Scientific Reports

https://doi.org/10.1038/s41598-026-38247-7

Article in Press

Multi-level attention DeepLab V3+ with EfficientNetB0 for GI tract organ segmentation in MRI scans

Received: 25 June 2025

Accepted: 29 January 2026

Published online: 06 February 2026

Cite this article as: Sharma N., Gupta S., Al-Yarimi F.A.M. *et al.* Multi-level attention DeepLab V3+ with EfficientNetB0 for GI tract organ segmentation in MRI scans. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-38247-7

Neha Sharma, Sheifali Gupta, Fuad Ali Mohammed Al-Yarimi, Upinder Kaur, Salil Bharany, Ateeq Ur Rehman & Belayneh Matebie Taye

Multi-Level Attention DeepLab V3+ with EfficientNetB0 for GI Tract Organ Segmentation in MRI Scans

Neha Sharma*
Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India; sharma.neha@chitkara.edu.in

Sheifali Gupta
Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India; sheifali.gupta@chitkara.edu.in

Fuad Ali Mohammed Al-Yarimi
Applied College of Mahail Aseer, King Khalid University, Muhayil Aseer 62529, Saudi Arabia; fuadalyarimi@gmail.com

Upinder Kaur
Department of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab 144411, India; upinderkaur45@gmail.com

Salil Bharany
Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India; salil.bharany@gmail.com

Ateeq Ur Rehman
Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamilnadu, India; ateqrehman@gmail.com
Applied Science Research Center, Applied Science Private University, Amman, Jordan
University Center for Research and Development, Chandigarh University, Mohali, Punjab 140413, India

Belayneh Matebie Taye*

Department of Computer Science, College of Informatics, University of Gondar, Gondar, Ethiopia; belayneh.matebie@uog.edu.et

***Corresponding author:** Belayneh Matebie Taye (e-mail: belayneh.matebie@uog.edu.et) and Neha Sharma (email: sharma.neha@chitkara.edu.in

Abstract

Gastrointestinal (GI) cancer is a fatal malignancy that affects the organs of the GI tract. The rising prevalence of GI cancer has recently influenced the health of millions of people. To treat GI cancer, radiation oncologists must carefully focus X-rays on tumors while avoiding other unaffected organs in the GI tract. This research proposes a novel approach to segment healthy organs within the GI tract from magnetic resonance imaging (MRI) scans using a multi-level attention DeepLab V3+ model. The proposed model aims to enhance segmentation performance by incorporating state-of-the-art approaches, such as atrous convolutions and EfficientNet B0 as an encoder, by leveraging hierarchical information present in the data. Here, the attention mechanism is applied at multiple levels of features, i.e., low, medium, and high, to capture and leverage hierarchical information present in the data. At the same time, EfficientNet B0 extracts deep and meaningful features from input images, providing a robust representation of GI tract structures. Hierarchical feature fusion combines local and global contextual information, resulting in more accurate segmentation with fine-grained details. The model is implemented using the UW-Madison dataset, comprising MRI scans from 85 patients with gastrointestinal cancer. To optimize the model, it has been simulated with different parameters, including optimizers, the number of epochs, and cross-validation folds. The model has achieved performance metrics such as a model loss of 0.0044, a dice coefficient of 0.9378, and an Intersection over Union (IoU) of 0.921.

Keywords: *Gastrointestinal Tract, Segmentation, Multi-level Attention, DeepLab V3+, EfficientNet B0, Deep Learning.*

1. Introduction

Medical image segmentation is the process of identifying and extracting specific areas of interest in an image, such as organs in the body and tumors [1]. The primary goal of medical image segmentation is to accurately identify and precisely locate critical anatomical regions necessary for efficient cancer treatment [2]. Nevertheless, the inefficiency of manual segmentation may be inferred from its repetitive, time-consuming nature, lower precision, and the variability of imaging techniques. This may be determined from the fact that it possesses a wide array of applications, such as the examination and identification of various medical conditions, including skin cancer [3-5], breast cancer [6], brain tumour [7-8], and gastrointestinal (GI) cancer [9-10].

In the last few years, many patients have been diagnosed with GI tract cancer worldwide [10-11]. Gastrointestinal cancer is a life-threatening condition that affects the digestive system. It has a survival rate of almost 30%. Radiation therapy is the most common treatment for GI cancers. During radiation treatment, oncologists direct X-rays on the affected area while avoiding the healthy organs [11]. Oncologists can view the tumor's location for exact dosages according to the presence of tumor cells, which may change daily, using linear accelerator devices and magnetic resonance imaging (MRI) [11]. The manual outlining of the organs takes a lot of time and effort, which can cause treatments to take up to an hour daily. The proposed work outlines the stomach and intestines to allow for changes in the X-ray beam's direction to improve the dosage distribution to the tumor while neglecting the healthy organs. By minimizing collateral damage to adjacent organs, automatic segmentation reduces treatment-related side effects and complications, enhancing patients' quality of life during and after therapy. Additionally, precise segmentation allows for the optimization of radiation dose distribution and the exploration of advanced treatment techniques, ultimately leading to improved tumor control probability, reduced recurrence rates, and enhanced long-term survival outcomes. More patients might receive effective care due to the automated segmentation procedure, which would speed up the healing process.

Deep learning methods have formed the foundation for many modern image segmentation and classification solutions [12]. In the proposed work, the multi-level attention DeepLab V3+ model has been implemented for GI tract segmentation. Earlier, it has been implemented for other tasks such as remote sensing [13], brain tumors [14], skin lesions [15], and kidney tumors [16]. This paper used the DeepLab V3+ to segment the GI tract organs for the first time.

The major findings of this research work are as follows:

- A multi-level attention DeepLab V3+ model is proposed to segment healthy organs in the GI tract. The model integrates state-of-the-art techniques such as atrous convolutions, EfficientNet B0 as an encoder, and a multi-level attention mechanism to enhance segmentation accuracy.
- A attention mechanism is applied at multiple levels of features, i.e., low level, medium level, and high level, to capture and leverage hierarchical information present in the data. A channel-wise attention

module focuses more on relevant channel features for every layer of atrous convolution used in DeepLab V3+.

- EfficientNet B0, as an encoder, facilitates the extraction of deep and meaningful features from input images, contributing to a better representation of gastrointestinal tract structures. By using EfficientNet B0 as the encoder, the proposed approach extracts deep and meaningful features from input images, providing a robust representation of GI tract structures.
- The proposed design has been implemented on the UW-Madison dataset with 38496 magnetic resonance imaging (MRI) scans of 85 patients. The proposed model has been trained with varying hyperparameters like optimizers, number of epochs, and cross-folds for optimizing the model. Also, the proposed design has been compared using various performance parameters, i.e., model loss, dice, and Intersection over Union (IoU) coefficient.

The rest of the paper has been divided as follows: section 2 provides the Related Work of the GI tract segmentation. Section 3 is the Dataset Description utilized for applying the proposed model. Section 4 will be devoted to the proposed multi-level attention deep lab V3+ model. Section 5 is the Results and Discussion after implementation; Section 6 illustrates the state-of-the-art comparison and section 7 Conclusion and Future work of the current research work.

2. Related Work

Automated segmentation of medical images has been an area of interest since the 19th century, fuelled by the growing demand for precise, efficient, and automated techniques to aid clinical diagnosis and treatment planning [17]. Over the past few years, segmentation of the GI tract has become a prominent area of interest, with applications varying from disease diagnosis and surgical planning to robotic navigation and cancer detection [17-19]. There have been a number of investigations into Convolutional Neural Network (CNN) based architectures. Ye et al. [20] proposed the SIA-Unet model, which uses a spatial attention mechanism to improve MRI scan segmentation by selectively filtering spatial data. While SIA-Unet demonstrated improved performance through uniform longitudinal guidance, it was limited by its reliance on conventional U-Net structures and lacked explicit multi-scale context fusion. In contrast, our method combines multi-level attention with atrous spatial pyramid pooling (ASPP), enabling it to effectively capture both fine and global semantic features across different scales. Nemani et al. [21] proposed a hybrid model to balance accuracy and computational cost. Their approach mitigates this by using EfficientNet B0 as a lightweight yet powerful encoder, along with an attention mechanism that enhances informative features without significantly increasing complexity. Chou et al. [22] employed the Mask R-CNN framework to segment human body parts in clinical images. Although it yielded a Dice score of 0.51, the method struggled with small or overlapping anatomical structures. By contrast, our proposed model achieved a higher Dice score of 0.73, attributed to its ability to focus on channel-wise salient features and maintain spatial context through decoder-based upsampling and skip connections. Niu et al. [23] presented a GI tract segmentation method using a hybrid of residual connections and U-Net, along with a feature fusion strategy. Their method improved the IoU by 2.5% over conventional approaches. While residual learning facilitates better gradient flow, it does not explicitly incorporate attention to refine features. In contrast, our multi-level attention framework enables the model to selectively enhance relevant features at low-, mid-, and high-levels, contributing to more accurate segmentation, particularly in complex or overlapping organ regions.

Li et al. [24] proposed a 2.5D model that combines adjacent slices to leverage spatial dependencies across slices. Their fusion method of 2.5D and 3D improved the Dice by 0.36% and IoU by 0.12%. Although beneficial for 3D segmentation, this approach requires higher computational resources and may not generalize well to single-slice datasets. Our model operates on individual 2D slices but achieves comparable or better accuracy due to its efficient multi-scale feature aggregation and attention-based refinement. Chia et al. [25] explored the use of FiLM in segmentation with ResNet50 and alternative backbones, identifying its effectiveness when test and training distributions align. Their results suggest performance dependency on data similarity. On the other hand, their model is robust in different folds of the UW-Madison dataset because it has deep semantic representation ability via ASPP and attentions. Georgescu et al. [26] proposed ensemble-based therapeutic image segmentation models via multi-network fusion. While ensemble models improve the performance, they make inference time longer and need a large amount of training. Our one-network solution provides competitive segmentation performance at the cost of ensemble learning overhead and hence is more appropriate for real-time or low-resource environment. Jiang et al. [27] proposed BiFTransNet, a transformer segementation model with a BiFusion decoder that combines global and local features. Their approach reported an IoU of 86.54% and a Dice of 89.51. Their structure achieves high accuracy with fewer parameters using compound-scaling EfficientNet and a slim attention module, achieving a better balance between accuracy and efficiency. Qiu et al. [28] employed a Swin Transformer-based UPerNet,

While transformer backbones are well-suited to capture global context, they tend to lose spatial precision in boundary areas. By contrast, our model, which uses its skip connections and multi-level attention-guided decoder, preserves boundary acuity without losing large semantic information. John et al. [29] used EfficientNet B7 and compound scaling for GI tract image segmentation. Even though EfficientNet B7 is deeper to extract features, it comes with increased memory and computational requirements. We rather use EfficientNet B0 to preserve computational efficiency without losing competitive performance. Our application of attention mechanisms still further sharpens the learned features, facilitating better region localization with lower model complexity. Wang et al. [30] investigated the application of soft robotic endoscopes for GI imaging, also highlighting the need for accurate segmentation in autonomous medical procedures. Though not algorithm-specific, their research highlights the clinical need for high-quality segmentation, which our proposed solution addresses directly via attention-augmented multiscale learning.

Li et al. [31] proposed the UCFNNet model with lesion learners and noise suppression gates for diagnosis of ulcerative colitis. While conceptually similar to our work with attention mechanisms, their model is designed specifically for disease-specific segmentation. Our model, on the other hand, addresses healthy organ boundary segmentation in the UW-Madison dataset and can be used as a baseline for broader extension into pathological analysis. Song et al. [32] introduced a transformer-based cluster center-augmented network for semantic segmentation, which showed better segmentation in intricate images. Though efficient, transformer-based clustering increases the complexity and needs significant GPU memory. Our model provides a less complex and computationally efficient structure while maintaining deep contextual understanding via multi-level attention. Lastly, more recent papers [33][34] highlight the growing overlap of encoder-decoder architectures, attention modules, and multi-scale feature learning. They have proposed a standard UNet-based method in our earlier work [35], which had encouraging results but failed to extract multi-scale and contextual information efficiently. Despite numerous such attempts that have progressed impressively, there are challenges due to the inherent complex nature, overlapping boundaries, and intensity variability of GI tract organs. This paper directly focuses on these difficulties by introducing an innovative Attention DeepLab V3+ model based on a multi-level attention scheme and EfficientNet B0 encoder, targeting segmentation of healthy anatomy regions in the UW-Madison dataset. The proposed Attention DeepLab V3+ model introduces several distinct contributions: (1) the use of EfficientNet B0 for lightweight yet powerful encoding, (2) the integration of a novel channel-based attention mechanism within ASPP to enhance multi-scale feature refinement, and (3) a decoder capable of recovering fine spatial resolution by fusing attention-enhanced semantic features with shallow encoder features.

3. Dataset Description

The anonymized MRI scans of radiotherapy at the UW-Madison Carbone Cancer Center [36] were a foundation of proposed study. The UW-Madison GI Tract MRI dataset is presently the only publicly available dataset providing annotated multi-organ gastrointestinal segmentation masks, that’s why it was selected for this study. The dataset comprises data for 85 patients having 38496 MRI scans in 16-bit PNG format, which is taken from Kaggle [36].

