Engineering
National Institute of Technology Andhra Pradesh, India
stalakola@outlook.com

Madhusudhan Suryaprathap Reddy
Department of Computer Science and Engineering
National Institute of Technology Andhra Pradesh, India
madhuman1012@gmail.com

Rishi Nagam
Department of Computer Science and Engineering
National Institute of Technology Andhra Pradesh, India
nrishi2310@gmail.com

Srilatha Chebrolu
Department of Computer Science and Engineering
National Institute of Technology Andhra Pradesh, India
srilatha.chebrolu@nitandhra.ac.in

***Abstract*—In this work, an approach to tackle the problem of image segmentation on MRI scans of cancer patients using deep learning techniques has been proposed. Radio oncologists are challenged with a tedious task of segmenting the stomach and intestine regions manually. Based on the segmented regions, oncologists need to deliver high doses of X-Ray beams to treat the tumor cells while avoiding the healthy intestine and stomach regions. Using deep learning based segmentation techniques this process can be automated, which would lead to better treatment and save a lot of time. This work proposes a methodology using Edge U-Net and U-Net VGG19 to efficiently segment the large bowel, small bowel and stomach regions from the MRI scans of cancer patients. Experiments were conducted on the proposed methodology and results were compared with U-Net, Feature Pyramid Network and U-Net++ with variant encoders such as ResNet50, Xception, EfficientNet-B0, ResNeXT50. Comparison results show that the proposed methodology Edge U-Net and U-Net VGG19 has achieved the highest weighted Dice coefficient and 3D Hausdorff distance value of 0.86194.**

***Index Terms*—Computer Vision, Deep Learning, U-Net, MobileNets, Holistically-Nested Edge Detection, VGG19, Segmentation, Inception-V4**

I. Introduction

In 2019, there have been around 5 million new gastrointestinal cancer cases worldwide and 3.4 million deaths, as of 2018 each year nearly $783,000$ deaths occur globally [1]. Gastrointestinal (GI) cancer accounts for nearly 26% for cancer cases and 35% of cancer related deaths globally. Radio therapy can help the patient's treatment significantly, and the process of delivering X-ray beams is by a radio oncologist for a period of 10-15 minutes everyday for about 1-6 weeks. Their objective is to target the tumour cells with strong doses of X-ray radiation while avoiding the healthy intestinal and stomach cells. Due to the regular collection of data, which might aid in tracking the patient's development progress, the oncologist must manually segment the intestine and stomach areas from the MRI images. Based on these segmented regions the oncologists need to choose particular directions to deliver the X-Ray beams.

The process of manually segmenting every patient's MRI scans is a time consuming and laborious task. Multiple MRI scans of each patient are taken regularly and they need to be manually segmented. And it takes a lot of time to fully segment the intestine and stomach regions of a single patient. And since humans are also prone to error, an automated process can improve the patient's treatment significantly. By using deep learning based segmentation techniques [2] [3] the MRI scans could be segmented and the regions of interest are determined in relatively less period. Hence it could be possible to provide a much safer and effective treatment.

In [3], Suigu Tang et al. have proposed a methodology on simultaneously classifying and segmenting GI tract MRI scans using a transformer based approach. The multi-task network utilises features learned locally using CNN and features that are learned globally using transformer architecture to attain higher performance. CVC-ClinicDB dataset [4] has been used for training, with accuracy and Dice Similarity Coefficient (DSC) as evaluation metrics. The methodology has attained 96.94% and 77.76% on classification accuracy and DSC score respectively. A modified Mask-RCNN [5] architecture has been proposed by Mehshan Ahmed Khan et al. in the work [2]. They have proposed a methodology to segment and classify gastrointestinal images. Segmentation is achieved using Mask-RCNN with Feature Pyramid Network (FPN) and ResNet50 as its backbone. Classification is done using a ResNet101 followed by fully connected and average pooling layers with ReLU as its activation function. Evaluation is performed over a private dataset [6] containing data of 30 patients. For segmentation the methodology achieves a precision value of 0.6666 and a recall value of 0.6666, using 0.5 IoU as threshold, 0.6222 and 0.6445 using an IoU threshold of 0.75.

In this work we built a model to accurately segment the intestine and stomach regions using an ensemble model of Edge U-Net and U-Net with VGG19 as backbone along with prior classification done using Inception-V4. The model takes MRI scans of patients as inputs and outputs a segmentation mask of the intestine and stomach regions. The MRI scans

are of patients who underwent 1 to 5 MRI scans on various days during their treatment period. The goal of the model is to produce a pixelated classification among the three classes - large bowel, small bowel and stomach.

