TransUNet for Precise and Robust GI tract Segmentation in MRI Images

Bindu Madhavi Tummala *, Rishitha Jaladi, Dasari Chinna Veeraiah, and Aruna Kumari Peruri

Department of Computer Science and Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India

Email: bindumadhavi@vrsiddhartha.ac.in (B.M.T.); rishitha.jaladi03@gmail.com (R.J.); veeranani222@gmail.com (D.C.V); arunakumarip1003@gmail.com (A.K.P.)

*Corresponding author

***Abstract*—Medical image processing is revolutionizing Gastrointestinal (GI) cancer radiation therapy by enabling precise targeting of high X-ray beam dosages to the tumors while protecting the surrounding healthy organs. However, the manual segmentation of healthy organs at risk from MR-Linac images is a time-consuming process that can delay treatment and increase patient suffering. Deep learning-based medical image processing algorithms have the potential to automate the segmentation of organs at risk, thereby improving the accuracy and efficiency of GI cancer radiation therapy. Every day, the MR-Linac, a cutting-edge MRI technology, tracks the ever-changing positions of tumors. This advanced tool empowers medical professionals to fine-tune cancer treatments with remarkable precision. This study presents a TransUNet model, a combination of transformer architecture with Convolutional Neural Networks (CNNs). TransUNet achieves remarkable results in segmenting and labeling different regions within images by integrating the spatial comprehension of CNNs with the self-attention mechanisms of transformers. Our research compares the TransUNet model with various combinations of loss functions. The model outperforms with Dice+BCE loss function.**

***Keywords*—magnetic resonance imaging, TransUNet, gastrointestinal tract segmentation**

I. Introduction

Millions of individuals are impacted by Gastrointestinal (GI) disorders each year, which constitute a major global health concern. As per the World Health Organization (WHO), gastrointestinal disorders are accountable for over 1.5 million deaths annually and roughly 9% of the worldwide disease burden. The diagnosis and management of these conditions depend on the precise and trustworthy segmentation of the Gastrointestinal (GI) organs, including the stomach, large bowel, and small bowel. Medical professionals have manually segmented the GI system, which is a laborious procedure. Segmentation techniques that are automated are required to overcome these obstacles. Medical imaging is crucial in today's healthcare for accurate diagnosis and treatment planning of a variety of diseases and conditions [1]. Magnetic Resonance Imaging (MRI) is one of the many imaging modalities that stands out for its non-invasive capacity to offer comprehensive anatomical information Within the field of Gastroenterology, Magnetic Resonance Imaging (MRI) is a useful diagnostic tool that facilitates the visualization of the Gastrointestinal (GI) tract and the assessment of various diseases, including tumors and inflammatory bowel disease. For accurate diagnosis and treatment planning and segmentation, the process of separating and dividing anatomical structures or regions of interest within medical images is essential. The job of accurately segmenting the GI tract from MRI scans is difficult because of differences in picture quality, intricate anatomical structures, and a variety of diseases that might mask borders. Conventional machine learning approaches and handmade features are frequently used in segmentation techniques, which may not be able to adequately capture the complex spatial dependencies and minute details seen within the GI tract. However, by utilizing self-attention mechanisms to successfully capture long-range dependencies and contextual information, recent advances in deep learning particularly the triumph of transformer-based models have revolutionized several computer vision tasks.

The main objective of this research is to explore the accuracy and reliability of GI tract segmentation in MRI images using transformer-based designs, which are widely recognized for their effectiveness in computer vision and natural language processing applications. We want to increase the precision, resilience, and effectiveness of GI tract segmentation in comparison to conventional techniques by utilizing the power of transformers, which are excellent at modeling intricate relationships in data sequences. The use of transformer-based models for GI tract segmentation in MRI images is thoroughly examined in this work. To address the difficulties caused by the variation in GI tract appearance, we study how transformer topologies can be modified and improved to ensure correct segmentation. Additionally, our goal is to show off these models' potential in clinical settings by offering dependable and accurate segmentation findings that support medical professionals in making diagnosis and treatment choices.

Manuscript received January 10, 2024; revised February 26, 2024; accepted March 20, 2024; published September 12, 2024

Our goal is to demonstrate the efficacy and potential of transformers in advancing medical image analysis, particularly in the context of GI tract segmentation from MRI images, through experimental evaluations and comparative analyses with cutting-edge segmentation techniques. Transformers have transformed medical picture segmentation by using self-attention processes to capture long-range dependencies, resulting in increased accuracy and efficiency in tasks such as organ segmentation and tumor identification. By fusing the advantages of both techniques, the integration of transformers with conventional methods improves their performance and makes them useful instruments for medical picture processing and diagnosis. Transformers can improve U-Net [2], which is well-known for its efficiency in segmenting structures with distinct boundaries. Transformers are a useful tool to supplement U-Net's local feature capture capabilities by capturing long-range dependencies and global context. This combination increases the accuracy of segmentation, particularly when structures of interest have intricate shapes or diverse appearances. The model becomes a useful tool in medical image analysis by incorporating Transformers, which gives it the capacity to extract significant characteristics, model spatial relationships, and utilize contextual information.

In the realm of medical image analysis, the pursuit of accurate segmentation has driven the development of novel models that leverage the strengths of both transformers and Convolutional Neural Networks (CNNs). One novel transformer model created for medical image segmentation problems is the Contoured Convolutional Transformer (CCT). It combines the benefits of transformers and Convolutional Neural Networks (CNNs) to improve segmentation accuracy. In CCT, transformers are used to capture global dependencies and context, while convolutional layers are utilised to extract local features.

CCT can handle complicated structures and appearance variations in medical pictures with ease thanks to its hybrid architecture. CCT is a potential method for precise and effective analysis of medical pictures since it combines CNNs with transformers to deliver state-of-the-art performance in medical image segmentation. Another transformer model for accurate segmentation is HRViT, or High-Resolution Vision Transformer. It is built on the ResNet architecture [3]. It attempts to bring together the advantages of Vision Transformers' long-range dependency modeling and ResNet's robust feature extraction capabilities. HRViT can extract detailed features from high-resolution images and capture global context and dependencies by adding transformer layers into the ResNet backbone. Due to its hybrid architecture, HRViT is a good choice for tasks like high-resolution image classification or segmentation that call for both global context modeling and fine-grained feature extraction. To improve its capacity for comprehensive feature extraction and long-range dependency modeling in image tasks, HRviT, which is built on HardNet, combines the high-resolution feature extraction capabilities of HardNet with the global context modeling of Vision Transformers. Tasks like dense image matching and retrieval, where both fine-grained characteristics and global context are critical, benefit greatly from this hybrid architecture.