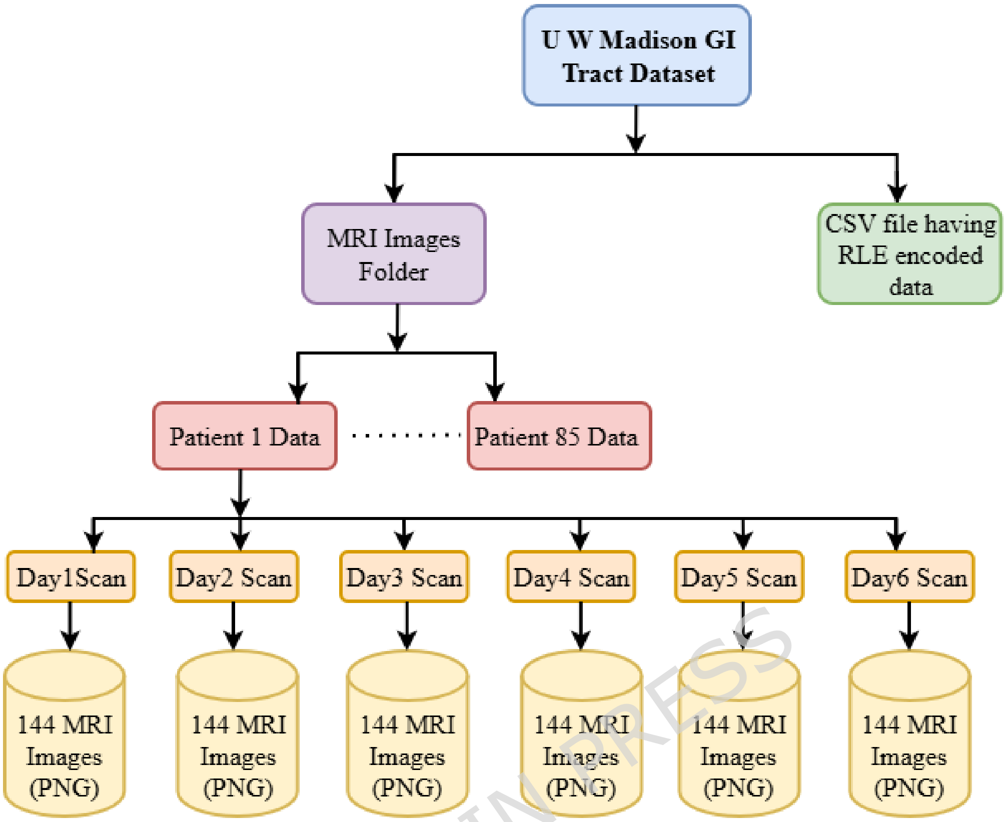


Figure 1: UW-Madison GI Tract Dataset

Figure 1 depicts dataset configurations that comprise an MRI images folder and CSV file with Run-length encoding (RLE) data. The MRI image folder consists of 85 folders for patients 1 to 85. The scans are performed over one to six days and are stored in folders 1 to 6. The MRI scans of the patients are stored in subfolders inside the day folder. The scans have variable dimensions such as length and width. Some images are rectangular, whereas the remaining images are square. To make the image size same all the images are resized to 224x224. The ground truth mask is in CSV format, where the segmented portions are shown in RLE form. RLE is a lossless image compression method that works well for images containing many homogenous regions, such as computer graphics or scanned texts. Here, RLE is used for encoding ground truth masks from MRI scans of GI tract. For instance, Figure 2(a) displays the scanned image of a 56-number slice of patient ID 111. The decoded RLE for the large intestine, the small intestine, and the stomach is shown in Figures 2(b), 2(c), and 2(d) for the same slice.

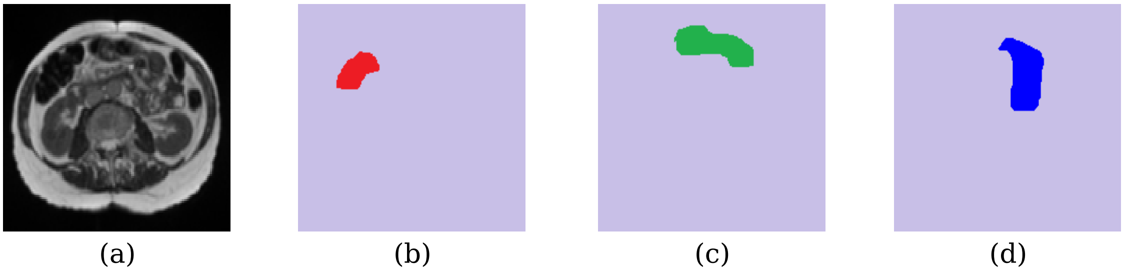


Figure 2: UW-Madison Dataset (a) Input Image, (b) RLE Decoded for Large Bowel, (c) RLE Decoded for Small Bowel, and (d) RLE Decoded for Stomach

3.1 Data Augmentation

Data augmentation is used to generate diversity in the images to increase the segmentation performance and combat overfitting. Figure 3 shows the results of data augmentation for two gastrointestinal medical images. One row represents one image, with the original and five augmentations: horizontal flip, vertical flip, rotation, brightness adjustment, and elastic transformation. These augmentations improve robustness of models by incorporating spatial and intensity variations while maintaining anatomical structures. The augmentations used were random horizontal and vertical flipping to mimic diversity in the direction of imaging, random rotations between ±15 degrees to correct for variation in patient positioning, and brightness and contrast changes to reflect diversity in imaging conditions. Elastic deformations were also employed to introduce smooth, localized distortions that maintain anatomical structure without compromising the network's capability to discern fine differences in organ shape and texture. These augmentations were used during training so that in each epoch, the model was exposed to a large number of different transformed samples.

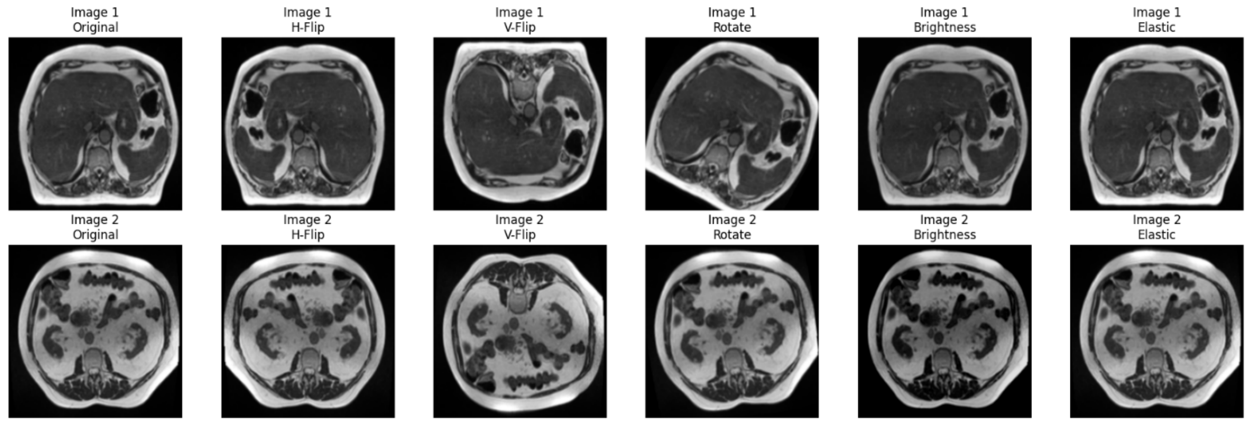


Figure 3: Augmented visualization of two gastrointestinal images showing original, horizontal flip, vertical flip, rotation, brightness adjustment, and elastic transformation.

3.2 Dataset Splitting

The UW-Madison GI Tract dataset used in this study comprises a total of 38,496 MRI images collected from 85 unique patients, each having between 1 to 6 imaging days, where every “day” folder corresponds to a separate imaging session of the same patient captured under consistent anatomical orientation and resolution. Each MRI image contains three organ annotations small intestine, large intestine, and stomach provided in RLE format within a CSV file, resulting in a total of 115,488 annotations (14,085 for the large intestine, 11,201 for the small intestine, 8,627 for the stomach, and 81,575 blank cases without organ presence). The ground truth masks were decoded from these RLE annotations for model training and evaluation.

Table 1: Dataset Splitting

Category	**Total Annotations**	**Training (80%)**	**Testing (20%)**
Large Intestine	14085	11989	2816
Small Intestine	11201	8961	2240
Stomach	8627	6903	1724

Blank	81575	65261	16314
Total	115488	93114	22374

To ensure independence between the training and testing sets and to eliminate any possibility of data leakage, the dataset was divided strictly at the patient level rather than at the patient-day or slice level. This patient-exclusive split ensured that all MRI slices corresponding to a single patient across all imaging days and sessions were allocated entirely to either the training or testing set, but never both. Consequently, the model never encountered any slices from the same patient during both training and evaluation, preventing memorization of patient-specific anatomical structures that could artificially inflate performance. The final data split consisted of 68 patients (approximately 80%) in the training set and 17 patients (approximately 20%) in the testing set. Within each subset, the internal directory structure (patient folders → day subfolders → slice images) was maintained to preserve the relationship between patient and imaging days. In total, the dataset contained 326 imaging days, with an average of 3.83 ± 1.2 days per patient, distributed as 262 imaging days (68 patients) for training and 64 imaging days (17 patients) for testing. Table 1 presents the corresponding distribution of annotated cases, including 11,989 training and 2,816 testing cases for the large intestine, 8,961 training and 2,240 testing cases for the small intestine, 6,903 training and 1,724 testing cases for the stomach, and 65,261 training and 16,314 testing blank cases. This proportional division maintained the diversity of anatomical and temporal variations across subsets. Since each "day" folder represents scans of the same patient with only minor physiological differences, splitting at the patient-day level could have led to feature leakage. Therefore, the patient-level split provides a more robust, unbiased, and generalizable evaluation of model performance while ensuring that no overlapping anatomical information is shared between the training and testing sets.

4. Proposed Multi-Level Attention DeepLabV3+ Model

The proposed work introduces a Multi-level Attention DeepLab V3+ [37] model specifically designed for the automatic segmentation of GI tract organs namely, small intestine, large intestine, and stomach using MRI scans. This model is particularly designed to overcome the limitations in GI organ segmentation through combining strong encoding, multi-scale context aggregation, attention-based feature improvement, and high-resolution decoding schemes. The model starts with an EfficientNet B0 encoder, which effectively extracts deep hierarchical features from the input MRI images. EfficientNet B0 has been chosen because it is optimized compound scaling strategy, keeping balance between depth, width, and resolution to achieve better performance with less computational cost. As the image passes through the encoder, its spatial dimension is progressively reduced step by step and it takes in more abstracted semantic representations. Upon leaving the encoder, the feature maps that have been extracted are processed using the ASPP module, the core module of the DeepLab V3+ network. There are five branches for the ASPP module: a 1×1 convolutional layer, three atrous convolutions of size 3×3 with dilation rate 6, 12, and 18 respectively, and global average pooling operation. Each branch is batch normalized and Rectified Linear Unit (ReLU) activated to stabilize and nonlinearize the features. With these multiple receptive fields, the network is able to learn both local details and the larger context, which is particularly useful in dealing with the intricate anatomical variations in the GI tract. To further enhance the semantic richness of the features learned at each scale, we introduce a channel-based attention mechanism within every ASPP branch. This attention mechanism computes channel-wise significant weights via global average pooling, and sigmoid activation function. Such learned weights enable the model to focus on more informative channels and dampen less informative channels, thus enhancing the quality of multi-scale features prior to concatenation. After concatenation, the multi-scale attention-enhanced feature maps are additionally refined using a spatial attention mechanism, which enables the model to focus on the most informative parts in the spatial domain. The objective here is to localize the anatomical borders of organs more accurately by taking into consideration where the most discriminative features are located. The ASPP module output is then fed to the decoder. The decoder uses skip connections from the encoder to reinstate fine spatial details, which tend to be lost during downsampling. Each decoding block has batch normalization, ReLU activation, 1×1 convolutions for channel alignment, and upsampling operations to gradually boost feature map resolution progressively. These blocks are constructed to combine contextual knowledge from the ASPP module and fine-grained information from previous encoding layers. During the decoding process, attention-enhanced features from various levels are combined such that the segmentation output preserves both high-level semantic correctness and low-level structural accuracy. The final prediction layer generates segmentation masks that outline the stomach, small and large intestines regions with improved anatomical structure. The complete architecture depicted in Figure 4 demonstrates how the combuned use of multi-level attention, EfficientNet encoding, ASPP-based multi-scale feature extraction, and a robust decoder lead to a highly accurate and computationally efficient GI tract segmentation pipeline.