II. Related Convolution Neural Networks Architectures

Convolution Neural Networks (CNN) architectures that are used in this work are i) Feature Pyramid Network (FPN), ii) Visual Geometry Group (VGG) and iii) U-Net. This section describes each of these architectures.

A. Feature Pyramid Network

A FPN [7] is a CNN architecture created to solve the object detection and segmentation [8] task. FPN creates a multiscale feature representation of an image by combining features from different levels of a CNN. It is made up of backbone networks such as SeResNet50, ResNet, Inception-V4, or Inception-V4. Using the input image, FPN constructs a pyramid of feature maps, with each level denoting a feature map of a different size. This FPN architecture consists of a top-down route and a bottom-up route. Using CNN layers, the bottom-up method extracts feature maps of different sizes. The bottom-up pathway's outputs are upsampled and combined with the top-down pathway's higher scale feature maps to create the pyramid of features. This pyramid is then used as input to the segmentation algorithm thus, the pyramid of features extracted are processed through CNN layers to perform segmentation.

B. Visual Geometry Group

VGG [9] is a deep CNN architecture design for image classification. VGG is known for its model simplicity and uniformity. VGG has two variants VGG16 and VGG19. Depending on the variant used it has 16 or 19 layers. The VGG19 model has sixteen convolutional layers and three fully connected layers. Though the architecture was designed for image classification tasks, it has been adopted for image segmentation [10]. VGG has been used to modify the encoder network of the segmentation architecture, thus creating reliable feature maps that can significantly enhance the segmentation model's performance.

C. U-Net

U-Net [11] is a deep CNN model introduced for image segmentation. The underlying structure of this CNN architecture is the Fully Convolutional Network (FCN) [12] architecture. U-Net is composed of an expanding path and a contracting path that are linked together by skip connections. The input image is reduced in size and its features are extracted. The output of the expanding path is improved in terms of spatial resolution by upsampling the feature maps created by the contracting path using a sequence of transpose convolutional layers also known as deconvolutional layers. Skip connections link the encoding path and decoding path of the model. These enable the model to use both top-level and bottom-level information from the input image to provide more precise segmentation predictions.

1) Encoding path: Encoding path refers to the contracting route of the U-Net architecture. The objective of the encoding path is to reduce the input image and extract information from it. The max pooling layers build feature maps from the convolutional layers by downsampling them by selecting the maximum value within the window size. The convolutional layers in the encoding stage apply a collection of filters to the input image in a sliding window approach. As it moves through the encoding phase, these procedures enable the model to extract even more intricate and abstract information from the input image.

2) Decoding path: The decoding path of the U-Net architecture is made up of a number of transpose convolutional layers. The upsampling of the feature maps created by the contracting path and the improvement of the output's spatial resolution are the goals of the decoding phase. In contrast to regular convolutional layers, which shrink the input image, transpose convolutional layers expand the input image. They are used in the decoding phase of the U-Net architecture to apply a collection of filters to the input feature maps in a sliding window form. The model is able to regain the spatial resolution that was lost during the network's encoding phase because of the upsampling process.

3) Bottleneck layer: The bottleneck layer in the U-Net architecture is the part where the feature maps are reduced to lower size and then subsequently increased to original size. It acts as a bridge between the encoding path and the decoding path. It consists of a series of convolutional layers with a large number of filters and smaller kernel size. These layers further decrease the spatial dimensions along with increasing the number of channels which allow the network to capture high-level features effectively.

4) U-Net++: U-Net++ [13] is a nested U-Net architecture introduced by Zongwei Zhou et al. for the objective of image segmentation. U-Net has the following limitations i) decrease in resolution caused by max-pooling and ii)the process of decoding results in the loss of low-level feature data. These limitations have been addressed by U-Net++. The U-Net++ has extra "nested" U-Net blocks in the encoding path. In each "nested" block, the input image is first downsampled. The resulting feature map is then used to perform another U-Net block. Then, a new input for the following "nested" block is created by appending the output of this "nested" block with the previous block output in the encoding route. This nested architecture allows the model to capture multi-scale context information and improve feature representation at different levels of abstraction. Moreover, U-Net++ utilizes convolutional layers rather than max-pooling layers in the encoding path to preserve spatial information and prevent resolution loss.