Four sections make up the remaining portion of the paper: A summary of pertinent literature is provided in Section II. Our techniques for creating and applying the models are covered in Section III. Our experiment and the findings of a comparison between models are covered in Section IV. Finally, Section V offers suggestions for more research.

II. Literature Review

In recent years, transformers have emerged as a transformative force in various fields, revolutionizing natural language processing, computer vision, and beyond. Nowadays transformers have been used in medical image segmentation. This section explores the usage of transformers in medical image segmentation. Transformers are used in many ways and they are combined with different models like U-Net. Khan *et al.* [4] have presented a comprehensive summary of transformer-based models across various medical imaging applications, encompassing tasks such as segmentation, detection, classification, reconstruction, synthesis, registration, and the generation of clinical reports. They analyzed the combination of CNNs and transformers in segmentation that allows for the capture of both local and global features in the input image, which can improve segmentation accuracy. Gao *et al.* [5] have introduced a hybrid transformer architecture called UTNet is used to segment medical images. It improves medical image segmentation by incorporating self-attention into a convolutional neural network. The network can capture long-range dependencies at various scales with minimum overhead owing to the self-attention method. Karimi and Vasylechko *et al.* [6] have suggested a network for convolution-free medical image segmentation. A novel approach to 3D medical image segmentation has been introduced. This model relies on self-attention between nearby 3D patches, in contrast to all other contemporary models that use convolution as their primary building component.

Wang *et al.* [7] have developed a model with Contoured Convolutional Transformer Network (CCTrans). The U-shaped structure used in this architecture is made up of skip connections, a decoder, and an encoder. The gated module consists of Batch Normalization (BN), convolution, ReLU, and sigmoid layers, all of which were inspired by ResNet. In clinics, the CCTrans network is utilized to help physicians accelerate organ segmentation activities and enhance diagnostic effectiveness. He *et al.* [8] have proposed a High-Resolution Vision Transformer (HRViT) based on HRNet. The CFNet network leads cross-scale fusion features to efficiently connect to decoder features to address semantic gaps and uses a multi-view attention method for feature extraction. This method is computationally more effective and enables improved processing of details in MRI pictures. Shen *et al.* [9] have used a transformer, HarDNet Structures, and designed an

encoder-decoder network structure U-Net++, utilizing an intensely supervised training strategy, enabling the model to undergo supervised learning for its multi-branch output. After multi-layer convolution, the local feature information of the input image is obtained using the HarDNet68 module. The transformer module slices the input image before acquiring global feature information for the medical image. Chen *et al.* [10] have utilized U-Net and transformers and introduced a new approach TransUNet. Transformers are used as encoder and U-Net is used as decoder. It encodes comprehensive global context by treating image features as sequential data but also maximizes the utilization of low-level CNN features through the implementation of a hybrid U-shaped architectural design. TransUNet leverages Transformers' self-attention to capture global context, improving segmentation accuracy.

Yan *et al.* [11] have introduced TransHRNet, a hybrid CNN Transformer model that uses parallel transformers to segment 3D medical images. It is based on the Effective Transformer (EffTrans) block. A parallel transformer refers to the use of multiple transformers operating in parallel on different resolution streams of the input image. The parallel transformers exchange information across the different streams to learn global information and capture long-range dependencies in the medical image. TransHRNet uses a Transformer to capture many contexts and can retain multiple-resolution representations from CNN characteristics. Madhavi *et al.* [12] have proposed a two-step training procedure using an encoder-decoder-based architecture to segregate liver tumors. The 3D-IRCADb1 dataset is taken into consideration for training and testing due to its tumor complexity, and MDICE, a combined loss function, is used to improve the learning potential. Nguyen *et al.* [13] have evaluated Transformer-based semantic segmentation models for tumor delineation in histopathological images, specifically employing the PAIP liver histopathological dataset. The investigation involves a comparative analysis between six widely used Transformer-based models and six conventional CNN-based models. The findings reveal that Transformer-based models, such as Segmenter [14], Swin-Transformer, and TransUNet, outperform CNN-based models in tumor segmentation. These Transformer models harness the global contextual information within an image, incorporating insights from long-distance relationships and dependencies, to achieve better segmentation results. Kelei *et al.* [15] had provided a comprehensive overview of the applications of Transformers in the field of medical image analysis. This paper delves into the core concepts of Transformers, reviews different Transformer architectures tailored for medical image applications, discusses their limitations, and explores key challenges and opportunities in utilizing Transformers for medical image analysis.

The observations we have found based on the related work are:

(1) The existing literature review lacks hyperparameter tuning on Transformers.
(2) Preprocessing techniques like data augmentation and resizing have not been done for medical images to rectify data imbalance.
(3) The implementation of this research in real-time could enable doctors to treat a larger number of patients, thereby offering substantial advantages across healthcare settings.

III. Methodology

This study aims to introduce an enhanced TransUNet model for the precise segmentation of healthy organs, providing valuable assistance to radio-oncologists. Various techniques were employed to optimize the model's performance while maintaining the core TransUNet design, resulting in reduced computational requirements.

(1) Preprocessing methods were utilized to improve the quality of the images in both the training and testing phases.
(2) To address the imbalance in the dataset, we scale the input photos to 256×256 pixels.
(3) The Dice+BCE loss function is utilized to assess the Model.

This section describes the proposed methods for the segmentation of the GI tract. Information regarding the input dataset used for segmentation will be presented in Section A. The various pre-processing techniques discussed in Section B were used to improve the dataset before further processing. The TransUNet model Architecture, which was used in our project is shown in Section C. Fig. 1 describes the proposed model for the segmentation of the GI tract.