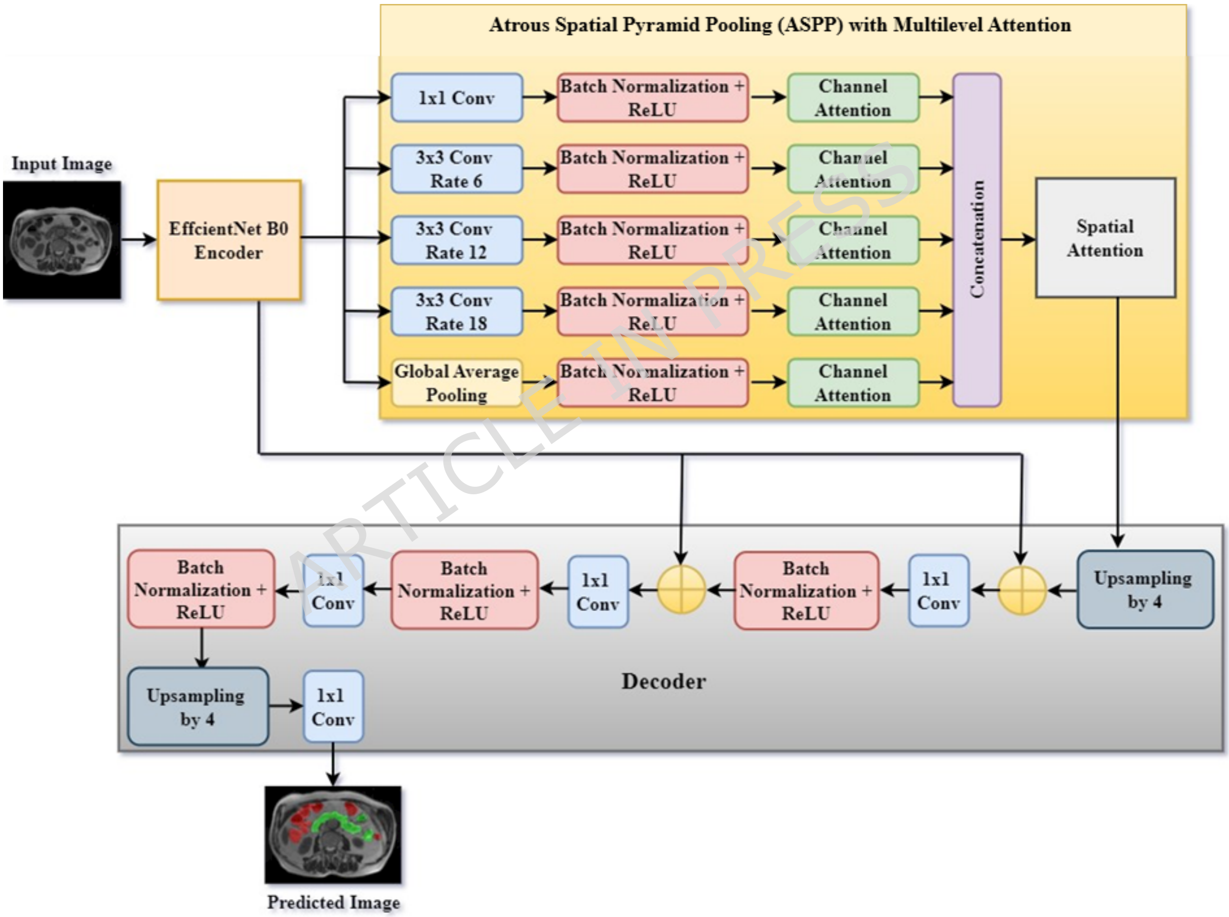


Figure 4: Architecture of Proposed Multi-level Attention DeepLab V3+ Model

The model comprises different convolution blocks that combine convolution, batch normalization, and activation layers. The convolutional layers (e.g., Conv1, Conv2_block1) apply convolutional operations to the input feature maps using learnable filters as shown in equation (1):

$$\text{Output}(x,y) = \sum_{i-1}^{h} \sum_{j-1}^{w} \sum_{k-1}^{c} \text{input}(x+i, y+j, k) * \text{Filter}(i,j,k) + \text{Bias} \quad (1)$$

Where Output(x,y) is the resulting feature map value at the spatial location (x,y) after applying the convolution operation, input(x + i,y + j,k), is the value of the input feature map at location (x+i,y+j) in the k-th channel. This represents a patch of the input data that the filter is sliding over. Filter (i, j, k) is the learnable weight in the filter (or kernel) at position (i,j) for the k-th channel. This defines how the filter interacts with the input data. Bias is a learnable bias term added after the multiplication and summation, helping the model to better fit the data. h and w is the height and width of the filter. These define the spatial dimensions of the convolutional kernel. C is the input channels (also called depth). i, j, k are Indices used for iterating over the height, width, and channels of the filter, respectively.

The Pooling layers (e.g., average_pooling2d) downsample the input feature maps to reduce spatial dimensions. The output of a pooling is shown in equation (2):

$$\text{Output }(x,y,k) = \max_{i=1}^{h} \max_{j=1}^{w} \text{FeatureMap}(x+i, y+j, k) \quad (2)$$

Where, Output (x, y, k) is the result of the max pooling operation at position (x,y) in the k-th channel. FeatureMap (x + i, y + j, k) is the value from the output of the convolution layer at spatial location (x+i,y+j) in the k-th channel. Pooling operates on these convolutional feature maps. h and w is the height and width of the pooling window. i, j are the indices that slide over the pooling window, and k is the channel index.

4.1 EfficientNet B0 as Encoder

EfficientNet B0 [38] is a CNN architecture designed to balance high model performance with computational efficiency by compound scaling. Traditional CNN architectures typically scale only one dimension of the model at a time either depth, width, or resolution. In contrast, EfficientNet B0 employs a compound coefficient that uniformly scales all three dimensions in a balanced manner. This principle scaling method allows the model to reach higher accuracy with a substantial decrease in parameters and FLOPs. The backbone of EfficientNet B0 is the Mobile Inverted Bottleneck Convolution (MBConv) block, which is basic block of the network. In contrast to normal convolutions, MBConv utilizes a bottleneck pattern that initially broadens the channel number using a pointwise 1×1 convolution (expansion layer), utilizes a depthwise separable convolution to reduce computational costs while utilizing efficient spatial filtering, and then projects the output back into a reduced-dimensional space using another 1×1 convolution (projection layer). This reverse pattern aids in lowering computational cost without losing necessary spatial and semantic information. The application of depthwise separable convolutions makes it possible for EfficientNet B0 to heavily reduce parameters and computations versus typical convolutions.

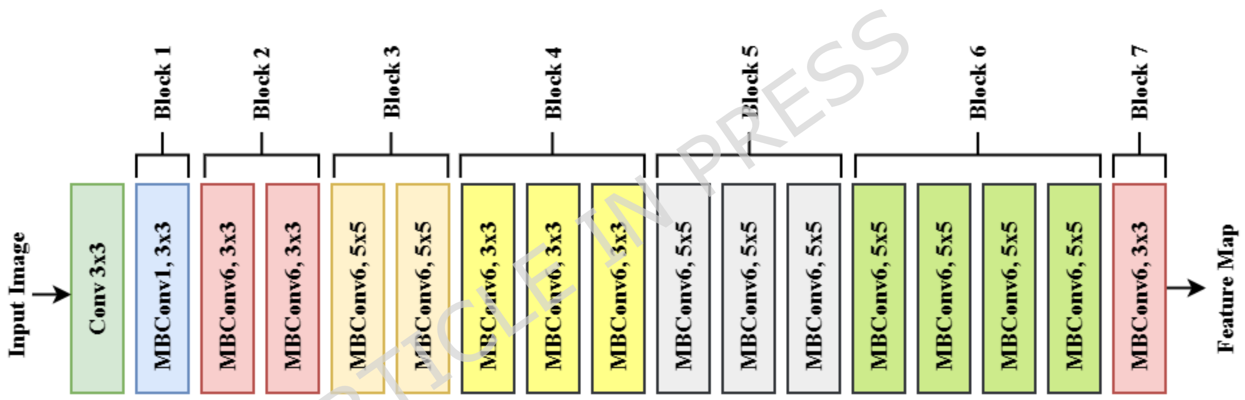


Figure 5: Architecture of EfficientNet B0

As shown in Figure 5, the architecture of EfficientNet B0 is divided into seven consecutive blocks with each block composed of several MBConv layers. These blocks employ a combination of 3×3 and 5×5 kernel sizes to detect diverse receptive fields. Expansion factors, strides, and layer counts are different across blocks, allowing the model to evolve to greater depth and complexity. The structure starts with an initial 3×3 convolutional layer for the extraction of low-level features followed by the stacked MBConv blocks that progressively process the input. With the advancing image in the architecture, spatial resolution is progressively diminished by strided convolutions, but the feature channel count grows, enabling the network to acquire deep, high-level abstractions required for sophisticated tasks like GI organ segmentation. The model is able to capture fine-grained edge details in the initial layers and rich semantic features in deeper layers through multi-resolution encoding. The output feature maps generated by EfficientNet B0 refined by MBConv blocks and enhanced through channel attention serve as the input to the ASPP module, where multi-scale semantic information is further extracted.

4.2 ASPP Module with Multi-level Attention

To enhance segmentation performance while preserving spatial resolution, the ASPP module in the proposed attention DeepLab V3+ architecture integrates both atrous (dilated) convolutions and a multi-level attention mechanism. Atrous convolution introduces gaps between kernel elements, enabling convolutional kernels to cover more receptive field without increasing the parameters or computation. This technique is particularly effective in semantic segmentation, where objects and anatomical structures may appear at varying scales. The mathematical formulation of atrous convolution is shown in equation (3)

$$z[i] = \sum_{n} a(i + r.n)f[n] \quad (3)$$

Where z is the output feature map, i shows the spatial domain location of z, a is the input feature map, r is the atrous convolution rate, and f is the convolution filter. The output from these levels is concatenated and sent to the following network block named as Multi-level attention mechanism in DeepLab V3+.

Figure 6 shows the detailed architecture of the ASPP module integrated with a channel-based attention mechanism. The ASPP consists of five parallel branches: A standard 1×1 convolution, Three 3×3 convolutions with atrous rates of 6, 12, and 18, respectively, and A global average pooling branch. Each branch is followed by batch normalization and ReLU activation to normalize and activate the outputs. To further improve the discriminative power of the features extracted at each scale, we introduce channel-based attention into each branch. This attention mechanism starts by performing global average pooling on every feature map to create a condensed descriptor that captures the global context. These descriptors are fed into a set of two fully connected layers and activated through a sigmoid activation function to produce channel-wise attention weights. Following attention refinement, all the five branches' outputs are concatenated. A spatial attention module is used with the combined feature map to assist the model in localizing spatial areas important for segmentation. The output of the ASPP module is a multi-scale, attentioned representation that is transmitted to the decoder.

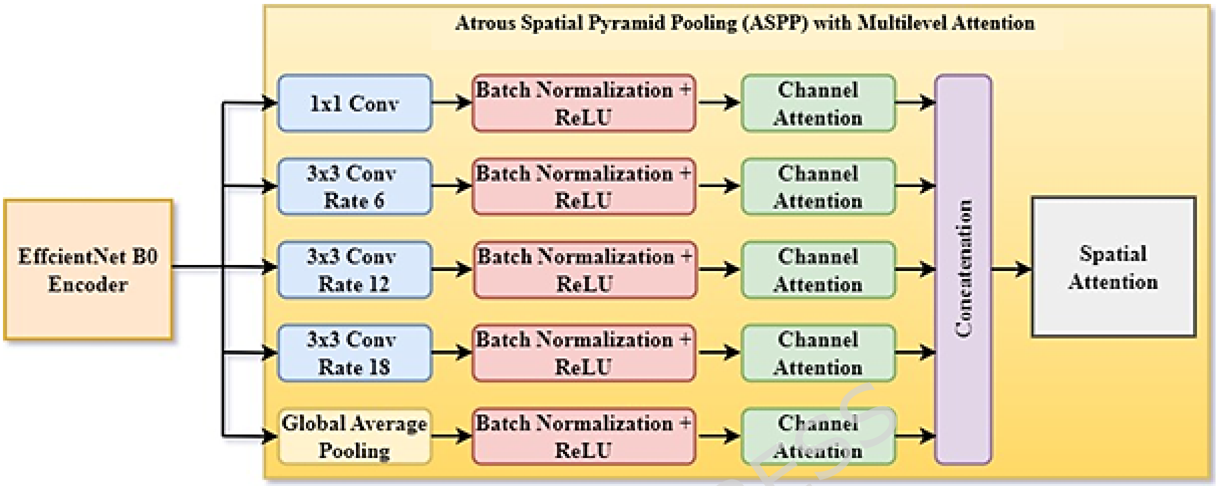


Figure 6: Architecture of the proposed Atrous Spatial Pyramid Pooling (ASPP) module.

To further enhance segmentation precision, particularly in intricate anatomical areas, we integrate a multi-level channel attention mechanism that takes effect at various stages of the network: low-level, mid-level, and high-level features. Motivated by Squeeze-and-Excitation networks [39], the mechanism allows the model to learn inter-channel dependencies and context relationships among the feature maps at different depths.

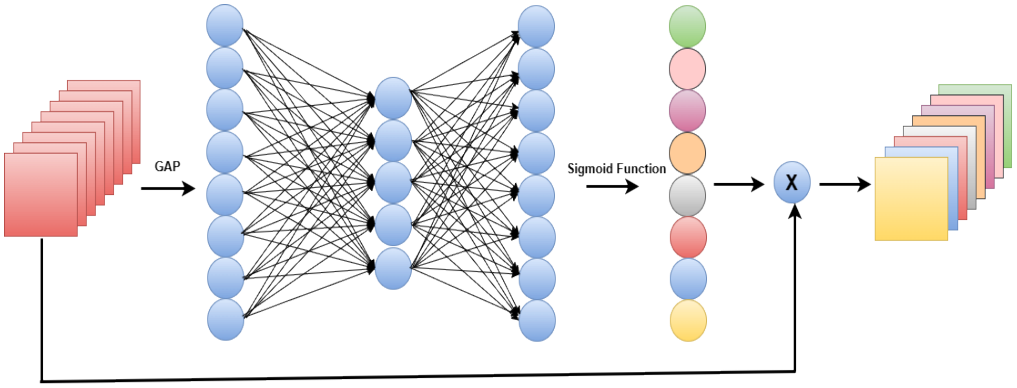


Figure 7: Channel-wise attention mechanism applied to ASPP branches.

As demonstrated in Figure 7, the channel attention process starts with global average pooling over all the channels of a feature map. The operation yields a descriptor vector that captures the global significance of each channel. The descriptors obtained after pooling are utilized to provide as input to two fully connected layers and a sigmoid activation function to produce channel-wise attention weights. These weights are utilized to modulate the original feature maps using element-wise multiplication as shown in equation (4):

$$Wn = \frac{1}{LxB}\sum_{i}^{l}\sum_{j}^{b} Yn(i,j) \tag{4}$$

Where, $Yn(i,j)$ is the pixel at position n-th channel, L×B is the spatial size of the channel, Wn is the global average value (attention score) for the n-th channel.

