Konya Mühendislik Bilimleri Dergisi, c. 12, s.3, 671-686, 2024
Konya Journal of Engineering Sciences, v. 12, n. 3, 671-686, 2024
ISSN: 2667-8055 (Elektronic)
DOI: 10.36306/konjes.1404420

A 3D U-NET BASED ON EARLY FUSION MODEL: IMPROVEMENT, COMPARATIVE ANALYSIS WITH STATE-OF-THE-ART MODELS AND FINE-TUNING

[1,*] **Beyza KAYHAN**, [2]**Sait Ali UYMAZ**

[1,2] Konya Technical University, Engineering and Natural Sciences Faculty, Computer Engineering Department, Konya, TÜRKİYE
[1]bkayhan@ktun.edu.tr, [2]sauymaz@ktun.edu.tr

Highlights

- A review on deep learning based multi-organ segmentation.
- Using the two-stage U-Net model
- Improving fusion approach combining different color channels for segmentation of CT images
- Performing parameter optimization
- Comparison of performances of different models in segmentation of abdominal organs

***Corresponding Author:** Beyza KAYHAN, bkayhan@ktun.edu.tr

Konya Mühendislik Bilimleri Dergisi, c. 12, s.3, 671-686, 2024
Konya Journal of Engineering Sciences, v. 12, n. 3, 671-686, 2024
ISSN: 2667-8055 (Elektronic)
DOI: 10.36306/konjes.1404420

A 3D U-NET BASED ON EARLY FUSION MODEL: IMPROVEMENT, COMPARATIVE ANALYSIS WITH STATE-OF-THE-ART MODELS AND FINE-TUNING

[1,*] Beyza KAYHAN, [2]Sait Ali UYMAZ

[1,2] *Konya Technical University, Engineering and Natural Sciences Faculty, Computer Engineering Department, Konya, TÜRKİYE*
[1]bkayhan@ktun.edu.tr, [2]sauymaz@ktun.edu.tr

(Received: 15.12.2023; Accepted in Revised Form: 20.06.2024)

ABSTRACT: Multi-organ segmentation is the process of identifying and separating multiple organs in medical images. This segmentation allows for the detection of structural abnormalities by examining the morphological structure of organs. Carrying out the process quickly and precisely has become an important issue in today's conditions. In recent years, researchers have used various technologies for the automatic segmentation of multiple organs. In this study, improvements were made to increase the multi-organ segmentation performance of the 3D U-Net based fusion model combining HSV and grayscale color spaces and compared with state-of-the-art models. Training and testing were performed on the MICCAI 2015 dataset published at Vanderbilt University, which contains 3D abdominal CT images in NIfTI format. The model's performance was evaluated using the Dice similarity coefficient. In the tests, the liver organ showed the highest Dice score. Considering the average Dice score of all organs, and comparing it with other models, it has been observed that the fusion approach model yields promising results.

Keywords: *Computed Tomograph, Multi Organ Segmentation, Deep Learning, Fusion Model, U-Net*

1. INTRODUCTION

Segmentation of organs in medical images is of crucial importance for diagnosing diseases, planning treatment, and locating target organs for radiotherapy [1]. Automated multi-organ segmentation is difficult because of structural complexity and volumetric differences of organs. In recent years, there has been a growing interest in using deep learning methods to address these difficulties [2]. These methods automatically extract feature vectors, which are used for tasks such as object detection and classification. This feature vector extraction is achieved through non-linear layers. By using multiple layers, deep learning can learn different features from the data. For example, basic features like edges and patterns are learned in the first layers, while more complex features are learned in subsequent layers [3], [4]. Deep learning has been successfully applied in various fields, including face recognition [5], voice recognition [6], robotic applications [7], and particularly in the biomedical applications [8]. This is due to the increasing availability of medical images and the ability of deep learning architectures to provide fast and reliable results [9].

In this study, a 3D U-Net based fusion model combining different color spaces was used to overcome the limitations of traditional methods in multi-organ segmentation and compared with state-of-the-art approaches. Roth et al. [10] increased the segmentation success by combining image inputs of different resolutions. This success shows that fusion models are an effective strategy, and based on this, the fusion model used in this study combines different color spaces. In combining different color spaces, Ghosh et al. (2018) was effective. Ghosh et al. [11] also found that combining different color spaces was effective in detecting bleeding areas in endoscopy images, with the HSV color space performing the best. This highlights the impact of color spaces on model performance. Additionally, using different color spaces can improve segmentation accuracy and reliability by highlighting different features in images [12]. One of the important aspects of this study is the inclusion of optimizations and fine-tuning to enhance the performance of the fusion model. Another crucial part is the integration of different slice selection methods to better capture contextual information from the 3D data. This approach aims to augment the data and

*Corresponding Author:** Beyza KAYHAN, bkayhan@ktun.edu.tr ultimately improve the accuracy and reliability of the segmentation.

2. RELATED WORK

Automatic segmentation of organs in computed tomography images is difficult due to differences in shape and size. Improving segmentation accuracy by overcoming these challenges has become an active area of research. When deep learning methods were not widespread, traditional and atlas-based methods were used in multi-organ segmentation. In these methods, mathematical and techniques methods are used to perform the segmentation process. Their differences in organs complicated the segmentation process. In recent years, deep learning-based methods that address organ differences more effectively have been used and have been observed to yield successful results [13].