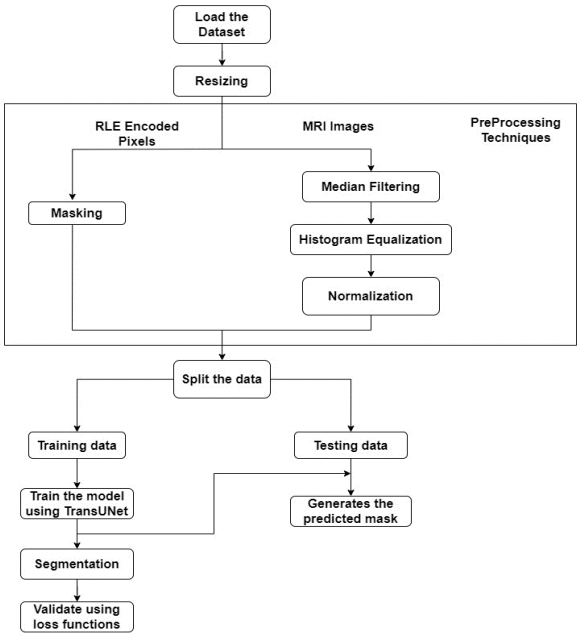


Fig. 1. Proposed workflow for segmentation of GI tract.

A. Dataset Description

This research makes use of the 'UW Madison GI Tract Image Segmentation dataset, comprising 38,496 MRI

scans from actual cancer patients. These scans were 16-bit grayscale PNGs [16]. The dataset encompasses 85 cases and includes a CSV file providing RLE-encoded pixels for the stomach, small bowel, and large bowel. Three columns make up the CSV file: id, class, and segmentation. "class" provides information about the object's expected class, "id" provides a unique identity for each scan in the dataset, and "segmentation" provides details on the RLE-encoded pixels for the discovered object. The dataset has a folder labeled "train" that contains slice images for specific cases on specific days. The names of the image files consist of four numerical values representing the dimensions of the image slices in terms of width and height (specified in pixels) and the pixel spacing in the horizontal and vertical directions (specified in millimeters as floating-point values). The MRI scan consists of 3 classes, namely large bowel, small bowel, and stomach. Our research utilizes 36,496 scans for training and 2,000 scans for testing. The dataset is sourced from the UW-Madison Carbone Cancer Centre, a leading institution in MRI-Linac-based radiotherapy [17]. The MRI scan in the UW Madison GI tract segmentation dataset is shown in Fig. 2.

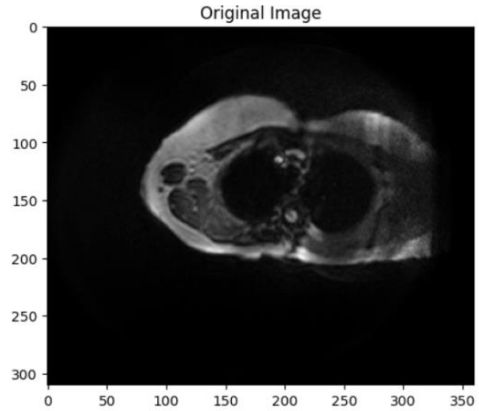


Fig. 2. A sample image from the dataset.

Table I represents the sample RLE-encoded pixels data for MRI Scans in the UW Madison GI tract segmentation dataset.

TABLE I. MRI SCANS DATA IN CSV FILE

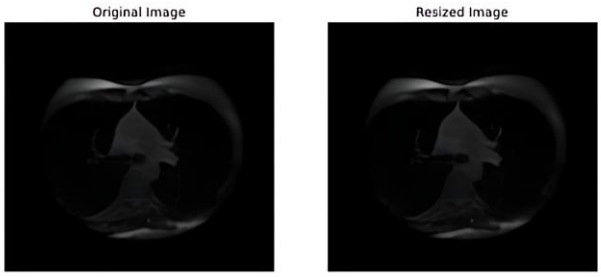
ID	class	segmentation
case123_day20_slice_0110	large_bowel	11845 10 12107 17
case123_day20_slice_0110	small_bowel	17403 5 17462 19
case123_day20_slice_0110	stomach	
case123_day20_slice_0111	large_bowel	11848 4 12110 12
case123_day20_slice_0111	small_bowel	16871 4 17135 8 17207
case123_day20_slice_0111	stomach	

B. Data Preprocessing

Data preprocessing is a crucial step in which raw data is cleaned, transformed, and organized. Data is pre-processed to enhance image quality and facilitate accurate analysis. It is the process of converting unprocessed data into a form that can be analyzed. It creates a standardized and optimized dataset, enabling more reliable diagnoses, precise analyses, and effective utilization. In this study, the pre-processing steps included are Resizing, Masking, Bias Correction, Histogram Equalization, Median Filtering, and Normalization. In the next sections, a detailed explanation of several pre-processing stages is provided.

1) Resizing

Resizing is the process of altering an image size. It modifies an image's dimensions by either increasing or decreasing them. In the context of MRI images, resizing is done to standardize the image dimensions for consistency in analysis or to meet specific requirements for input into machine learning models. This process typically involves interpolation techniques to adjust pixel values and spatial relationships, allowing images of different sizes to be brought to a uniform size for ease of comparison, analysis, or model training. The most prevalent image size in the sample is 266×266, occurring in 67.33% of cases. Following in descending order of frequency are dimensions 276×276, 310×360, and 234×234. To maintain consistency in image sizes for mask generation of segmented areas, we standardized all images to 256×256 pixels. Fig. 3 displays the MRI scan both before and after the resizing [18].

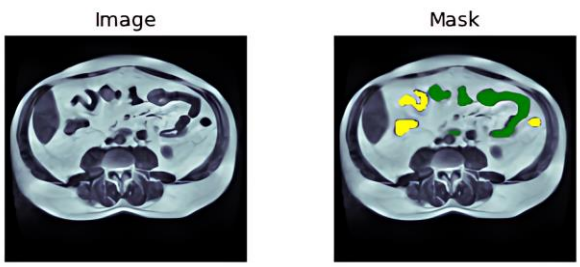


(a) (b)

Fig. 3. MRI image of GI tract. (a) Before resizing, (b) After resizing.