This mechanism enables the network to selectively highlight informative feature channels and downweight redundant ones, improving its attention to semantically significant structures like the small intestine, large intestine, and stomach. The multi-level attention mechanism allows features of various semantic levels to be adaptively tuned before passing them to the decoder. It encodes hierarchical dependencies and enables strong feature fusion through attention-refined alignment of features from low, mid, and high-level layers. Such features are concatenated and passed to a scale attention module that captures relationships between scales as well. The model integrates channel-wise and multi-level attention and greatly enhances its capacity to identify organs of different shapes and sizes with delicate boundaries. This attention-enhanced ASPP output enables more accurate, context-aware segmentation downstream in the decoder.

4.3 Decoder

The decoder within the suggested Attention DeepLab V3+ architecture is pivotal in reconstructing high-resolution mask from the compressed, high-level feature representations generated by the encoder and the ASPP block.

While semantic features are progressively downsampling in the encoder, spatial resolution is drastically decreased. To overcome this, the decoder employs progressive upsampling while combining multi-scale, attention-weighted features to recover fine spatial details.

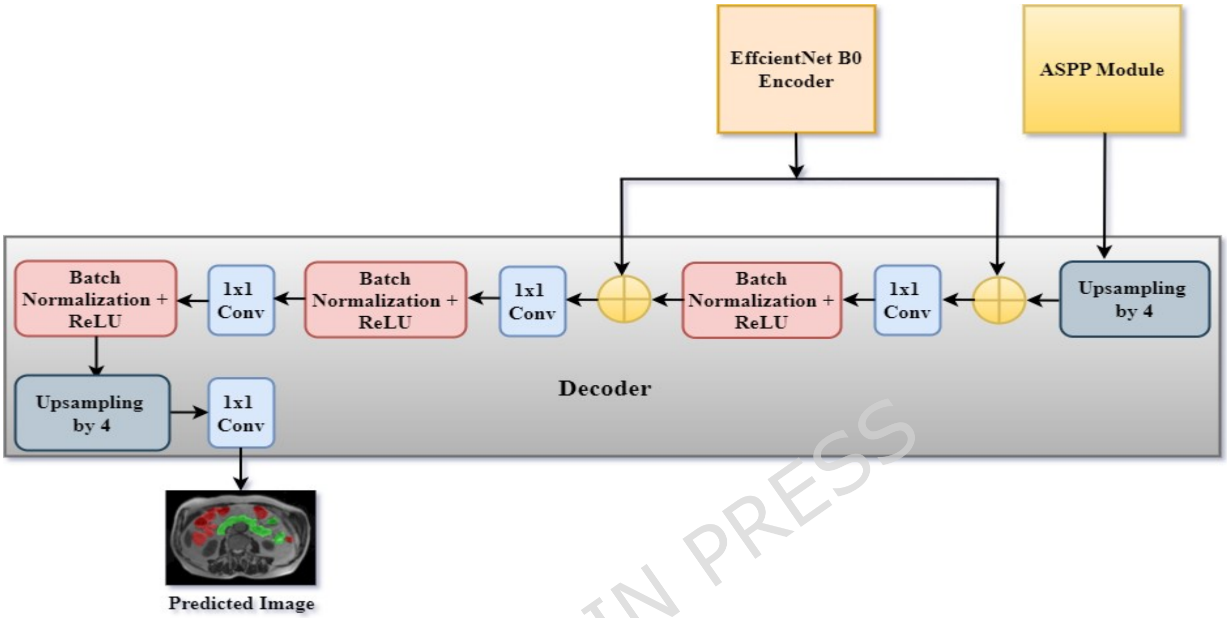


Figure 8: The decoder architecture of the proposed Attention DeepLab V3+ model.

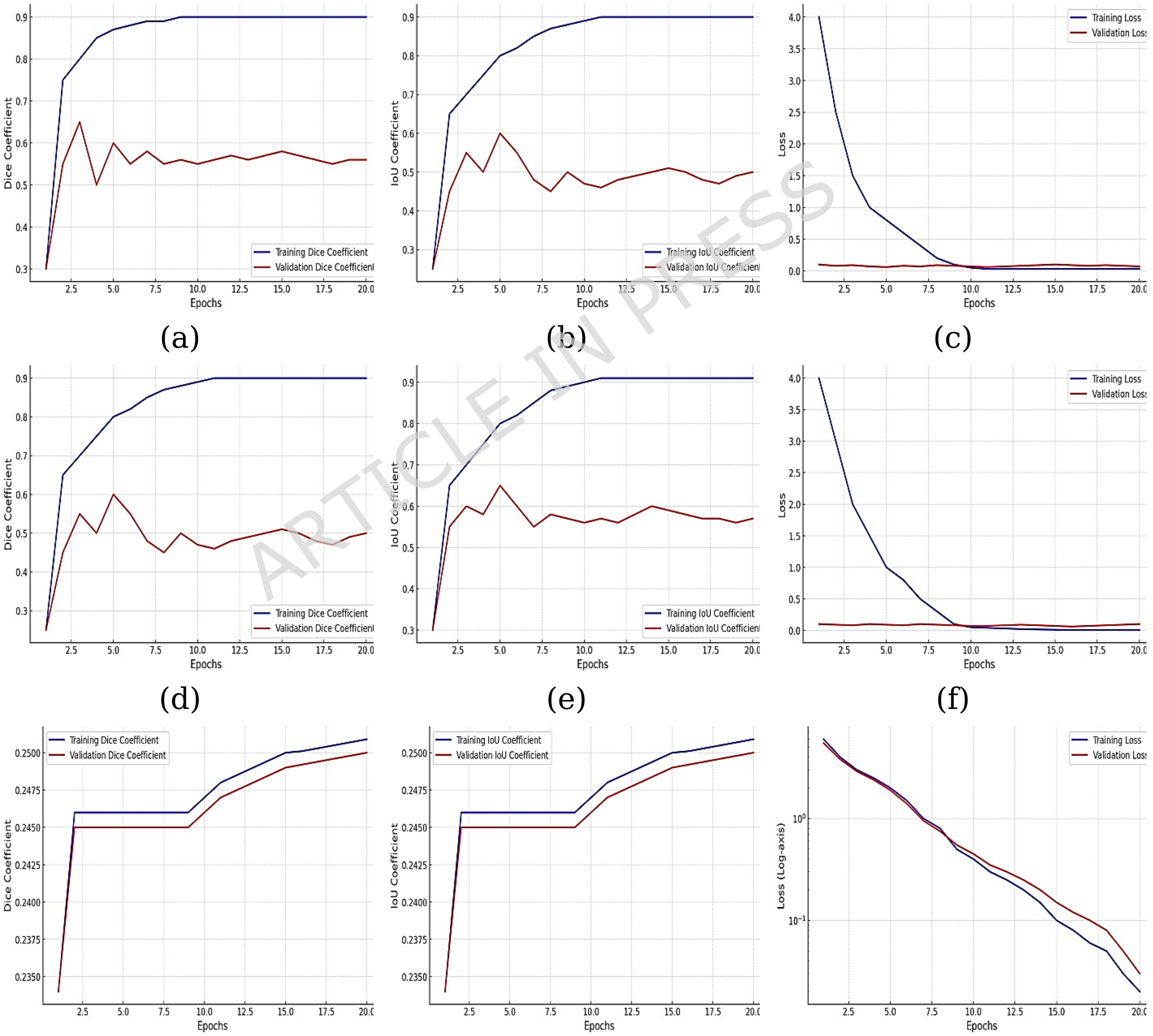
As shown in Figure 8, the decoder starts by taking multi-scale feature maps from both the ASPP module and the EfficientNet B0 encoder through skip connections. The skip connections act as routes for directly transmitting low-level features to the decoder so it can regain fine spatial details, which tend to be lost in deeper layers of the network. The ASPP module feature maps are then upscaled by a factor of 4 to be comparable with the spatial dimensions of features in the previous stages of the encoder. Alignment allows fusion of structural (low-level) and semantic (high-level) information. The fusion is done through a series of operations: Batch normalization stabilizes training and normalizes feature distributions. Activation using ReLU injects non-linearity to enhance the learning ability of the network. 1×1 convolutions are employed for matching channel dimensions and further detailing the feature representation per stage. Through conducting convolution followed by upsampling per stage, the decoder progressively reconstructs the feature maps to the original input resolution. This produces a high-resolution segmentation map that precisely outlines target anatomical structures such as the small bowel, large bowel, and stomach. Inclusion of multi-level attention mechanisms in ASPP output ensures decoder pays attention to semantically important and spatially significant regions. It improves the network's performance in delineating finer structural borders, particularly in overlapping and confusing gastrointestinal anatomy areas. The final output is a segmented prediction map which identifies regions of interest with room for identifying relevant patterns, anatomical variations, or disease affected areas.

5 Results and Discussions

This study introduced a DeepLab V3+ model integrating EfficientNet B0, ASPP and multilevel attention to segment the GI tract organ. All experiments were performed on a workstation with an NVIDIA RTX A5000 GPU (24 GB VRAM), Intel Core i7-11700 CPU, and 32 GB RAM, under Windows 10 (64-bit) operating system with CUDA 11.2 and cuDNN 8.1. The suggested Multi-Level Attention DeepLab V3+ model with EfficientNet-B0 encoder has around 8.3 million trainable parameters and has ≈ 21.7 FLOPs per forward pass, providing the best balance between accuracy and computational cost. The model was trained for 30 epochs with a batch size of 16, RMSprop optimizer, and an initial learning rate of 0.0001. Categorical cross-entropy was used as the loss function for multi-class segmentation. The overall training time was around 4.5 hours, and each epoch took around 540 seconds. In inference, the model obtained an average inference time of ≈ 31 milliseconds for every 224 × 224 MRI slice, which is a throughput rate of ~32 frames per second (FPS). The highest GPU memory usage for inference was 2.4 GB at batch size 16, proving that the model is light and computationally viable for near real-time clinical use. Experiments were run on Python 3.8 using TensorFlow and Keras libraries for absolute reproducibility of the results.

5.1 Analysis based on Different Optimizers

The model proposed in this research has been tested with three optimizers with the rest of the hyperparameters remaining the same. Optimizers employed in this research are Adaptive Moment Estimation (Adam) [40], RMS prop [41], and Stochastic Gradient Descent (SGD) [42-43]. Various ways of optimization are applied as each of them has its strengths and weaknesses and can be better or worse suited for different models. The optimizers are executed for ten epochs and employ two cross-folds [44]. Figure 9 demonstrates the relative performance of these optimizers using Dice coefficient, IoU, and loss. The graphs show that the curves of Dice coefficient and IoU reach their best values when the RMSprop optimization is used, which suggests better precision in segmenting organs in the GI tract when compared to using Adam and SGD optimizers. Meanwhile, the loss curve shows its smallest path when RMSprop is used, revealing faster convergence and reduced error in training as well as validation stages. This implies that RMSprop performs better than Adam and SGD in optimizing the model's parameters in order to obtain the best segmentation outcome.



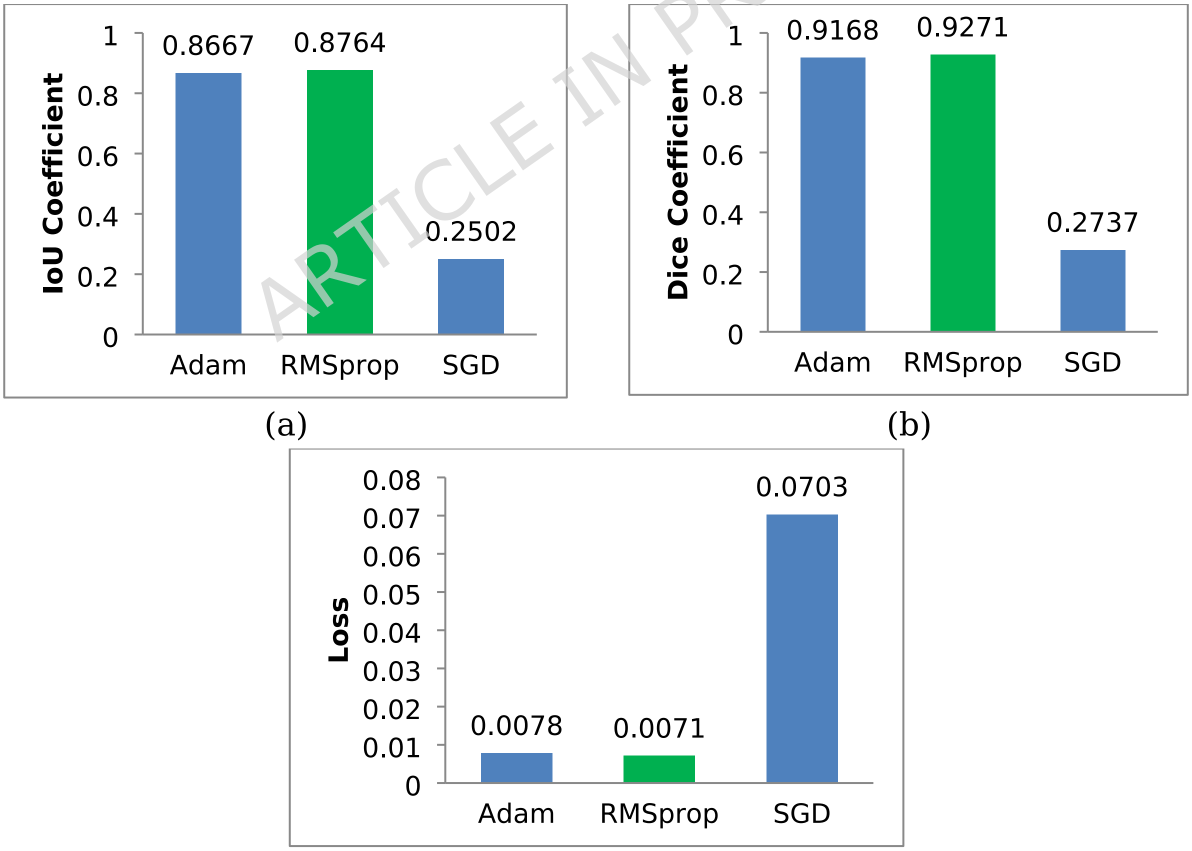
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 9: Graphical Analysis of Dice, IoU, and Loss using Different Optimizers: For Adam- (a) Dice (b) IoU, (c) Loss, For RMSprop- (d) Dice (e) IoU, (f) Loss, For SGD- (g) Dice, (h) IoU, (i) Loss

Figure 10 presents the values of the performance parameters such as IoU coefficient shown in figure 10(a), Dice Coefficient shown in figure 10(b) and Loss shown in figure 10(c) for optimizers Adam, RMSprop, and SGD. RMSprop is the best optimizer and delivers the highest Dice coefficient (0.9271), the lowest Loss (0.0071), and the highest IoU (0.8764) of the three. Adam is very close to competitive values in all three categories, signifying its effectiveness in identifying semantic segmentation trends. Conversely, SGD is far behind with the lowest Dice coefficient (0.2737), highest Loss (0.0703), and lowest IoU (0.2502). The significant performance metrics gap demonstrates the difference that the choice of optimizer makes in the proposed model to define semantic regions in the current task. From Figure 10, the proposed attention DeepLab V3+ using the RMSprop optimizer is seen performing better than Adam and SGD.