III. Proposed Methodology

In this section, the proposed methodology for performing the task of image segmentation of large bowel, small bowel and stomach on the MRI scans of GI tract is described. The task of image segmentation starts with data preprocessing.

A. Data Preprocessing

The task of data preprocessing is as shown in Figure 1. Two different preprocessing pipelines are used for data preprocessing. In the first pipeline the input images are resized to 320x384. In the second pipeline, 2.5D Image preprocessing is applied on the input image. Even though 2D training can be performed, additional depth data from the MRI slices are used to produce data of 3D volumes [14], [15]. To create a 3D volume, 3 consecutive MRI slices are stacked. Methodologies proposed in medical image segmentation using 2.5D [16], [17] image processing have shown improvement in performance. Both these pipelines pass through image augmentation techniques [18]. The image augmentation techniques include i) Horizontal flip ii) Image rotation iii) Elastic transform iv) Coarse dropout. Grayscale images are generated from the first pipeline and 2.5D images from the second pipeline.

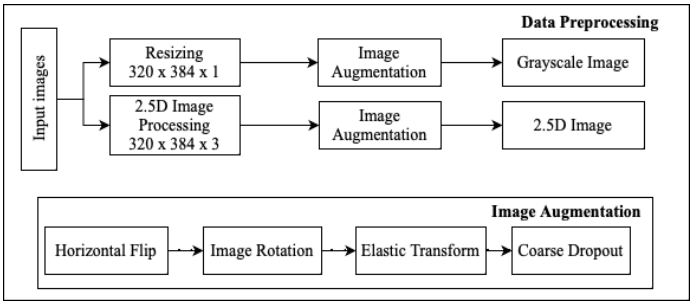


Figure 1: Data Preprocessing

B. Proposed ensemble architecture for Image segmentation

The proposed ensemble architecture for image segmentation is as shown in Figure 2. After preprocessing, the input images are given as input to an Inception-V4 [19] binary classification model. This model predicts the existence of healthy organs i.e., large bowel, small bowel and stomach in the input slices. Segmentation of the image can be omitted if there are no healthy organs present in it. In this case an empty mask is generated as a segmentation mask. Otherwise, the input images are processed in two pipelines. In the first pipeline, 2.5D images are given as input to a U-Net with VGG19 as backbone network. In the second pipeline grayscale images are given as input to Edge U-Net. The final predicted segmentation mask is the mean computed from the segmentation masks predicted from both the pipelines.

The following subsections describe Inception-V4 classification model, Edge U-Net and U-Net VGG19 segmentation models which are been used in the proposed architecture for GI tract image segmentation.

1) Inception-V4: Inception-V4 is a state-of-the-art performance classification algorithm developed by researchers of Google. The fundamental idea behind Inception-V4 is to reduce computation cost while attaining a higher accuracy, which was accomplished by replacing larger CNN's with sequences of smaller CNN's. The architecture is a series of inception blocks [20], with each of them containing CNN's with

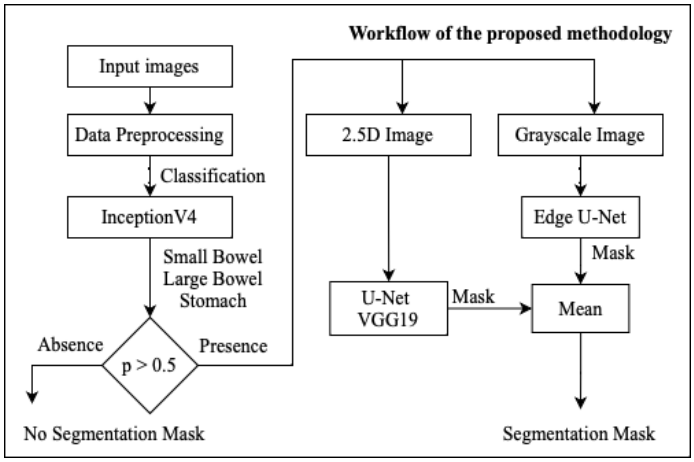


Figure 2: Workflow of the proposed methodology

1x1, 3x3 and 5x5 filter sizes and max-pooling layer with 2x2 filter size. Inception-V4 integrates with batch normalization [21] and residual connections [22] to improve performance.

2) Edge U-Net: In the Edge U-Net, convolution blocks of the encoder part of U-Net are replaced with MBconv blocks [23] and extra skip connections are added that capture edge data from the input images. The architecture of Edge U-Net is as shown in Figure 3. Edge detection is done using Holistically-Nested Edge Detection (HED) [24]. Extra skip connection is computed by applying Hadamard product [25] across every channel of the encoder output at corresponding levels.