2) Masking

The dataset contains Run-Length Encoding (RLE) encoded pixels. RLE represents the consecutive runs of pixels for each segmented region. To generate masks, decode the RLE information for identifying the precise pixel positions and their corresponding regions within the image. By reconstructing the pixel positions from the RLE data, a binary mask is created, where pixels belonging to the segmented area are marked, facilitating accurate delineation of structures or objects in medical images [19]. Fig. 4 represents the MRI scan and corresponding Mask generated from RLE-encoded pixels.



(a) (b)

Fig. 4. (a) MRI image, (b) Mask.

3) Bias correction

Bias Correction is a preprocessing technique that enhances the overall quality and reliability of MRI images. This technique is designed to compensate for non-uniformities in signal intensity caused by factors like magnetic field inhomogeneities or acquisition variations. Magnetic Resonance Imaging (MRI) can suffer from spatially varying signal intensities, known as bias or shading, which may obscure important anatomical details. Bias correction methods aim to rectify these variations, ensuring a more consistent and accurate representation of tissue intensities. Fig. 5 displays the MRI scan both prior to and following bias correction.

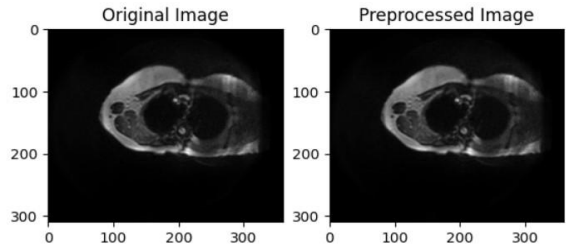


Fig. 5. MRI Image. (a) Before bias correction, (b) After bias correction.

4) Histogram equalization

Histogram Equalization is a commonly employed method in image processing, designed to enhance the contrast of an image by reassigning the intensity values throughout its histogram. It accomplishes this by efficiently extending the image's intensity range or spreading out the most frequent intensity values. When the useful data of an image is represented by close contrast values, this method typically enhances the global contrast of the image. This process improves the overall image quality, potentially aiding subsequent medical image analysis tasks by highlighting relevant anatomical features in MRI scans. Fig. 6 shows the MRI scan pre- and post-histogram equalization.

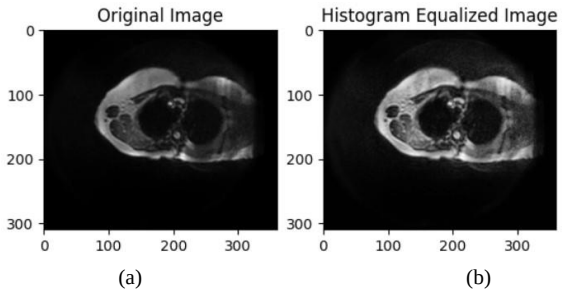


Fig. 6. MRI image (a) Before histogram equalization, (b) After histogram equalization.

5) Median filtering

Median filtering is a common image processing technique used for noise reduction while preserving the edges and fine details of the image [20]. The median filter functions by systematically examining each pixel in the image and replacing its value with the median value of its neighboring pixels through an iterative process., contributing to a cleaner, more interpretable image. Fig. 7 represents the MRI Scan before and after median filtering.

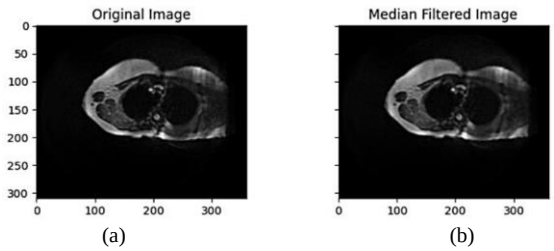


Fig. 7. MRI image. (a) before median filtering, (b) after median filtering.

6) Normalization

Normalization in MRI image processing involves scaling pixel intensities to a standardized range, often [0, 1]. This technique ensures consistency and comparability across images, mitigating variations in intensity due to acquisition differences [21]. Normalization enhances the interpretability of MRI images, facilitating accurate analysis and diagnosis. Fig. 8 displays the MRI scan both prior to and following Normalization.

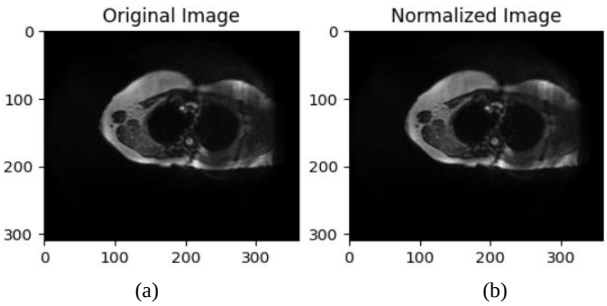


Fig. 8. MRI image (a) Before normalization, (b) After normalization.

C. Segmentation using TransUNet model

A proposal called TransUNet is made, which is beneficial to both Transformers and U-Net. In order to extract global contexts, the transformer takes tokenized patches of images from a CNN feature map as its input, while the decoder enhances the encoded features and integrates them with the high-resolution CNN feature maps for accurate localization. The task of locating and defining structures or regions of interest within medical pictures, such as organs or abnormalities, is known as medical image segmentation, and TransUNet was created expressly for this purpose. By using the advantages of transformer models which are renowned for their capacity to identify global dependencies within sequences and applying them to picture data, it seeks to get beyond the drawbacks of traditional CNN-based models [22]. Fig. 9 represents the TransUNet Architecture.

TransUNet is an encoder-decoder architecture, that effectively captures various levels of abstraction and incorporates skip connections to fuse detailed information from initial layers. By implementing self-attention mechanisms, it selectively focuses on relevant regions within the image, thereby enhancing segmentation

accuracy. TransUNet adeptly combines the strengths of both transformers and U-Net architectures, resulting in a highly effective model. By integrating transformer components, it harnesses the power of self-attention mechanisms, allowing for global context understanding and long-range dependencies modeling. The skip connections within the U-Net architecture enable the seamless fusion of high-resolution details from the encoder with contextual information from the decoder, thereby preserving fine-grained information crucial for accurate segmentation. By incorporating self-attention methods through the use of transformers, the model is able to concentrate on pertinent areas of the picture, improving segmentation accuracy. TransUNet's performance in segmenting medical pictures is improved by its ability to capture both local details and global contextual information. TransUNet can adapt to pictures of different sizes because of its patch-based methodology, in contrast to typical CNN-based models that are dependent on fixed image dimensions. In a variety of medical image segmentation tasks, TransUNet has proven to function at the cutting edge, demonstrating its efficacy in precisely identifying structures inside medical pictures.