(a)

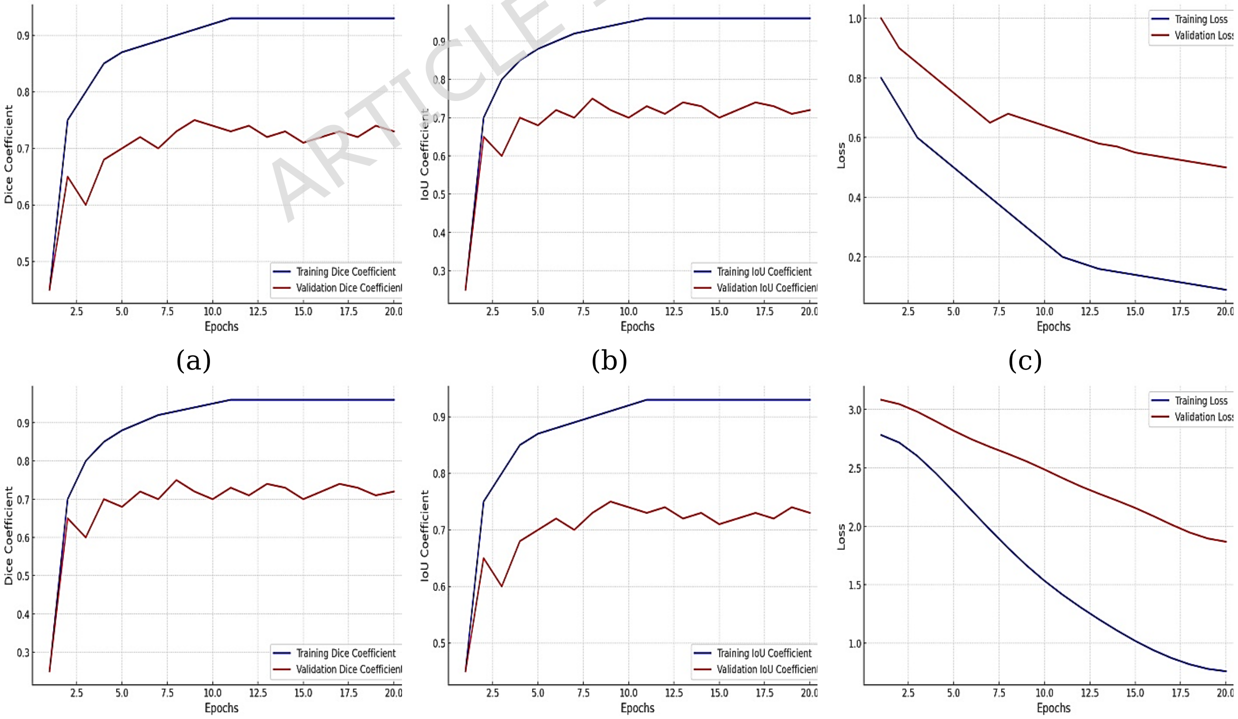
(b)

(c)

Figure 10: Performance Parameters using Different Optimizers (a) IoU Coefficient Comparison, (b) Dice Coefficient Comparison, and (c) Loss Comparison

5.2 Analysis based on the Number of Folds

The proposed method has been tested with various cross-folds to evaluate results better with the other hyperparameters unchanged. In this analysis, RMSprop was used as concluded from the previous experiment. The model has been tested with 2, 4, and 8 cross-folds. In Figure 11, the performance evaluation of segmentation over varying cross-folds numbers is illustrated, with emphasis on the Dice coefficient, IoU, and loss in the context of an approach model. The plots show that the curves for Dice coefficient and IoU reach their highest values when using four cross-folds, implying maximum segmentation accuracy as compared to configurations using two and eight cross-folds. Further, the loss curve path at its minimum is when using four cross-folds, which shows better convergence and less error in both the training phase and validation phase. This denotes the usefulness of using four cross-folds in improving segmentation accuracy and reducing loss, for the improvement of the effectiveness of radiation treatment planning in GI cancer.



(a)

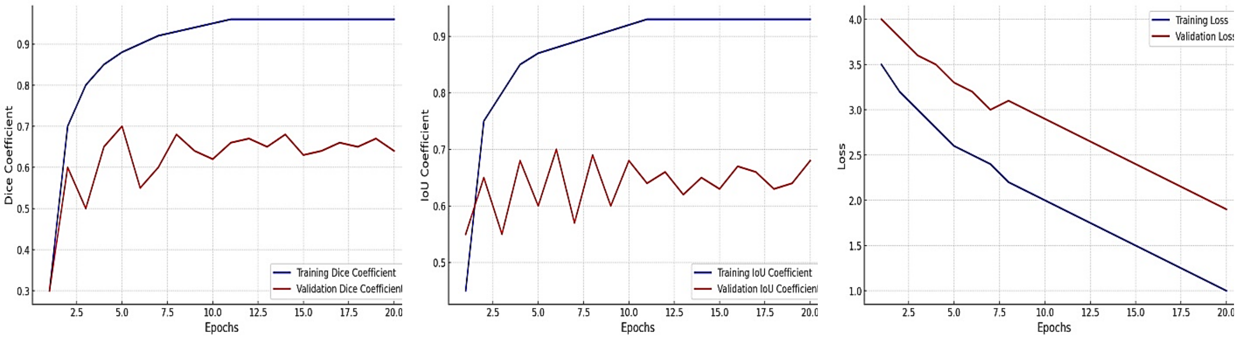
(b)

(c)

(d)

(e)

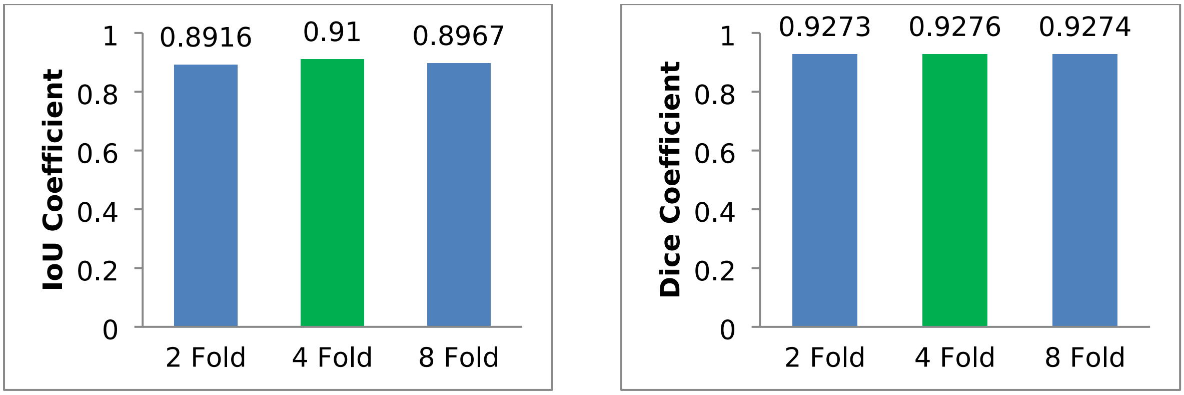
(f)



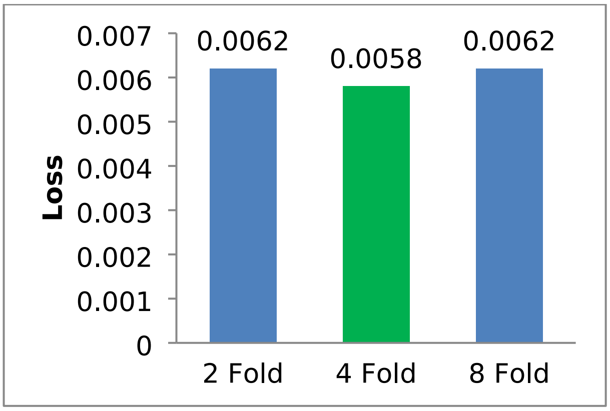
(g) (h) (i)

Figure 11: Graphical Analysis of Dice, IoU, and Loss using Different Cross Folds: For 2 Folds- (a) Dice (b) IoU (c) Loss, For 4 Folds - (d) Dice (e) IoU (f) Loss, For 8 Folds - (g) Dice, (h) IoU (i) Loss

Figure 12 demonstrates the performance parameter values such as IoU coefficient shown in figure 12(a), Dice Coefficient shown in figure 12(b) and Loss shown in figure 12(c) for various cross folds 2, 4, and 8. The outcomes indicate that the performance is consistent and similar in all folds. The Dice coefficient is constant, with values of 0.9273, 0.9276, and 0.9274 for folds 2, 4, and 8, respectively, showing a high degree of performance in detecting overlap in predicted and ground truth masks. The loss values also demonstrate a slight variation, with 0.0062, 0.0058, and 0.0062 for folds 2, 4, and 8, respectively. The IoU values also vary from 0.8916 to 0.910, demonstrating uniform and good boundary delineation of segmented regions. The segmentation performance of the model is strong and also generalizes well across different folds, demonstrating a solid and consistent performance on different sets of the dataset. Figure 12 concludes that the suggested model performed better with four folds than 2 and 8 folds.



(a) (b)

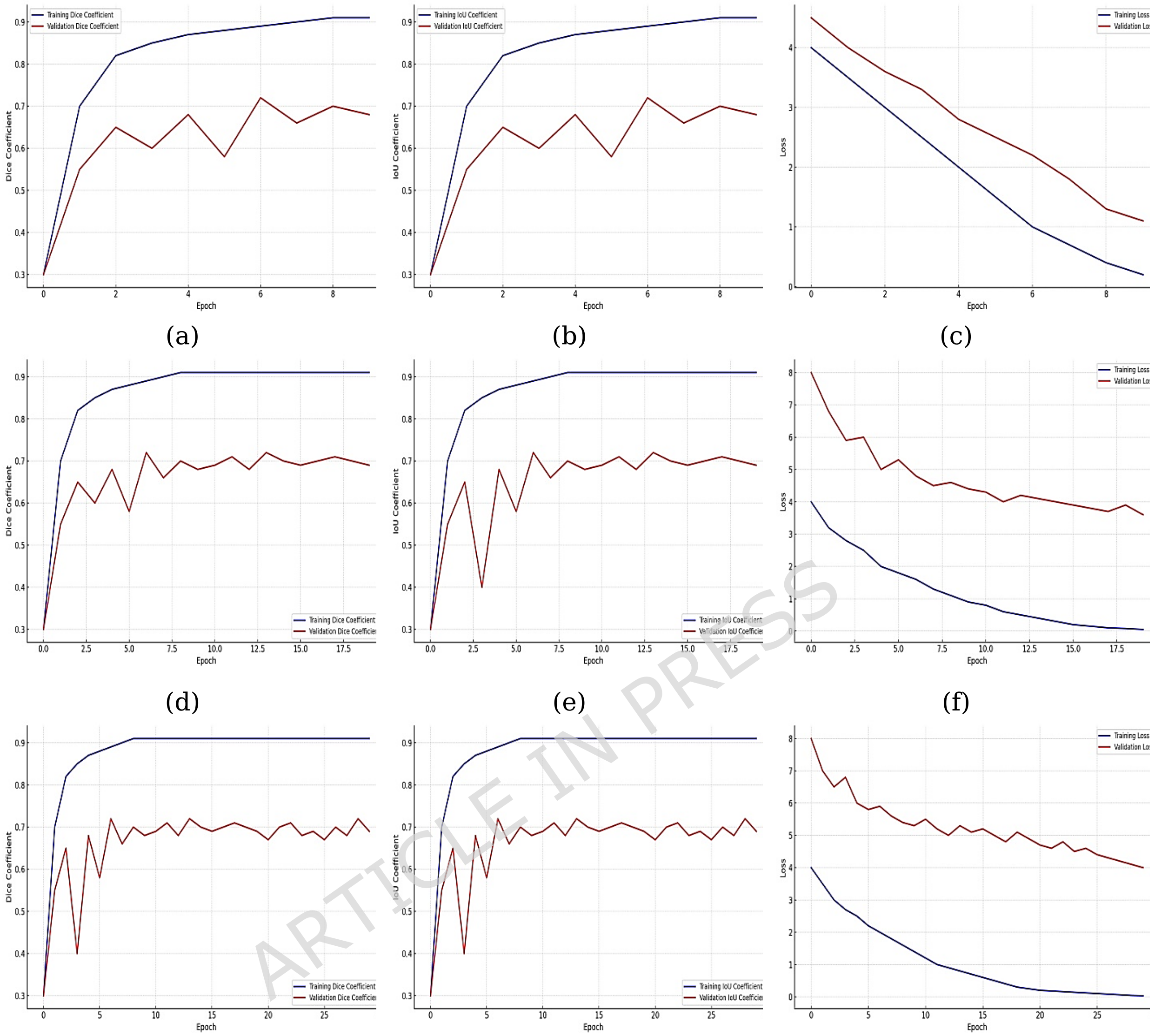


(c)

Figure 12: Performance Parameters using Different Cross Folds (a) IoU Coefficient Comparison, (b) Dice Coefficient Comparison, and (c) Loss Comparison

5.3 Analysis based on Number of Epochs

The research work proposed is trained with the number of epochs while keeping all the other hyperparameters the same. The model was trained with 10, 20, and 30 epochs with RMSprop optimizer and four folds. The dice, IoU, and loss plots are shown in Figure 13 using 10, 20, and 30 epochs. The graphs prove that the Dice coefficient and the IoU curves both attain their highest values at training the model for 30 epochs, signifying ideal segmentation accuracy in contrast with training periods of 10 and 20 epochs. The loss curve also attains its lowest point at 30 epochs, denoting better convergence and less error in the training and validation periods. In addition, the results indicate that increasing the training period beyond 30 epochs does not result in additional improvements in segmentation performance. Therefore, the choice of restricting the training time to 30 epochs is justified to help utilize resources effectively while achieving maximum segmentation accuracy for radiation therapy planning in GI cancer treatment.