HED is a deep learning based edge detection architecture. HED architecture is based on VGG without the final fully connected layer used for classification. HED contains side-outputs produced at intermediate layers which are used to produce edge maps at different scales. The side-outputs along with the final edge feature map are fused with each other after upscaling to form the final output. Further training the network along with side-outputs helps the model to produce much accurate edge maps at various scales. HED attains Optimal Dataset Scale (ODS) F-Score of 0.746 on NYU Depth dataset [26] and ODS F-Score 0.782 on BSD500 dataset [27], with an improved speed of inference over the other CNN based edge detection architectures such as, DeepContour [28], CSCNN [29], DeepEdge [30].

3) U-Net VGG19: The U-Net with VGG19 as encoder is an image segmentation model that combines the U-Net architecture and the VGG19 network. The VGG19 network serves as an encoder and the segmentation process is performed using the U-Net. In the U-Net, the decoder component creates the segmentation mask and The encoder component extracts features from the input image. The convolutional layers of the VGG19 CNN network make up the U-Net encoder. The decoder component is made up of a number of convolutional and up-convolutional layers that are connected to the decoder via skip connection from the encoder. The skip connections enable the decoder to utilise the encoder's high-resolution

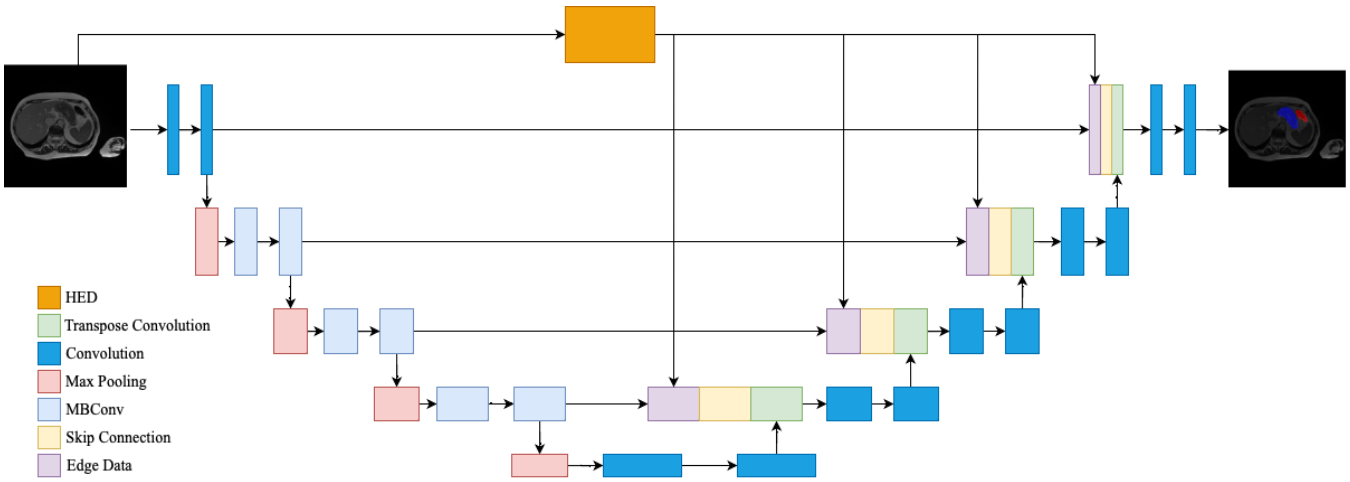


Figure 3: Edge U-Net architecture

characteristics for more precise segmentation.

IV. Experimental results and analysis

This section displays the experimental results from work done using the GPU P100 accelerator on the Kaggle platform. This accelerator has GPU memory of 16GB and provides 1.6x more GFLOPs than K80. Experiments were conducted using the proposed methodology on the dataset described as follows

A. Dataset

In this work, the proposed methodology for GI tract Image Segmentation has been analysed using the dataset [31]. This dataset consists of anonymized MRI scans of patients who received care at the Carbone Cancer Center at the University of Madison, Wisconsin. This dataset is a collection of MRI scans of Gastro-Intestinal Tract captured during radiation treatment. This dataset has 1-5 MRI scans captured on different days of 85 cancer patients undergoing radiation treatment for training purposes. MRI scans taken on a day are represented by a set of slices. Thus, each patient will have multiple sets of slices. This dataset contains a total of $38,496$ slices of size 2.47 GB. There are around 50 patient MRI scans for testing purposes. Each slice is a 16-bit gray-scale image. Images are of varying sizes from 234x234 to 310x360. If the training image contains a large bowel, small bowel and stomach, then segmentation masks for each class are provided individually as Run Length Encoding (RLE) masks. Otherwise, no mask is provided for the image. The proposed methodology predicts segmentation mask for the test data, if the slice consists either of large bowel, small bowel or stomach. Otherwise an empty mask will be predicted.