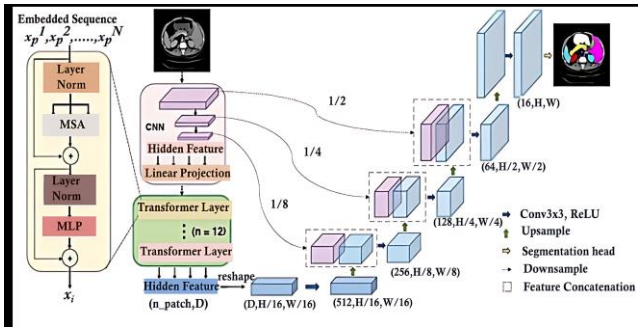


Fig. 9. TransUNet architecture.

IV. Result and Discussion

A. Hyper Parameter Tuning

This is the process of selecting the best configuration of hyperparameters for a machine-learning model. They control aspects of the learning algorithm and influence the model's performance and behavior. Hyperparameter tuning involves systematically searching through a predefined set of hyperparameter values to identify the combination that maximizes the model's performance on a validation dataset [23], and finding the ideal combination of hyperparameter values that produces the highest performance on a particular dataset. Effectively tuning hyperparameters often leads to improved model accuracy, generalization, and efficiency. Hyperparameters include batch size, learning rate, resource optimization, and early stopping, number of epochs and adaptive strategies.

Optimizing the learning rate can contribute to sustainable training by efficiently guiding the model towards convergence. A well-tuned learning rate reduces the overall computational resources needed for training, promoting energy efficiency and minimizing the carbon footprint associated with model development.

Batch size influences the hardware resource requirements during training. Choosing an optimal batch size can lead to more energy-efficient training processes, aligning with sustainability goals by minimizing the overall computational power consumption. Tuning the number of epochs contributes to sustainable model development by avoiding unnecessary computational costs associated with excessive training. Optimizing this hyperparameter ensures that the model achieves satisfactory performance within a reasonable training time, conserving resources.

1) Resource optimization

Efficient hyperparameter tuning targets resource optimization by identifying the optimal configuration with minimal computational resources. This not only enhances model accuracy and generalization but also reduces the overall computational load and energy consumption during the training and evaluation phases. The process aligns with sustainability practices by minimizing the environmental impact associated with machine learning model development.

2) Early stopping and adaptive strategies

Incorporating early stopping criteria and adaptive learning techniques during hyperparameter tuning further contributes to sustainability. Early stopping terminates the training process when the model's performance plateaus, preventing unnecessary computations. This strategic approach not only conserves computational resources but also aligns with sustainable practices, emphasizing the importance of minimizing energy consumption in machine learning model optimization.

3) Sustainability benefits

The integration of hyperparameter tuning and novel sampling methodologies yields significant sustainability benefits. By optimizing the utilization of computational resources, this approach reduces unnecessary computational load, thereby minimizing energy consumption throughout the model development lifecycle. This resource efficiency not only enhances the model's performance but also aligns with sustainable practices, mitigating the environmental impact associated with training and optimizing machine learning models.

The hyperparameters used in this proposed model are batch_size, Epochs, and n_splits. A batch size of 64, epochs of 25, and n_spits of 5, early stopping, and adaptive strategy were selected based on considerations of computational efficiency and model convergence.

B. Comparison of Segmentation Performance Using Different Combinations of Loss Functions

The performance of the TransUNet model is evaluated on the UW Madison GI tract Segmentation Dataset using Various Metrics including Model Loss, Dice coefficient, and IOU Coefficient. Our Proposed TransUNet Model was evaluated using the Dice+BCE Loss Function. Model loss evaluates training and validation loss.

1) Analysis of training and validation loss

Training loss measures the error between predicted and actual values during the model's training phase, indicating how well it fits the training data. Validation loss assesses

the model's performance on a separate dataset not used for training, gauging its ability to generalize to new, unseen data. Our Model achieves a training loss of 0.0923 and a validation loss of 0.1512 using Dice + BCE loss function. The reported training and validation losses, along with the use of the Dice + BCE Loss Function, provide insights into the performance of the TransUNet model on the UW Madison GI tract Segmentation Dataset. These metrics are essential for evaluating the effectiveness of the model in learning and generalizing from the training data to new, unseen data.

The low training loss (0.0923) indicates that the model is fitting the training data well, capturing the underlying patterns and features. Fig. 10 depicts a comparison of the loss incurred by various loss functions.

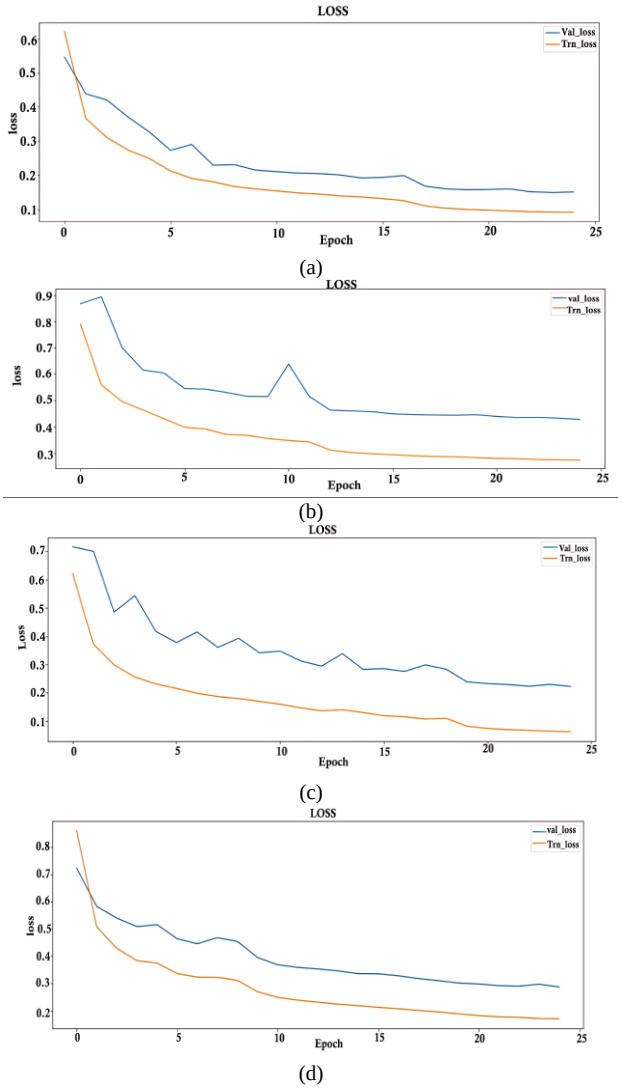


Fig. 10. Training and validation loss curves for (a) Dice + BCE loss, (b)Tverskey loss, (c) Dice loss, (d) Focal + Dice loss.