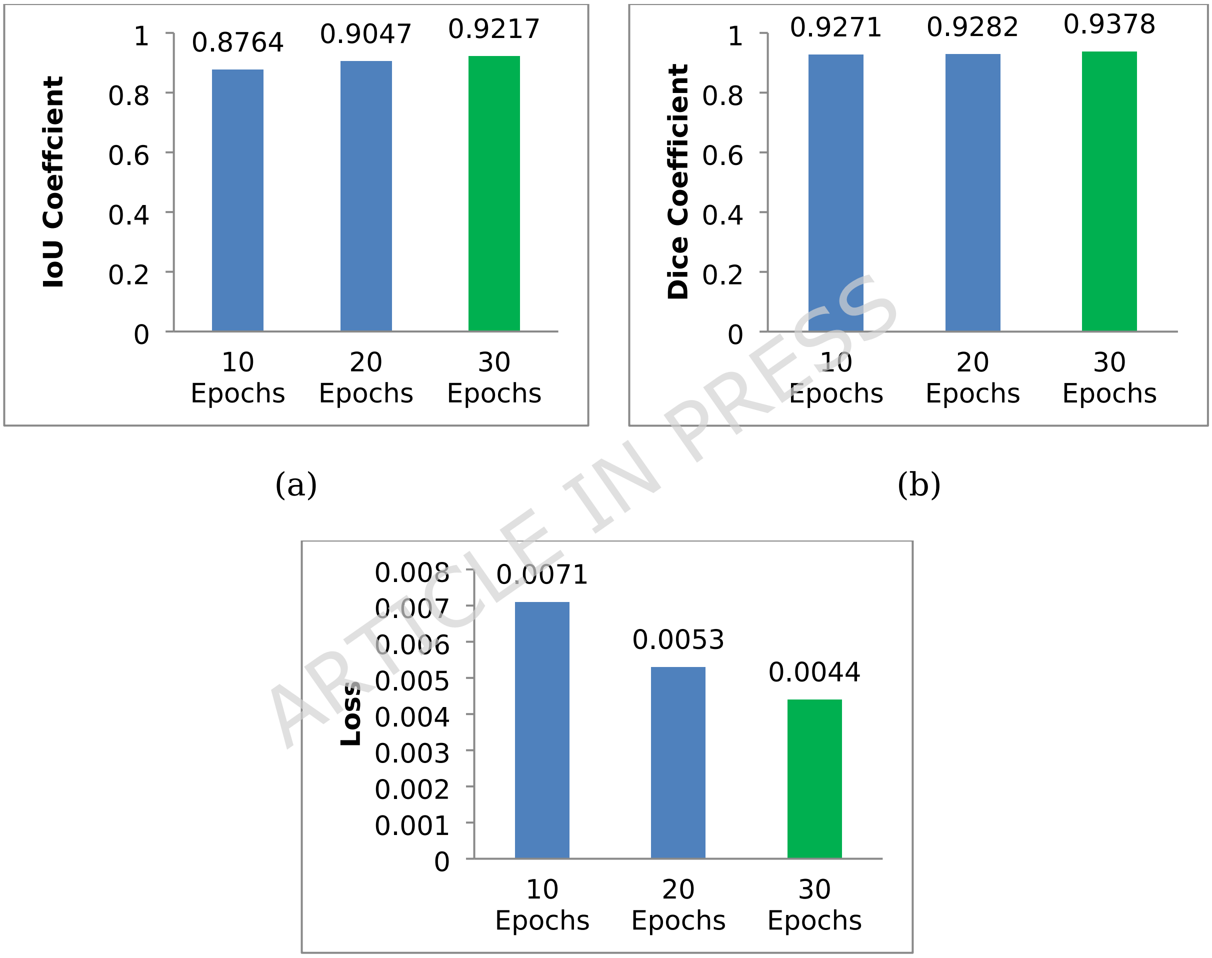
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 13: Graphical Analysis of Dice, IoU and Loss using Different Numbers of Epochs: For epochs 10- (a) Dice (b) IoU (c) Loss, For epochs 20- (d) Dice (e) IoU (f) Loss, For Epochs 30- (g) Dice, (h) IoU, (i) Loss

Figure 14 depicts the values of performance parameters such as IoU coefficient shown in figure 10(a), Dice Coefficient shown in figure 10(b) and Loss shown in figure 10(c) for various epochs 10, 20, and 30. In all parameters, the model shows a noticeable improvement with an increase in the number of training epochs. The Dice coefficient, which is a parameter of segmentation performance, improves steadily from 0.9271 for 10 epochs to 0.9378 for 30 epochs. Similarly, the Loss measure goes down from 0.0071 to 0.0044, which represents better convergence and less dissimilarity in predicted vs. actual values. The IoU, a pixel-wise measure of overlap, also increases steadily from 0.8764 to 0.9217, reflecting better segmentation boundary delineation. These findings highlight the necessity of adequate training epochs in optimizing the model's segmentation accuracy, with significant improvements in accuracy and convergence metrics when the training time increases. Figure 14 concludes that the proposed model performs better for 30 epochs.



(a)

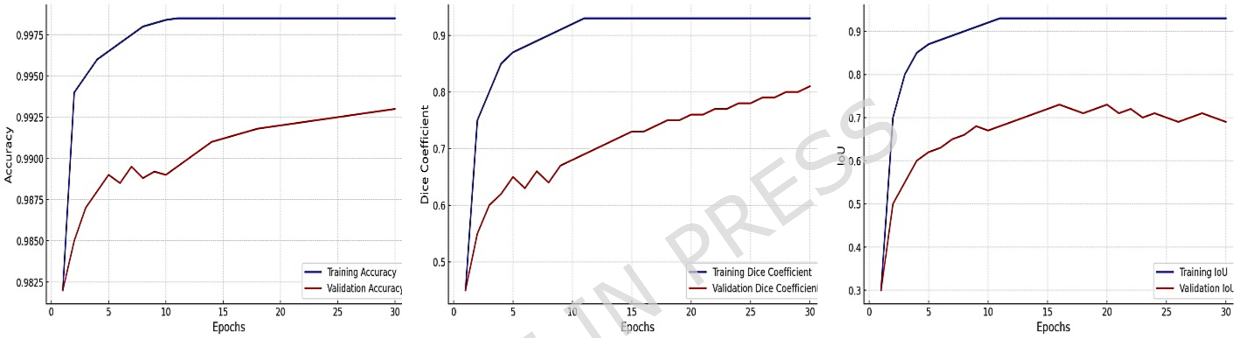
(b)

(c)

Figure 14: Performance Parameters using Different Number of Epochs (a) IoU Coefficient Comparison, (b) Dice Coefficient Comparison, and (c) Loss Comparison

5.4 Quantitative Analysis of the Proposed Model

The final optimized model undergoes a rigorous evaluation process involving 30 epochs and four cross-folds, utilizing the RMSprop optimizer. Figure 15 shows the quantitative analysis of the proposed model. Figure 15(a) illustrates the actual result while training which indicates the general overall distribution of the model. Figure 15(b) gives an idea of Dice score. For a good analysis of the segmentations, how much is it similar between predicted and the actual segmentation. It gives an estimate of overlap in between predicted and true segmentation by showing IoU in figure 15(c). Combined, this data gives in-depth information regarding the performance trend of the various indicators, providing an idea regarding the validity and reliability.



(a) (b) (c)

Figure 15: Qualitative Analysis of the Proposed Model (a) Accuracy, (b) Dice and, (c) IoU

Table 2 reports the overall and per-class performance metrics of the proposed Multi-Level Attention DeepLab V3+ model. The model achieved an accuracy of 0.9976, confirming that a high proportion of pixels in the MRI images were correctly segmented. The average Dice score of 0.9378 and IoU of 0.9217 reflect strong agreement and substantial spatial overlap between the predicted and ground-truth segmentation masks. The model loss value of 0.0044 further indicates stable convergence and low prediction error during optimization. Moreover, the class-wise outcomes reflect balanced segmentation performance over all three gastrointestinal organs, with the large intestine, small intestine, and stomach obtaining Dice scores of 94.12%, 93.47%, and 93.75% and their respective IoU values of 92.36%, 91.18%, and 91.02%. These uniform values validate that the model performs well uniformly across various anatomical areas without class bias. All of the metrics were calculated per-slice and per-class with a 0.5 threshold applied to softmax outputs and then macro-averaged over classes. The minimal numerical difference between the Dice and IoU scores stems from class-wise averaging instead of a global binary sum. Overall, these findings prove that the model proposed here attains high segmentation accuracy, robust generalization, and stable convergence, justifying its efficiency and reliability for gastrointestinal organ segmentation tasks.

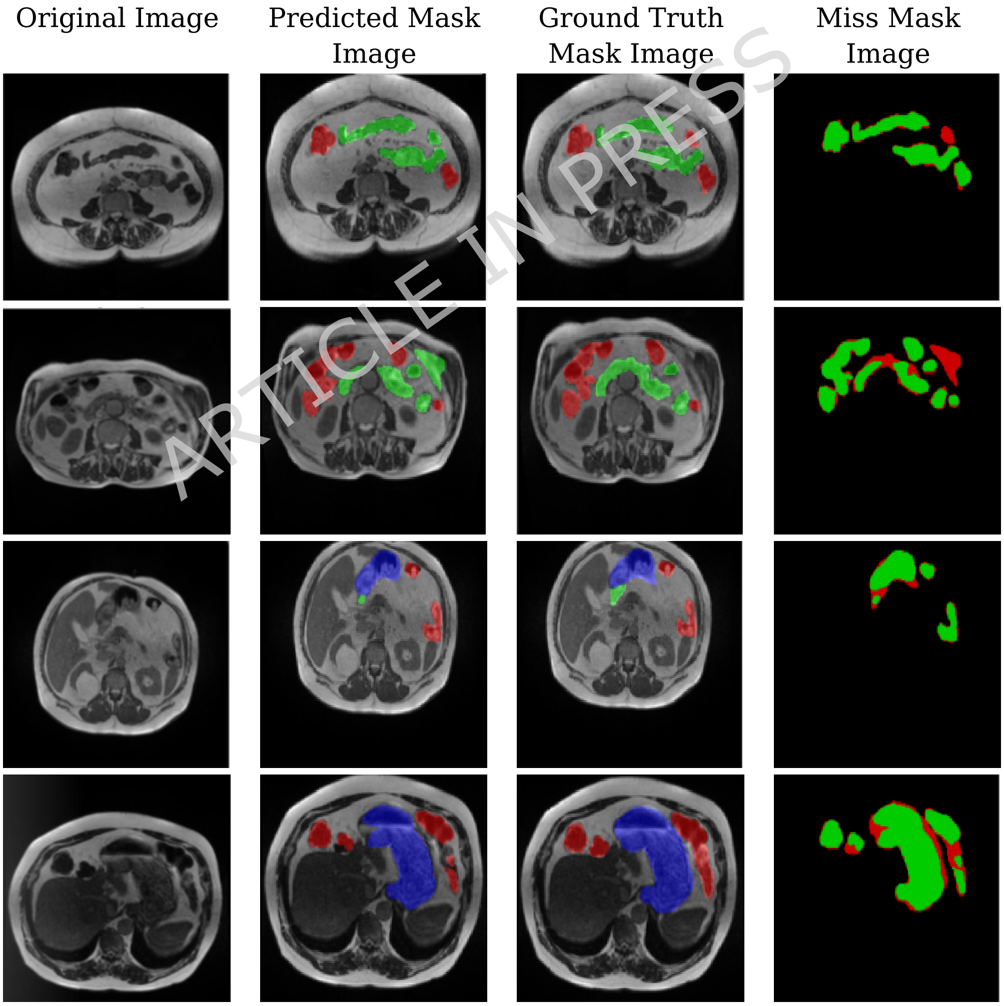
Table 2: Performance Parameters of Final Optimized Model

Class	Dice (%)	IoU (%)
Large Intestine	94.12	92.36
Small Intestine	93.47	91.18
Stomach	93.75	91.02
Mean (Macro-Average)	**93.78**	**92.17**

5.5 Qualitative Analysis of the Proposed Model

The final optimized model, which was trained for 30 epochs and used 4 cross-folds with the RMSprop optimizer. The qualitative analysis of the proposed model is presented in Figure 16, comprising the original image, predicted image, ground truth masks, and the missed mask, where the model predictions failed to agree with the ground truth. Notably, the predicted ground truth masks and the initial ground truth masks are presented in a three-color mode to provide easy visualization: red for the large bowel, green for the small bowel, and blue for the stomach. In addition, the missed mask image uses three colors: green indicates agreement or matching areas between the predicted and original masks, red indicates regions of disagreement or misprediction, and black indicates the background. Visual representation highlights the potential benefit of the proposed model in predicting the segmentation of the stomach, small intestine, and large intestine to be useful in GI cancer treatment planning using radiation therapy.