B. Classification Evaluation Metrics

Accuracy, recall, precision, and F1-score are used in this work as evaluation measures to assess the performance of different classification models and are defined as follows:

$$Accuracy = \frac{TH + TNH}{TH + FH + TNH + FNH} \quad (1)$$

Classification Model	Accuracy	Precision	Recall	F1-Score
SEResNeXT50	0.9842	0.9232	0.9494	0.9361
VGG19	0.9908	0.9405	0.9514	0.9459
Xception41	**0.9939**	0.5193	0.9587	0.6737
VIT Base Patch-16-384	0.9861	0.9367	0.948	0.9423
ResNet50	0.9873	0.9384	0.943	0.9407
SEResNet50	0.9835	0.9339	0.9415	0.9377
DenseNet121	0.9836	0.9183	0.9478	0.9328
ConvNeXT Base	0.9932	0.9495	0.9487	0.9491
InceptionV4	0.9913	**0.9892**	**0.9587**	**0.9737**

Table I: Accuracy, Precision, Recall and F1-Score obtained using various classification models on GI tract image segmentation dataset

$$Recall\ (r) = \frac{TH}{TH + FNH} \quad (2)$$

$$Precision\ (p) = \frac{TH}{TH + FH} \quad (3)$$

$$F1 - Score = 2 \times \frac{r \times p}{r + p} \quad (4)$$

where True Positives (TH) are a number of correctly classified images with healthy organs. True Negatives (TNH) are a number of correctly classified images no healthy organs. False Positives (FH) are a number of incorrectly classified images healthy organs. False Negatives (FNH) are a number of incorrectly classified images no healthy organs.

C. Segmentation Evaluation Metrics

In this work, the performance of the proposed methodology is assessed using the Dice Coefficient (DC) [32] and 3D Hausdorff distance. A weightage of 0.4 is given for DC and 0.6 is given for 3D Hausdorff Distance.

1) Dice Coefficient: The DC is used to determine how similar original mask and predicted mask are to one another. It ranges from 0-1, where 0 denoting significant differences and 1 denoting total overlap between the two masks. Higher DC values represent the higher overlap of the two masks.

Network Model	Encoder	Validation Data		Test Data
		Dice Coefficient	IoU Coefficient	Score
U-Net++ (320x320 grayscale)	ResNet50	0.9022	0.8706	0.7899
	Inception-V4	0.9112	0.8809	0.80717
	SEResNet50	0.9077	0.8769	0.80095
	ResNeXT50	0.8955	0.8631	0.78462
	Xception	0.9109	0.8806	0.79711
	EfficientNet-B0	0.8128	0.7753	0.71372
FPN (320x384 grayscale)	ResNet50	0.8432	0.805	0.71599
	Inception-V4	0.849	0.8107	0.71002
	SEResNet50	0.8846	0.8495	0.75076
	ResNeXT50	0.7983	0.7608	0.68941
	Xception	0.8824	0.847	0.73761
	EfficientNet-B0	0.7696	0.7331	0.68033
U-Net (320x384 grayscale)	ResNet50	0.8567	0.82	0.7473
	Inception-V4	0.8899	0.8555	0.77229
	SEResNet50	0.885	0.8502	0.76943
	ResNeXT50	0.8823	0.8476	0.7628
	Xception	0.9071	0.8749	0.77563
	DenseNet161	0.9114	0.8801	0.78925
	VGG16	0.8552	0.8183	0.7464
U-Net (320x384 2.5D)	Xception	0.923	0.8928	0.80138
	DenseNet201	0.912	0.8806	0.7961
	VGG19	0.941	0.9143	0.84984
Edge U-Net (320x384 grayscale)	-	0.9308	0.9051	0.84046
Proposed methodology (320x384 grayscale and 2.5D)	-	**0.9504**	**0.9259**	**0.86194**

Table II: Dice Coefficient and IoU coefficient values obtained on validation data and Score obtained using Equation 9 on the test data

$$DC(PM, OM) = \frac{2|PM \cap OM|}{|PM| + |OM|} \tag{5}$$

where PM is the predicted mask and OM is the original mask, $|PM|$ and $|OM|$ are their respective pixel count, and $|PM \cap OM|$ is the number of overlapping pixels.