2) *Analysis of the dice coefficient*

The dice coefficient measures the agreement between the predicted and true segmentation masks. The Dice coefficient ranges from 0 to 1, where 1 indicates perfect overlap between the predicted and true segmentation. Higher dice coefficients imply better segmentation accuracy and model performance. It is a measure of the similarity between the predicted and ground truth segmentation masks, and a higher value indicates better overlap and suggests that the model is effective at learning the patterns and structures in the training data Our TransUNet Model achieves a dice score of 0.854 on the training dataset and 0.7398 on the validation dataset. Fig. 11 represents the comparison of dice coefficient metrics among various loss functions.

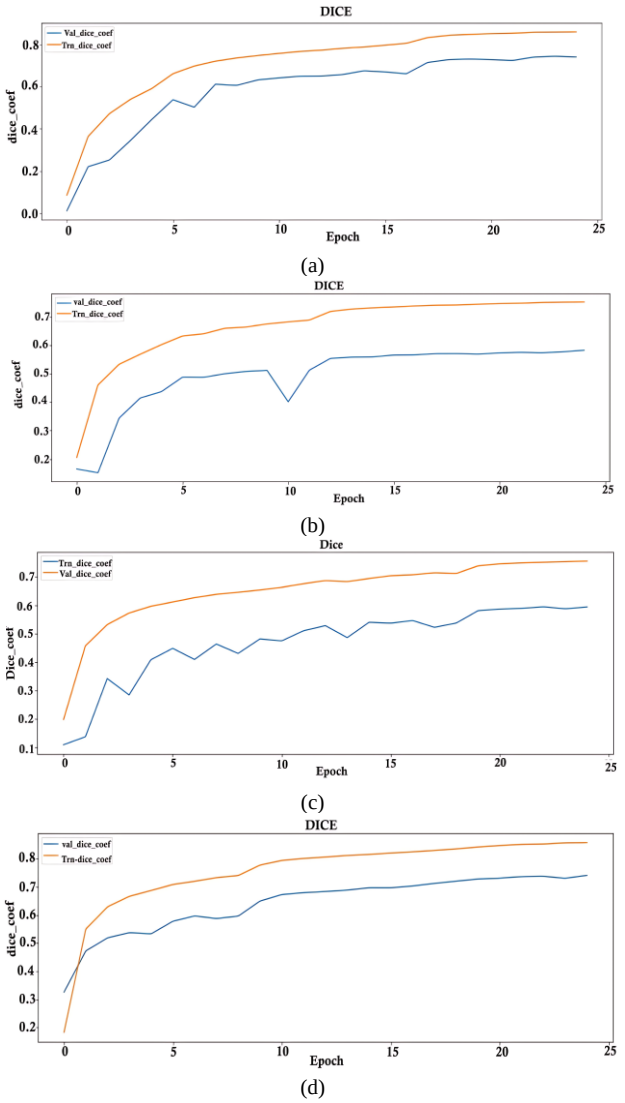


Fig. 11. Dice metrics for (a) Dice + BCE loss, (b)Tverskey loss, (c) Dice loss, (d) Focal + Dice loss.

3) *Analysis of the IOU coefficient*

IOU is calculated as the intersection of the predicted and true regions divided by the union of these regions. The IoU coefficient ranges from 0 to 1, where 0 indicates no overlap, and 1 corresponds to perfect overlap. IoU is a valuable metric for assessing the accuracy of segmentation models, providing insight into how well the predicted and true regions align in relation to their combined area. Our Model achieves an IOU Score of 0.7964 on the training dataset and 0.7962 on the Validation dataset. Fig. 12 presents a comparison of Intersection over Union (IOU) coefficient metrics among various loss functions.

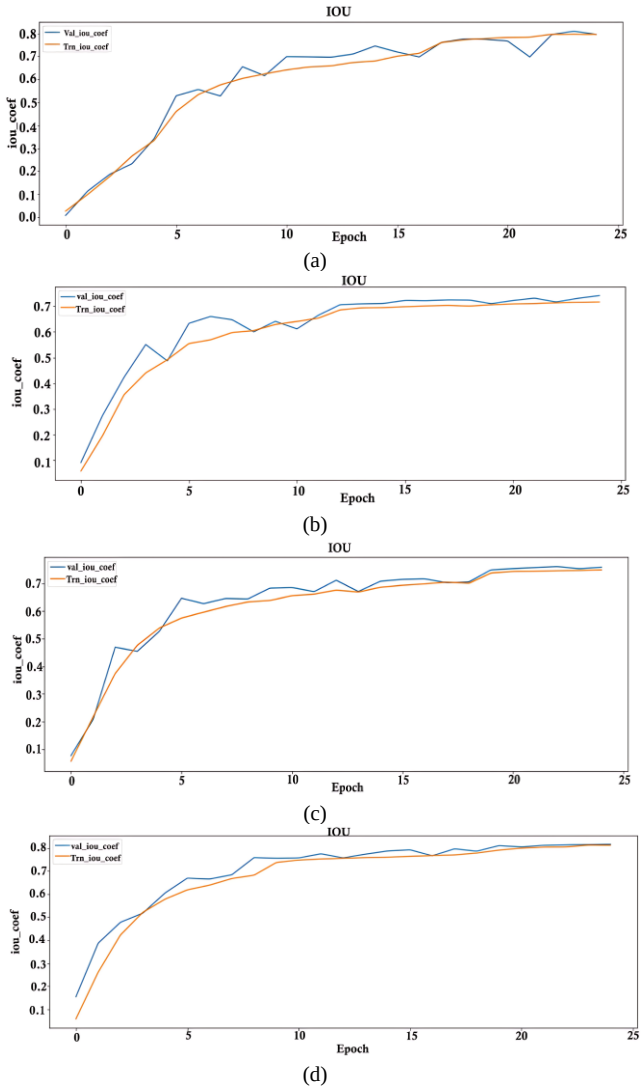


Fig. 12. IOU metrics for (a) Dice+BCE loss, (b)Tverskey loss, (c) Dice loss, (d) Focal+Dice loss.