Although the overall good performance of the proposed model, some limitations were noted in certain segmentation instances. Most significant were the segmentation errors along organ boundary regions, for example, the juncture of the small and large intestine, where anatomical structures tend to overlap or resemble each other visually. In a few instances, the model wrongfully labeled low-contrast areas or did not identify organ edges when the intensity gradient was weak. Moreover, images with motion blur, noise, or artifacts yielded incomplete or fragmented segmentations. These problems arise from an amalgamation of factors, such as class imbalance due to the underrepresentation of smaller organs or narrow structures in the training set, as well as the inherent texture and intensity similarity between gastrointestinal organs. Future enhancements may involve the application of edge-aware or boundary refinement loss functions, adaptive weighting of classes, and more drastic data augmentation to improve robustness. Adding extra layers of attention or multi-modal imaging data can further help in resolving ambiguities between overlapping structures and enhance overall segmentation accuracy.



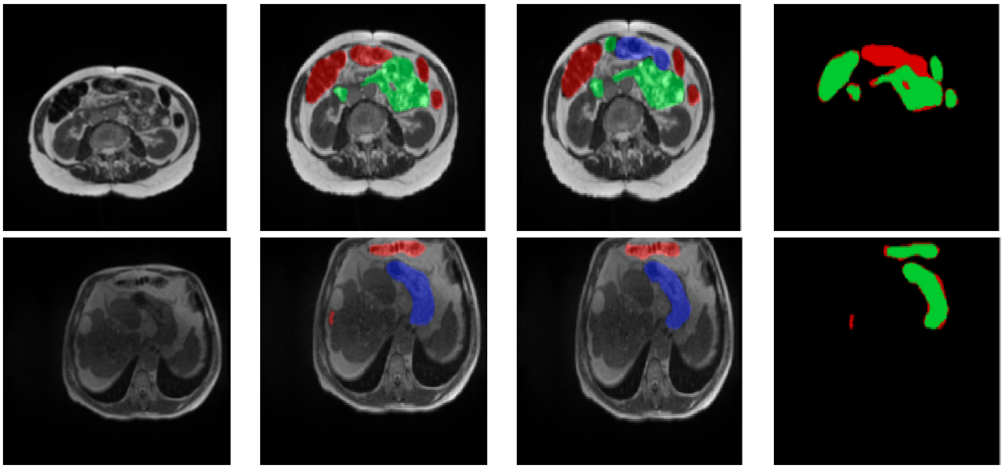


Figure 16: Visualization of Results (Here red color represents the large bowel, the green color shows the small bowel, and the blue color represents the stomach)

The model was trained solely on the UW-Madison dataset, which includes only healthy anatomical regions, limiting its generalizability to pathological cases or multi-institutional datasets compared to our previous model [35], which used a basic UNet architecture without attention mechanisms, the proposed Multi-level Attention DeepLab V3+ framework demonstrates substantial improvements. In particular, our former method obtained a Dice score of 0.8984% and an IoU of 0.8697%, while the present model obtained a Dice score of 93.78% and an IoU of 92.17%. This is because the current model utilized an EfficientNet B0 encoder for better feature extraction, an ASPP module for capturing multi-scale context, and a multi-level attention mechanism for feature refinement over semantic layers. These developments allow for improved boundary specification and structural precision in gastrointestinal organ segmentation. There are still issues in proper segmentation of overlapping or low-contrast structures, especially in complex anatomy regions. While EfficientNet B0 provides an effective balance between performance and computational costs, real-time clinical application and robustness under diverse environments remain unexplored. Furthermore, potential biases in the dataset and class imbalance can compromise segmentation accuracy for structures with low representation. Future research will aim to generalize the model to multi-modal and disease datasets, adding transformer-based modules in order to learn more contextual dimensions, and using domain adaptation or semi-supervised learning methods in order to increase robustness. Clinical testing and integration into real-time clinical workflows will also be critical to making practical impact.

5.6 Ablation Study of the Proposed Model

An extensive ablation study was performed to thoroughly assess the contribution of various architectural components and different attention mechanisms in the proposed model; the results are of the ablation study are summarized in Table 3. The ablation is started with the baseline DeepLab V3+ architecture using a standard ResNet-50 encoder without any attention mechanism, which provided a Dice score of 87.21% and an IoU score of 83.04%. Replacing ResNet-50 with a more efficient encoder, EfficientNetB0, resulted in notable improvements as Dice as 89.73%, and IoU as 85.92%, thereby confirming the advantages of compound scaling and lightweight computation during feature extraction. Such performance was further improved with the addition of the ASPP module to achieve Dice as 91.05% and IoU as 88.01%, due to its functionality of capturing multi-scale contextual information with dilated convolutions. To evaluate the effectiveness of the attention mechanism, we incorporate various widely adopted modules in a systematic manner after the ASPP block.

Implementing Squeeze-and-Excitation (SE) attention resulted in a Dice as 92.33% and IoU as 89.47%, thus proving the efficiency of channel reweighting. Utilizing Convolutional Block Attention Module (CBAM) [45], which combines spatial and channel attention, enabled further improvement such as Dice of 92.56% and IoU of 89.84%. Triplet Attention [46] with Dice of 92.41% and IoU of 89.58%, and Permute Squeeze-and-Excitation (PSE) [47] with Dice of 92.28% and IoU of 89.35%, also represent other successful alternatives with marginal improvements. The proposed model, which includes multi-level channel and spatial attention applied across different semantic levels, achieved the best overall performance as Dice of 93.78% and IoU of 92.17%. This confirms that hierarchically applied attention mechanisms provide superior feature refinement in capturing global context and fine-grained spatial information, which is especially critical in segmenting complex anatomical structures like the gastrointestinal organs.

Table 3: Ablation Study of the Proposed Model

Model Configuration	**Attention Mechanism**	**Dice Coeff (%)**	**IoU Coeff (%)**	**Remarks**
Baseline DeepLab V3+	None	87.21	83.04	Standard baseline; lacks attention of lightweight design
DeepLab V3+ with EfficientNetB0 Encoder	None	89.73	85.92	Improved efficiency and feature representation
DeepLab V3+ with EfficientNetB0 Encoder and ASPP	None	91.05	88.01	ASPP enhances multi-scale contextual understanding
DeepLab V3+ with EfficientNetB0 Encoder, ASPP, and SE	SE	92.33	89.47	Channel attention emphasizes important feature maps.
DeepLab V3+ with EfficientNetB0 Encoder, ASPP, and CBAM	CBAM	92.56	89.84	Improves spatial focus but increases complexity
DeepLab V3+ with EfficientNet B0 Encoder, ASPP, and Triplet Attention	Triplet Attention	92.41	89.58	Captures inter-dimensional relations
DeepLab V3+ with EfficientNetB0 Encoder, ASPP, and PSE	PSE	92.28	89.35	Focus on spatial sensitivity at pixel level
Proposed Model (DeepLab V3+ with EfficientNet B0	**Multi-Level Channel +**	**93.78**	**92.17**	**Best overall performance; efficient and**

Encoder, channel and attention)	**ASPP, attention, multilevel**	**Spatial Attention**		**accurate segmentation**

6. State-of-the-Art Comparison

To put the performance of the suggested Multi-level Attention DeepLab V3+ model into context, there was a thorough comparative analysis performed against the variety of recent state-of-the-art methods in gastrointestinal tract segmentation, as illustrated in Table 4. Models range across diverse architectural designs, from traditional U-Net-based structures to ensemble models, transformer-based models, and hybrid encoder-decoder structures. The aim is to show the positioning of suggested model in the overall research. Each of the comparative models included in Table 4 was originally trained and tested on the same UW–Madison GI Tract MRI dataset, the sole publicly available benchmark for GI organ segmentation. Reported values for Dice and IoU were directly extracted from the corresponding studies. The cited value is the performance metric reported by authors wherever full metrics are not available. The model that was suggested was trained and tested under the same dataset conditions and preprocessing procedures to ensure that all the comparisons were fair and directly comparable.

Approaches like SIA-UNet [20] and hybrid CNN Transformer networks [21] have contributed to GI tract segmentation to a great extent, with the latter using transformer blocks to improve long-range contextual awareness. Yet, these approaches either lack high-performance segmentation capability or require huge computational resources. For example, the hybrid CNN–Transformer model has a dice score of 79% and an IoU of 72% with roughly 18.6 M parameters and 56.2 FLOPs, which shows only slight improvement in capturing global features but poor boundary accuracy. Ensemble-based models like [26] yield higher performance (Dice = 91.30%) but require several pretrained backbones, resulting in 45 M parameters, over 120 FLOPs, and inference times of over 90 ms, which limits their applicability for real-time or clinical settings. Similarly, transformer-dense architectures like BiFTransNet [29], Swin Transformer-based UPerNet [30], and EfficientNet-B7 models [31] provide comparable Dice scores of 86.8% to 89.9%, but with 25–33 M parameter count and 70–95 FLOPs computational requirement, they have slow inference times and more memory-intensive usage. The opposite, the suggested Multi-Level Attention DeepLab V3+ model achieves the best dice coefficient of 93.78% and IoU coefficient of 92.17% at having only 8.3 M trainable parameters, 21.7 FLOPs, and an average inference time of 31 ms per image. This reflects the best trade-off for segmentation performance and efficiency. The advancement is the result of a few architectural breakthroughs: (1) using EfficientNet-B0 as a lightweight yet powerful encoder to learn deep hierarchical features with compound scaling; (2) using multi-level channel and spatial attention in both the ASPP module and decoder to selectively highlight informative features at low, mid, and high semantic levels; and (3) using attention-refined skip connections that improve boundary localization and structural consistency during upsampling. Together, these design choices enable the model to achieve state-of-the-art accuracy with the lowest parameter footprint, validating its claim of being a truly lightweight and efficient segmentation framework suitable for clinical environments.

Although nnU-Net [48] has established strong and standardized performance across diverse medical image segmentation challenges, no official results or benchmark implementation currently exist for the UW-Madison GI Tract MRI dataset. Given that nnU-Net dynamically adapts its architecture to dataset-specific characteristics such as voxel spacing and modality, reproducing its results without the original 3D volumetric MRI data would not yield a fair comparison. Nevertheless, based on its consistent success in other organ segmentation tasks (e.g., brain, prostate, liver), it can be expected that nnU-Net would achieve high performance on this dataset. Future work will include a full retraining of nnU-Net on the UW-Madison dataset to enable a standardized performance comparison

Table 4: Performance Comparison of the Proposed Attention DeepLab V3+ Model with Recent State-of-the-Art Segmentation Methods on the UW-Madison GI Tract Dataset.

Ref. No.	**Year**	**Technique**	**Trainable Parameters (M)**	**Flops (G)**	**Inference Time per Image (ms)**	**Results**	**Summary**
[20]	2022	SIA-Unet	23.1	61.4	48	IoU-0.65	Spatial attention; limited context modeling
[21	202	Hybrid	18.6	56.2	52	Dice-	Transformer

]	2	CNN-transformer Architecture				0.79 IoU-0.72	improves long-range capture
[22]	2022	U-Net and Mask R-CNN	25.4	64.9	58	Dice-0.73	Combines detection and segmentation
[24]	2022	Unet on 2.5D	28.3	68.7	63	Dice-0.63 IoU-0.56	Used adjacent slices; lower accuracy
[25]	2022	UNet trained with a ResNet50 backbone and a more economical and streamlined UNet	33.2	71.5	65	IoU-0.78	Light weight, but lacks attention
[26]	2022	Ensembles of different transfer learning architectures as the backbone of UNet	45.6	120.8	94	Dice-0.9130	High accuracy but computationally expensive
[27]	2022	Different encoders ResNet, EfficientNet, VGG16, and MobileNet for U-Net	29.5	70.4	61	IoU-0.84	Comparison of backbones
[29]	2023	BiFTransNet transformer-based model	25.4	82.6	64	Dice-0.8951	Global-local fusion via BiFusion
[30]	2023	EfficientNet 4B + Swin Transformer	28.3	95.1	72	Dice-0.8682	Strong transformer backbone
[31]	2023	EfficientNet B7	33.2	88.4	70	Dice-0.8991 IoU-0.8693	High performance, high computational cost

[32]	2024	Multiscale CNN with Deep Feature Fusion	21.8	58.9	46	Dice-0.9041 IoU-0.8817	Combines multi-resolution streams
[33]	2024	Swin-ViT+ Channel SE Attention	24.9	66.3	51	Dice-0.9102 IoU-0.8906	Efficient with channel attention for medical imaging
Proposed model		**Multi-level Attention DeepLab V3+ Model**	**8.3**	**21.7**	**31**	**Dice-0.9378,** **IoU-0.9217**	**Highest accuracy with lightweight encoder and attention integration**

7. Conclusion and Future Work

This research study proposed a multi-level attention DeepLab V3+ model. The proposed model employs EfficientNet B0 as the encoder, atrous convolution, and a multi-level attention mechanism to improve model effectiveness. This study assists doctors in segmenting the stomach, small intestine, and large intestine to adjust the X-ray beam and ensure drug delivery to the tumor in the GI tract. The multi-level attention mechanism introduced in DeepLab V3+ is highly effective for cancer treatment by improving segmentation. The enhanced segmentation accuracy supports early tumor detection, tumor evaluation, and clinical monitoring, ultimately leading to more precise treatment decisions and better patient outcomes. Moreover, the model was also tested with Adam, RMSprop, and SGD optimizers. The model was also run with 2, 4, and 8 cross-folds and 10, 20, and 30 epochs. The optimized model was trained with the RMSprop optimizer, 4 cross-validation folds, and 30 epochs. The optimized model's outcomes are dice, IoU, and loss of 0.9378, 0.9217, and 0.0044, respectively. Future directions may involve optimizing the computational model and computational resources to make it more accessible to researchers with limited computational capabilities.