2) 3D Hausdorff distance: A metric for comparing the similarity of two 3D segmentation masks is the 3D Hausdorff distance [33]. The Hausdorff distance is the greatest separation between any two pixels on one mask and their nearest counterparts on the other mask. To generate a bounded 0 to 1 3D Hausdorff distance, the picture area is normalised between the expected and actual pixel positions. Value 0 indicates high dissimilarity while 1 indicates high similarity.

$$HD1(PM, OM) = \max_{pm \in PM} \min_{om \in OM} |pm - om| \tag{6}$$

$$HD2(PM, OM) = \max_{om \in OM} \min_{pm \in PM} |pm - om| \tag{7}$$

$$HD(PM, OM) = \max(HD1, HD2) \tag{8}$$

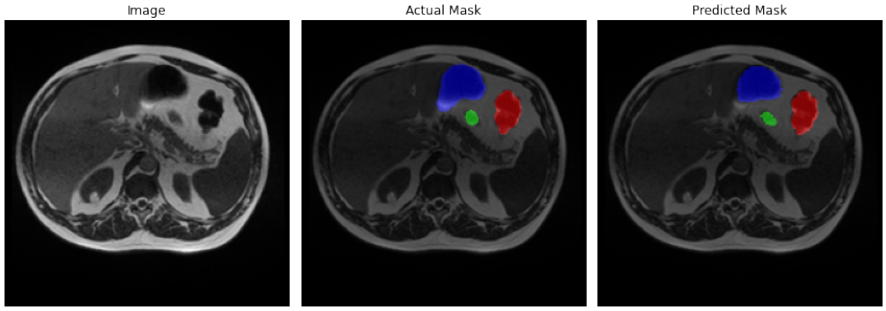
where PM is the predicted mask and OM is the original mask, pm represents a pixel from the predicted mask PM and om represents a pixel from the ground truth mask OM. The Euclidean distance between two pixels pm and om is represented by $|pm - om|$.

D. Experiment Results

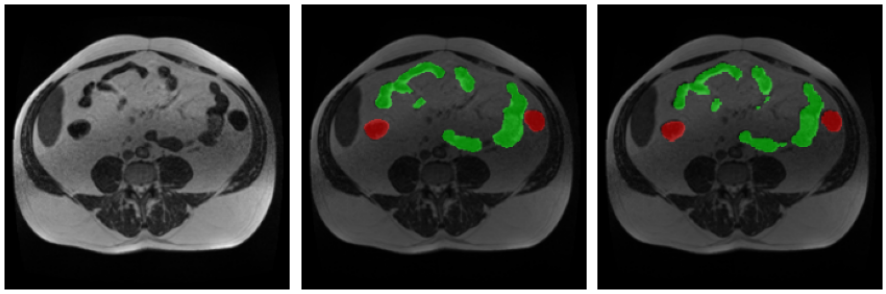
The images of the dataset are resized to 320x384 and are given as input to the classifier. Table I shows the accuracy, recall, precision and F1-score values obtained on the GI tract image segmentation dataset with various classification models i) SEResNeXT50 [34] ii) VGG19 iii) Xception41 [35] iv) VIT Base Patch-16-384 [36] v) ResNet50 vi) SEResNet50 vii) DenseNet121 [37] viii) ConvNeXT Base [38] ix) Inception-V4. Except for VIT Base Patch-16-384 every other classifier is trained on 320x384 images. VIT Base Patch-16-384 is trained on 384x384. Among all the classification models, Inception-V4 has obtained highest precision, recall and F1-score values for the given dataset. Inception-V4 has been chosen as the classifier in the proposed methodology. Table II shows the values of DC and IoU coefficient obtained on validation data and Score on the test data where Score is defined as follows:

$$\text{Score} = 0.4 \times \text{Dice Coefficient} + 0.6 \times \text{3D Hausdorff distance} \tag{9}$$