4) Testing for various loss functions

Testing involves evaluating the performance of a model on a set of images distinct from those used during training and validation. The MRI images serve as input data, the ground truth masks provide a reference for the actual anatomical structures, and the predicted masks are the model's segmentation outputs. Testing MRI Scans with Ground truth and predicted mask for various loss functions are shown in Fig. 13.

Testing MRI Scans with Ground truth and predicted mask are shown in Fig. 14(a) shows the correctly predicted mask when compared with the ground truth, whereas in Fig. 14(b) shows the wrongly predicted mask compared to the ground truth.

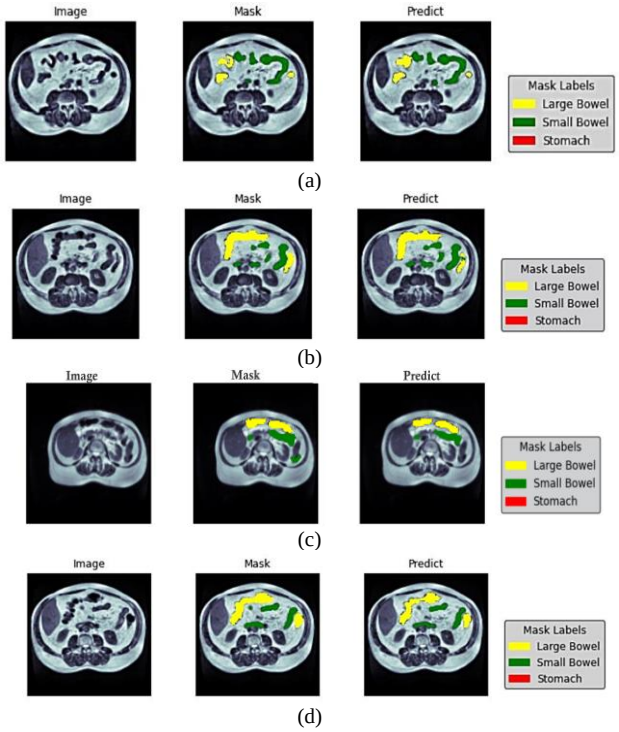


Fig. 13. Testing images for (a) Dice + BCE loss, (b)Tverskey loss (c) Dice loss (d) Focal + Dice loss.

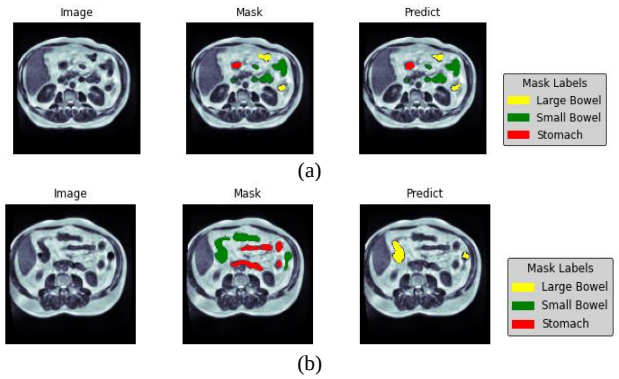


Fig. 14. Testing images for Dice + BCE loss, (a) correctly predicted mask, (b) wrongly predicted mask.

Table II shows the Comparison of training loss, IOU Coefficient and Dice Coefficient among different combination of loss functions such as Dice+BCE loss, Focal loss, Dice loss, and Focal+Dice loss.

TABLE II. COMPARISON OF TRAINING LOSS, IOU COEFFICIENT AND DICE COEFFICIENT FOR DIFFERENT COMBINATION OF LOSS FUNCTION

Metrics	Training loss	IOU Coefficient	Dice Coefficient	Validation loss	Val_IOU Coefficient	Val_Dice Coefficient
Dice + BCE loss	0.0923	0.7964	0.8584	0.1512	0.7962	0.7398
Focal loss	0.2743	0.7157	0.7517	0.4283	0.7409	0.5823
Dice loss	0.2616	0.7494	0.7557	0.4226	0.7589	0.5945
Focal + Dice loss	0.1715	0.8126	0.8561	0.2873	0.8163	0.7405

Table III shows the Comparison of dice and IOU metrics for various models on the UW Madison GI tract Segmentation Dataset.

TABLE III. COMPARISON OF DICE AND IOU METRICS FOR VARIOUS MODEL

No.	Model	IOU coefficient	Dice Coefficient
1	U-Net with EfficientNet-B3 as backbone [24]	85.3%	-
2	U-Net with Resnet50 as backbone [24]	84.9%	-
3	U-Net with VGC16 [24]	82.7%	-
4	U-Net [25]	-	51%
5	Mask R-CNN [25]	-	73%
6	Inception V3 [26]	76.87%	60.49%
7	SeResNet 50 [26]	75.88%	58.2%
8	DenseNet121 [26]	74.86%	55.58%
9	VGG 19 [26]	66.46%	47.41%
10	**Proposed Model**	**85.84%**	**79.64%**

V. CONCLUSION

We have proposed a TransUNet model for the segmentation of stomach and intestines in the GI tract. TransUNet consists of an encoder and decoder, the encoder extracts hierarchical features from the input image, capturing contextual information, while the decoder translates these features into a pixel-wise segmentation map. The primary emphasis lies in optimizing hyperparameters for the TransUNet model, including parameters like learning rate, batch size, and epochs. Data preprocessing methods have been applied to rectify dataset imbalances and enhance the quality of MRI scans. Assessing the TransUNet model through various combinations of loss functions. Dice+Binary Cross-Entropy (BCE) loss function offers significant insights into its effectiveness in handling medical image segmentation tasks. The TransUNet model achieves a dice score of 0.8584 and an IOU score of 0.7964 on the UW Madison GI tract segmentation dataset. We hypothesize that potential future advancements could further improve our results and address the limitations inherent in the current network architecture. Future developments can be an exploration of advanced data augmentation strategies to diversify the training dataset, implementation of ensemble methods by combining models like U-Net and Mask R-CNN possibly with U-Net acting as the foundation of Mask R-CNN and can enhance the accuracy of the model by using curriculum Learning [27].