Although the novel Attention DeepLab V3+ model achieves excellent performance on the UW-Madison GI Tract dataset, several limitations should be considered. The model was trained and tested solely on healthy anatomical zones from a single dataset, which may make its application to pathological examples or data from other institutions or imaging modalities, such as CT or endoscopy, challenging. Further, the dataset's class imbalance and limited range of image quality can introduce bias into feature learning, potentially compromising segmentation in underrepresented areas. Real-time performance and clinical deployment have not been pursued as yet. In the future, the model will be extended to multi-institutional datasets, include pathological cases, and incorporate transformer-based modules for enhanced global context learning. Additional enhancements may include domain adaptation methods, real-time inference optimization, and clinical workflow evaluation to assess practicability. The UW–Madison GI Tract dataset is the only existing publicly available MRI dataset with pixel-level annotations for segmentation of gastrointestinal organs. Thus, the model suggested was trained and validated solely on this dataset. Even though cross-dataset or multi-center validation was not possible because comparable datasets were unavailable, comprehensive data augmentation, cross-fold validation, and testing for robustness against noise and motion artifacts were used to enable the model to generalize. Collaboration with clinical centers to build and test the model on multi-institutional gastrointestinal datasets will be pursued in the future, once such datasets become accessible.

List of Abbreviations	
GI	Gastrointestinal
CNN	Convolutional Neural Network
ASPP	Atrous Spatial Pyramid Pooling
RLE	Run-Length Encoding
MBConv	Mobile Inverted Bottleneck Convolution
ReLU	Rectified Linear Unit
IoU	Intersection over Union
Adam	Adaptive Moment Estimation
SGD	Stochastic Gradient Descent
CBAM	Convolutional Block Attention Module
PSE	Permute Squeeze-and-Excitation
SE	Squeeze-and-Excitation
MRI	Magnetic Resonance Imaging

Author Contributions

Neha Sharma: Conceived the study, designed the model architecture, and led the experimental implementation and manuscript writing.
Sheifali Gupta: Contributed to data preprocessing, model optimization, and assisted in drafting and reviewing the manuscript.
Fuad Ali Mohammed Al-Yarimi was responsible for literature review, dataset preprocessing, and assisted in the formulation of the evaluation metrics and performance analysis
Upinder Kaur: Participated in literature review, performance evaluation, and analysis of segmentation results.
Salil Bharany: Provided technical guidance, contributed to model evaluation metrics, and reviewed the manuscript for technical accuracy.
Ateeq Ur Rehman: Supported model tuning, cross-validation experiments, and interpretation of results from a clinical perspective.
Belayneh Matebie Taye: Supervised the research, provided critical revisions to the manuscript, and managed overall coordination, communication, and final submission.

Acknowledgement: The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through the Large Research Project under grant number RGP2/27/46.

Data availability statement: The dataset used in this study, the "UW-Madison GI Tract Image Segmentation" dataset, is publicly available on Kaggle. It can be accessed at https://www.kaggle.com/competitions/uw-madison-gi-tract-image-segmentation/data.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Ethics Approval Statement: No animals or human subjects were involved in this study. The study utilized publicly available datasets, and all methods were carried out in accordance with relevant guidelines and regulations.

Consent to Publish Declaration: not applicable.

Funding Declaration section: No funding.

References

[1] Zhou, S. K. et al. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proc. IEEE Inst. Electr. Electron. Eng. 109, 820–838 (2021).

[2] Ozdemir, B., Aslan, E. & Pacal, I. Attention Enhanced InceptionNeXt Based Hybrid Deep Learning Model for Lung Cancer Detection. (IEEE Access, 2025).

[3] Pacal, I., Ozdemir, B., Zeynalov, J., Gasimov, H. & Pacal, N. A novel CNN-ViT-based deep learning model for early skin cancer diagnosis. Biomed. Signal Process. Control 104, 107627 (2025).

[4] Ozdemir, B. & Pacal, I. A robust deep learning framework for multiclass skin cancer classification. Sci. Rep. 15, 4938 (2025).

[5] Ozdemir, B. & Pacal, I. An innovative deep learning framework for skin cancer detection employing ConvNeXtV2 and focal self-attention mechanisms. Results Eng. 25, 103692 (2025).

[6] Yi, S. et al. IDC-Net: Breast cancer classification network based on BI-RADS 4. Pattern Recognit. 150, 110323 (2024).

[7] Bayram, B., Kunduracioglu, I., Ince, S. & Pacal, I. A systematic review of deep learning in MRI-based cerebral vascular occlusion-based brain diseases. Neuroscience 568, 76–94 (2025).

[8] İnce, S., Kunduracioglu, I., Bayram, B. & Pacal, I. U-Net-based models for precise brain stroke segmentation. Chaos Theory and Applications 7, 50–60 (2025).

[9] Heavey, S. F., Roeland, E. J., Tipps, A. M. P., Datnow, B. & Sicklick, J. K. Rapidly progressive subcutaneous metastases from gallbladder cancer: insight into a rare presentation in gastrointestinal malignancies. J. Gastrointest. Oncol. 5, E58-64 (2014).

[10] Rawla, P. & Barsouk, A. Epidemiology of gastric cancer: global trends, risk factors, and prevention. Gastroenterology Review/Przegląd Gastroenterologiczny 14, 26–38 (2019).

[11] Jaffray, D. A., & Gospodarowicz, M. K. Radiation therapy for cancer. Cancer: disease control priorities, 3,2015, 239-248.

[12] Yi, S., Qin, S., She, F. & Shao, D. BSD: A multi-task framework for pulmonary disease classification using deep learning. Expert Syst. Appl. 259, 125355 (2025).

[13] Du, S., Du, S., Liu, B. & Zhang, X. Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high-resolution remote sensing images. International Journal of Digital Earth 14, 357–378 (2021).

[14] Choudhury, R., Vanguri, A., Jambawalikar, R. & Kumar, S. R. Segmentation of brain tumors using DeepLabv3+. in Brainlesion: Glioma, Multiple Sclerosis, Stroke, and Traumatic Brain Injuries: 4th International Workshop, BrainLes 154–167 (Springer International Publishing, Granada, Spain, 2018).

[15] Azad, R., Asadi-Aghbolaghi, M., Fathy, M. & Escalera, S. Attention Deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation. in Lecture Notes in Computer Science 251–266 (Springer International Publishing, Cham, 2020).

[16] da Cruz, L. B. et al. Kidney tumor segmentation from computed tomography images using DeepLabv3+ 2.5D model. Expert Syst. Appl. 192, 116270 (2022).

[17] Bernal, J., Sánchez, J., & Vilarino, F. (2012). Towards automatic polyp detection with a polyp appearance model. Pattern Recognition, 45(9), 3166-3182.

[18] Poorneshwaran, J. M., Santhosh Kumar, S., Ram, K., Joseph, J. & Sivaprakasam, M. Polyp Segmentation using Generative Adversarial Network. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 2019, 7201–7204 (2019).

[19] Lafraxo, S. & El Ansari, M. GastroNet: Abnormalities recognition in gastrointestinal tract through endoscopic imagery using deep learning techniques. in 2020 8th International Conference on Wireless Networks and Mobile Communications (WINCOM) (IEEE, 2020).

[20] Ye, R., Wang, R., Guo, Y. & Chen, L. SIA-Unet: A Unet with Sequence Information for Gastrointestinal Tract Segmentation. in Pacific Rim International Conference on Artificial Intelligence 316–326 (Springer, Cham, 2022).

[21] Nemani, P. & Vollala, S. Medical image segmentation using LeViT-UNet++: A case study on GI tract data. arXiv [cs.NE] (2022).

[22] Chou, A., Li, W. & Roman, E. GI Tract Image Segmentation with U-Net and Mask R-CNN. Image Segmentation with U-Net and Mask R-CNN.

[23] Niu, H. & Lin, Y. SER-UNet: A Network for Gastrointestinal Image Segmentation. in Proceedings of the 2022 2nd International Conference on Control and Intelligent Robotics (ACM, New York, NY, USA, 2022).

[24] Li, H. & Liu, J. Multi-view unet for automated GI tract segmentation. in 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI) (IEEE, 2022).

[25] Chia, B., Gu, H. & Lui, N. Gastrointestinal Tract Segmentation Using Multi-Task Learning.

[26] Georgescu, M.-I., Ionescu, R. T. & Miron, A.-I. Diversity-promoting ensemble for medical image segmentation. arXiv [eess.IV] (2022).

[27] Jiang, X. et al. BiFTransNet: A unified and simultaneous segmentation network for gastrointestinal images of CT & MRI. Comput. Biol. Med. 165, 107326 (2023).

[28] Qiu, Y. Upernet-Based Deep Learning Method For The Segmentation Of Gastrointestinal Tract Images. in Proceedings of the 2023 8th International Conference on Multimedia and Image Processing 34–39 (2023).

[29] John, S. V. & Benifa, B. Automated segmentation of tracking healthy organs from gastrointestinal tumor images. in Smart Innovation, Systems and Technologies 363–373 (Springer Nature Singapore, Singapore, 2023).

[30] Wang, B. et al. Low-friction soft robots for targeted bacterial infection treatment in gastrointestinal tract. Cyborg Bionic Syst. 5, 0138 (2024).

[31] Li, H. et al. UCFNNet: Ulcerative colitis evaluation based on fine-grained lesion learner and noise suppression gating. Comput. Methods Programs Biomed. 247, 108080 (2024).

[32] Song, W. et al. CenterFormer: A novel cluster center enhanced transformer for unconstrained dental plaque segmentation. IEEE Trans. Multimedia 26, 10965–10978 (2024).

[33] Jiang, X. et al. BiFTransNet: A unified and simultaneous segmentation network for gastrointestinal images of CT & MRI. Comput. Biol. Med. 165, 107326 (2023).

[34] Nobel, S. M. N., Sifat, O. F., Islam, M. R., Sayeed, M. S. & Amiruzzaman, M. Enhancing GI cancer radiation therapy: Advanced organ segmentation with ResECA-U-Net model. Emerg. Sci. J. 8, 999–1015 (2024).

[35] Sharma, N., Gupta, S., Gupta, D., Gupta, P., Juneja, S., Shah, A., & Shaikh, A. (2024). UMobileNetV2 model for semantic segmentation of gastrointestinal tract in MRI scans. Plos one, 19(5), e0302880.

[36] https://www.kaggle.com/competitions/uw-madison-gi-tract-image-segmentation/data

[37] Wu, Z. et al. Segmentation of abnormal leaves of hydroponic lettuce based on DeepLabV3+ for robotic sorting. Comput. Electron. Agric. 190, 106443 (2021).

[38] Koonce, B. EfficientNet. in Convolutional Neural Networks with Swift for Tensorflow 109–123 (Apress, Berkeley, CA, 2021).

[39] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7132-7141, doi: 10.1109/CVPR.2018.00745.

[40] Zhang, Z. Improved Adam optimizer for deep neural networks. in 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS) (IEEE, 2018).

[41] Wichrowska, O. et al. Learned Optimizers that Scale and Generalize. arXiv [cs.LG] (2017).

[42] Keskar, N. S. & Socher, R. Improving generalization performance by switching from Adam to SGD. arXiv [cs.LG] (2017).

[43] Wang B, Chen Y, Ye Z, Yu H, Chan KF, Xu T, Guo Z, Liu W, Zhang L. Low-Friction Soft Robots for Targeted Bacterial Infection Treatment in Gastrointestinal Tract. Cyborg Bionic Syst. 2024;5: Article 0138. https://doi.org/10.34133/cbsystems.0138

[44] Li, H., Wang, Z., Guan, Z., Miao, J., Li, W., Yu, P., Molina Jimenez, C. (2024). UCFNNet: Ulcerative colitis evaluation based on fine-grained lesion learner and noise suppression gating. Computer Methods and Programs in Biomedicine, 247, 108080. doi: https://doi.org/10.1016/j.cmpb.2024.108080

[45] Woo, S., Park, J., Lee, JY., Kweon, I.S. (2018). CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11211. Springer, Cham. https://doi.org/10.1007/978-3-030-01234-2_1

[46] D. Misra, T. Nalamada, A. U. Arasanipalai and Q. Hou, "Rotate to Attend: Convolutional Triplet Attention Module," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2021, pp. 3138-3147, doi: 10.1109/WACV48630.2021.00318.

[47] Yiran Wang, Yuxin Bian, Shenlu Jiang, PSE: Enhancing structural contextual awareness of networks in medical imaging with Permute Squeeze-and-Excitation module, Biomedical Signal Processing and Control, Volume 100, Part B, 2025, 107052, doi: https://doi.org/10.1016/j.bspc.2024.107052.

[48] Luu, H.M. and Park, S.H., 2021, September. Extending nn-UNet for brain tumor segmentation. In International MICCAI brainlesion workshop (pp. 173-186). Cham: Springer International Publishing.