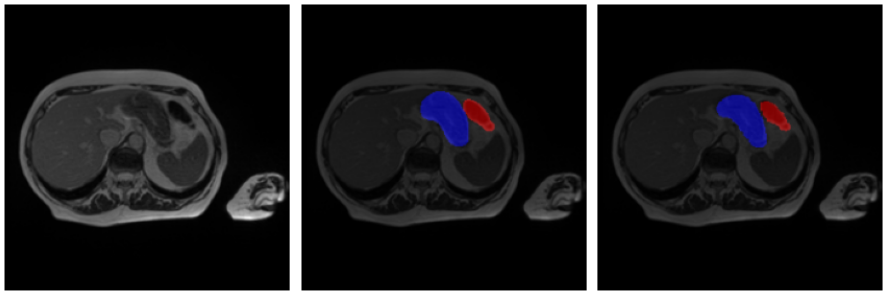
Experiments are conducted on various network models i) U-Net++ ii) FPN iii) U-Net iv) Edge U-Net v) Proposed methodology with various encoders i) ResNet50 ii) Inception-V4 iii) SEResNet50 iv) ResNeXT50 v) Xception vi) EfficientNet B0 vii) DenseNet viii) VGG19. The proposed methodology has obtained the highest DC 0.9504, IoU coefficient 0.9259



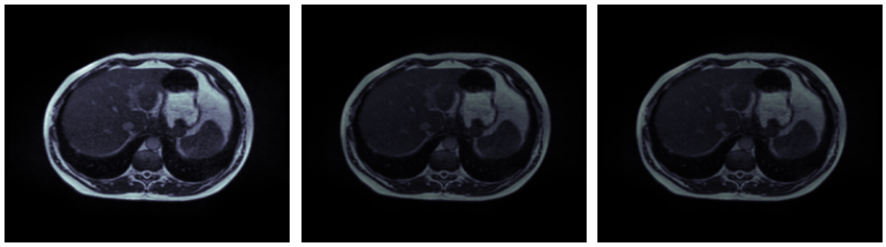
(a) Case 1



(b) Case 2



(c) Case 3



(d) Case 4

Figure 4: Original MRI slice, ground truth mask and predicted mask respectively for various cases

on validation data and Score value of 0.86194 on test data. Figure 4. shows four cases of the original MRI slice, ground truth segmentation mask and predicted segmentation mask obtained from the proposed methodology on the GI tract image segmentation dataset. In the figure red represents the large bowel, green represents the small bowel and blue represents the stomach. Case 1 of Figure 4 contains three organs. Original mask and predicted mask are showing the presence of all the three organs. Case 2 contains two organs i.e., large bowel and small bowel. Original mask and predicted mask shows their presence. Case 3 contains two organs i.e., small bowel and stomach. Original mask and predicted mask shows their presence. Case 4 contains no organ thus both original mask and predicted mask are empty. The obtained higher values of DC and 3D Hausdorff distance on test data for GI tract image segmentation data shows that proposed methodology is suitable for the problem of GI tract image segmentation.

V. Conclusion

The proposed methodology Edge U-Net and U-Net VGG19 performs the task of segmenting the large bowel, small bowel and stomach from the GI tract image segmentation

dataset. This proposed model has a Inception-V4 classifier prior to the segmentation model which has been evaluated and compared with other models such as ResNet50, Xception41, DenseNet121, ConvNeXT base. Inception-V4 has the highest precision 0.9892, recall 0.9587 and F1-Score 0.9737 values. The proposed segmentation model has been evaluated by weighted DC and 3D Hausdorff distance metrics. It has achieved high scores of DC 0.9504 and 3D Hausdorff distance 0.9259 on the validation data and 0.86194 Score on the test data. This model was compared with other network models such as FPN, U-Net++, U-Net with their encoder variants like ResNet50, Inception-V4, SEResNet50, Xception, DenseNet201. With this work, radio oncologists can reduce the time constraint and efficiently segment the large bowel, small bowel and stomach of the cancer patients.

REFERENCES

[1] P. Rawla and A. Barsouk, "Epidemiology of gastric cancer: global trends, risk factors and prevention," *Prz Gastroenterol*, vol. 14, no. 1, pp. 26–38, Nov. 2018.

[2] M. A. Khan, M. A. Khan, F. Ahmed, M. Mittal, L. M. Goyal, D. Jude Hemanth, and S. C. Satapathy, "Gastrointestinal diseases segmentation and classification based on duo-deep architectures," *Pattern Recognition Letters*, vol. 131, pp. 193–204, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786551930399X

[3] S. Tang, X. Yu, C. F. Cheang, Y. Liang, P. Zhao, H. H. Yu, and I. C. Choi, "Transformer-based multi-task learning for classification and segmentation of gastrointestinal tract endoscopic images," *Computers in Biology and Medicine*, vol. 157, p. 106723, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482523001889

[4] S. F. J. F.-E. G. G. D. R. C. . V. F. Bernal, J., "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *arXiv preprint arXiv:1703.06870*, 2018.