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Bindu Madhavi Tummala, Dasari Chinna Veeraiah Conducted the research, Jaladi Rishitha and Dasari Chin Veeraiah had analyzed the data and worked on the Methodology, Bindu Madhavi Tummala and Aruna Kumari Peruri worked on Testing. Rishitha Jaladi and Aruna Kumari Peruri wrote the paper, all authors approved the final version.

REFERENCES

[1] M. Arnold, C. C. Abnet, R. E. Neale, R. E. Vignat, E. L. Giovannucci, K. A. McGlynn, and F. Bray, "Global burden of 5 major types of gastrointestinal cancer," *Gastroenterology*, vol. 159, no. 1, pp. 335–349, 2020.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany*, 2015, pp. 234–241.

[3] A. Taha, "High-resolution images and efficient transformers," *Medium*, 2022.

[4] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, 102802, 2023. https://doi.org/10.1016/j.media.2023.102802

[5] Y. Gao, M. Zhou, and D. N. Metaxas, "UTNet: A hybrid transformer architecture for medical image segmentation," in *Proc. 24th International Conference on Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*, Strasbourg, France, 2021, pp. 61–71. https://doi.org/10.1007/978-3-030-87199-4_6

[6] D. Karimi, S. D. Vasylechko, and A. Gholipour, "Convolution-free medical image segmentation using transformers," in *Proc. 24th International Conference on Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*, Strasbourg, France, 2021, pp. 78–88. https://doi.org/10.1007/978-3-030-87193-2_8

[7] J. Wang, H. Zhang, and Z. Yi, "CCTrans: Improving medical image segmentation with contoured convolutional transformer network," *Mathematics*, vol. 11, no. 9, 2023. https://doi.org/10.3390/math11092082

[8] K. He, F. Gou, and J. Wu, "Image segmentation technology based on the transformer in medical decision-making system," *IET Image Processing*, vol. 17, no. 10, pp. 3040–3054, 2023. https://doi.org/10.1049/ipr2.12854

[9] T. Shen and H. Xu, "Medical image segmentation based on Transformer and HarDNet structures," *IEEE Access*, vol. 11, pp. 16621–16630, 2023. doi:10.1109/ACCESS.2023.3244197

[10] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint, arXiv:2102.04306, 2021.

[11] Q. Yan, S. Liu, S. Xu *et al.*, "3D Medical image segmentation using parallel transformers," *Pattern Recognition*, vol. 138, 109432, 2023. https://doi.org/10.1016/j.patcog.2023.10943

[12] B. M. Tummala and S. S. Barpanda, "Liver tumor segmentation from computed tomography images using multiscale residual dilated encoder-decoder network," *International Journal of Imaging Systems and Technology*, vol. 32, no. 2, pp. 600–613, 2022.

[13] C. Nguyen, Z. Asad, R. Deng, and Y. Huo, "Evaluating transformer-based semantic segmentation networks for pathological image segmentation," in *Proc. Medical Imaging 2022: Image Processing*, 2022, vol. 12032, pp. 942–947.

[14] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262–7272.

[15] K. He, C. Gan, Z. Li *et al.*, "Transformers in medical image analysis," *Intelligent Medicine*, vol. 3, no. 1, pp. 59–78, 2023.

[16] Kaggle. (2022). UW-Madison GI tract image segmentation. [Online]. Available: https://www.kaggle.com/competitions/uw-madison-gi-tract-image-segmentation

[17] D. A. Jaffray and M. K. Gospodarowicz, "Radiation therapy for cancer," *Cancer: Disease Control Priorities*, vol. 3, pp. 239–248, 2015.

[18] O. Rukundo, "Effects of image size on deep learning," *Electronics*, vol. 12, no. 4, 2023. https://doi.org/10.3390/electronics12040985

[19] S. Jagadeesh, T. Venkateswarlu, and M. Ashok, "Run length encoding and bit mask-based data compression and decompression using verilog," *Int. J. Eng. Res. Technol.(IJERT)*, vol. 1, no. 7, pp. 1–7, 2021. doi: 10.17577/IJERTV1IS7515

[20] S. Suhas and C. R. Venugopal, "MRI image preprocessing and noise removal technique using linear and nonlinear filters," in *Proc. 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, 2017. pp. 1–4. doi: 10.1109/ICEECCOT.2017.8284595

[21] H. S. Obaid, S. A. Dheyab, and S. S. Sabry, "The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning," in *Proc. 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, 2019, pp. 279–283. doi: 10.1109/IEMECONX.2019.8877011

[22] S. Sang, "Review: TransUNet—Transformers make strong encoders for medical image segmentation," *Medium*, 2023.

[23] A. Brenning, "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data," *Ecological Modelling*, vol. 406, 2021. https://doi.org/10.1016/j.ecolmodel.2019.06.002

[24] M. Sharma, "Automated GI tract segmentation using deep learning," arXiv preprint, arXiv: 2206.11048, 2022. https://doi.org/10.48550/arXiv.2206.11048

[25] A. Chou, W. Li, and E. Roman, "GI tract image segmentation with U-Net and mask R-CNN," *Image Segmentation with U-Net and Mask R-CNN*, 2022.

[26] N. Sharma, S. Gupta, D. Koundal, S. Alyami, H. Alshahrani, Y. Asiri, and A. Shaikh, "U-Net model with transfer learning model as a backbone for segmentation of gastrointestinal tract," *Bioengineering*, vol. 10, no. 1, 2023. https://doi.org/10.3390/bioengineering10010119

[27] B. M. Tummala and S. S. Barpanda, "Curriculum learning based overcomplete U-Net for liver tumor segmentation from computed tomography images," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1620–1629, 2023.