[6] M. Sharif, M. Attique Khan, M. Rashid, M. Yasmin, F. Afza, and U. J. Tanik, "Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 33, no. 4, pp. 577–599, Jul. 2021.

[7] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," *arXiv preprint arXiv:1612.03144*, 2017.

[8] Seferbekov, Selim and Iglovikov, Vladimir and Buslaev, Alexander and Shvets, Alexey, "Feature pyramid network for multi-class land segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 272–2723.

[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2015.

[10] M. Fradi, E. hadi Zahzah, and M. Machhout, "Real-time application based CNN architecture for automatic USCT bone image segmentation," *Biomedical Signal Processing and Control*, vol. 71, p. 103123, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421007205

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv preprint arXiv:1505.04597*, 2015.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *arXiv preprint arXiv:1411.4038*, 2015.

[13] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," *arXiv preprint arXiv:1807.10165*, 2018.

[14] H. Haque, M. Hashimoto, N. Uetake, and M. Jinzaki, "Semantic Segmentation of Thigh Muscle using 2.5D Deep Learning Network Trained with Limited Datasets," *arXiv preprint arXiv:1911.09249*, 2019.

[15] Y. Ou, Y. Yuan, X. Huang, K. Wong, J. Volpi, J. Z. Wang, and S. T. C. Wong, "LambdaUNet: 2.5D Stroke Lesion Segmentation of Diffusion-weighted MR Images," *arXiv preprint arXiv:2104.13917*, 2021.

[16] J. Li, G. Liao, W. Sun, J. Sun, T. Sheng, K. Zhu, K. M. von Deneen, and Y. Zhang, "A 2.5D semantic segmentation of the pancreas using attention guided dual context embedded U-Net," *Neurocomputing*, vol. 480, pp. 14–26, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222000650

[17] L. B. da Cruz, D. A. D. Júnior, J. O. B. Diniz, A. C. Silva, J. D. S. de Almeida, A. C. de Paiva, and M. Gattass, "Kidney tumor segmentation from computed tomography images using DeepLabv3+ 2.5D model," *Expert Systems with Applications*, vol. 192, p. 116270, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421015797

[18] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," *arXiv preprint arXiv:1712.04621*, 2017.

[19] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *arXiv preprint arXiv:1602.07261*, 2016.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," *arXiv preprint arXiv:1409.4842*, 2014.

[21] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv preprint arXiv:1502.03167*, 2015.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017.

[24] S. Xie and Z. Tu, "Holistically-Nested Edge Detection," *arXiv preprint arXiv:1504.06375*, 2015.

[25] S. Kazenas, "The Hadamard product and recursively defined sequences," *arXiv preprint arXiv:1911.01175*, 2019.

[26] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images." *ECCV (5)*, vol. 7576, pp. 746–760, 2012.

[27] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2010.

[28] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3982–3991.

[29] J.-J. Hwang and T.-L. Liu, "Pixel-wise Deep Learning for Contour Detection," *arXiv preprint arXiv:1504.01989*, 2015.

[30] G. Bertasius, J. Shi, and L. Torresani, "DeepEdge: A Multi-Scale Bifurcated Deep Network for Top-Down Contour Detection," *arXiv preprint arXiv:1412.1123*, 2015.

[31] P. C. P. Y.-S. L. L. happyharrycn, Maggie, "UW-Madison GI Tract Image Segmentation ," 2022. [Online]. Available: https://kaggle.com/competitions/uw-madison-gi-tract-image-segmentation

[32] R. R. Shamir, Y. Duchin, J. Kim, G. Sapiro, and N. Harel, "Continuous Dice Coefficient: a Method for Evaluating Probabilistic Segmentations," *arXiv preprint arXiv:1906.11031*, 2019.

[33] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on medical imaging*, vol. 39, no. 2, pp. 499–513, 2019.

[34] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *arXiv preprint arXiv:1709.01507*, 2019.

[35] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *arXiv preprint arXiv:1610.02357*, 2017.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2021.

[37] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *arXiv preprint arXiv:1608.06993*, 2018.

[38] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *arXiv preprint arXiv:2201.03545*, 2022.