Trullo et al. [14] proposed two common deep architectures to jointly separate all organs, including aorta, heart, esophagus, and trachea, instead of separating them separately. The second deep architecture, using the Sharp Mask network, is trained to distinguish each target organ from the background. In this study, initial segmentation was found to be useful for the segmentation of target organs. Larsson et al. [15] proposed a two-stage convolutional neural network for organ segmentation. In this network, each organ is segmented independently. The central voxel of the organ is obtained using the feature-based multiple atlas approach, and a prediction mask is placed around it. Subsequently, a 3D convolutional neural network (CNN) is applied for voxel-wise classification. This initialization method enables the training of regional networks, where the voxel only needs to distinguish between a specific organ and the background.

Roth et al. [16] propose a stepwise approach using a 3D fully convolutional network (FCN) trained on CT images. In the first stage, a mask of the patient's internal structure is obtained by applying simple thresholding with morphological operations. The FCN architecture is then trained using this mask, resulting in a reduction in the number of voxels required to calculate the loss function of the network. Additionally, the number of regions in the 3D image input to the convolutional neural network (CNN) is reduced by approximately 40%. In the second stage, the FCN architecture is trained with the mask obtained from the first stage. This architecture was tested on 150 CT images containing three organs (liver, spleen, and pancreas). Roth, Sugino, et al. [10] propose a multi-scale 3D FCN approach for high-resolution segmentation. The 3D FCN predictions of low-resolution inputs are combined with high-resolution 3D FCN inputs.

Shen et al. [17] show that the performance of multi-organ segmentation depends on the loss function as well as the network architecture. They compared the effects of Dice-based loss functions on CT images for multi-organ segmentation. In addition, they examined the impact of three different weighting types (uniform, simple, and square) and initial learning rates on segmentation using a 3D FCN. The models were evaluated on a random subset of 340 training and 37 test patients. The network produced a predictive map with eight classes, including seven organs (liver, stomach, spleen, gallbladder, artery, portal vein, and pancreas) and background.

Kakeya et al. [18] proposed a new deep learning model using transfer learning for automatic multi-organ segmentation. This model, called 3D U-JAPA-Net, in addition to the raw CT data, also uses a probability atlas of organs (PA), which provides information about the positions of the organs. The 3D U-JAPA-Net model utilizes transfer learning to effectively incorporate PA information. During the model training process, a 3D U-JAPA-Net with nine output classes (including eight organs and a background class) is trained using data from organs in their bounding boxes.

Vesal et al. [19] utilized a deep learning architecture to segment organs at risk (OARs) in thoracic CT images. The architecture combines a 2D U-Net and Dense Residual (DR) network, consisting of four downsampling and upsampling convolution blocks in the encoder and decoder branches. Due to limited sample size, a deeper 2D version of the network was used. In each block, two 3x3 convolutions and ReLU activation function were applied.

Mietzner and Mastmeyer [20] have developed an automated method for detecting and segmenting abdominal organs in CT scans. It is challenging to detect the pancreatic organ in particular. Using a combination of the random forest regression method and the 2D U-Net architecture, the segmentation mask and bounding box of five organs, namely liver, kidneys, spleen, and pancreas were detected. A dataset of 50 CT scans was used in this study. Rister et al. [21] trained a deep neural network to perform multi-organ segmentation. 140 CT scans were used, including six organs: liver, lung, bladder, kidney, bone, and brain. First, the lungs and bones were segmented a 3D Fourier transform, followed by the use of a 3D U-Net architecture to segment the remaining organs. Liu et al. [22] aimed to develop a deep learning-based method for multi-organ segmentation. Eight organs, namely the large intestine, small intestine, duodenum, left kidney, right kidney, liver, spinal cord, and stomach, were labeled by experts in CT images. The segmentation process was performed using a 3D U-Net architecture. Fang and Yan [23] performed multi-organ segmentation using a multi-scale neural network. The system with pyramid input is integrated into the U-Net network to combine features at different scales. Finally, the pyramid outputs are combined to achieve improved segmentation. This proposed network is called PIPO-FAN. Zhang et al. [24] proposed a full volume-based method, the efficientSegNet network, for multi-organ segmentation. This method takes full advantage of the 3D context and aims to reduce computational costs.

Kaur et al. [25] present a systematic literature review for multi-organ segmentation in the study. Previous studies have shown that the most used architectures for abdominal multi-organ segmentation are CNN, FCN, and U-Net. Generally, segmentation of large organs such as liver, kidney and spleen has been performed. It is not preferred due to the difficulty of segmenting small organs such as duodenum, esophagus, pancreas, and gallbladder. More research is needed to segment small organs and improve segmentation accuracy in the future.

3. MATERIAL AND METHODS

In this study, the Python (3.6) programming language was utilized for automatic multi-organ segmentation. The Simple ITK library was used to read and process 3D tomography images, The Numpy library was used for numerical operations. The Pytorch library was also utilized for developing deep learning models. The fusion model used in this study was executed on NVidia GeForce RTX 2070 with 8 GB of memory.

In this section, the dataset used for training the model is discussed. The data preprocessing and data augmentation processes performed on this data set are explained. Additionally, details of the fusion model used for multi-organ segmentation are given.

3.1. Dataset

In this study, the dataset containing abdominal CT images provided by Vanderbilt University Medical Center (VUMC) was used. The dataset consists of 30 images. Volume dimensions of CT images are 512 x 512 x 85 and 512 x 512 x 198, resolution 0.54 x 0.54 mm^2 and 0.98 x 0.98 mm^2 and slice thickness 2.5 mm and 5.0 mm varies between. Trained individuals labeled a total of 13 organs in each CT image, which were then validated by a radiologist. Some patients do not have a right kidney or gallbladder. For this reason, it was not labeled. The data was recorded in the NIfTI file format [26]. Figure 1 shows each labeled organ.

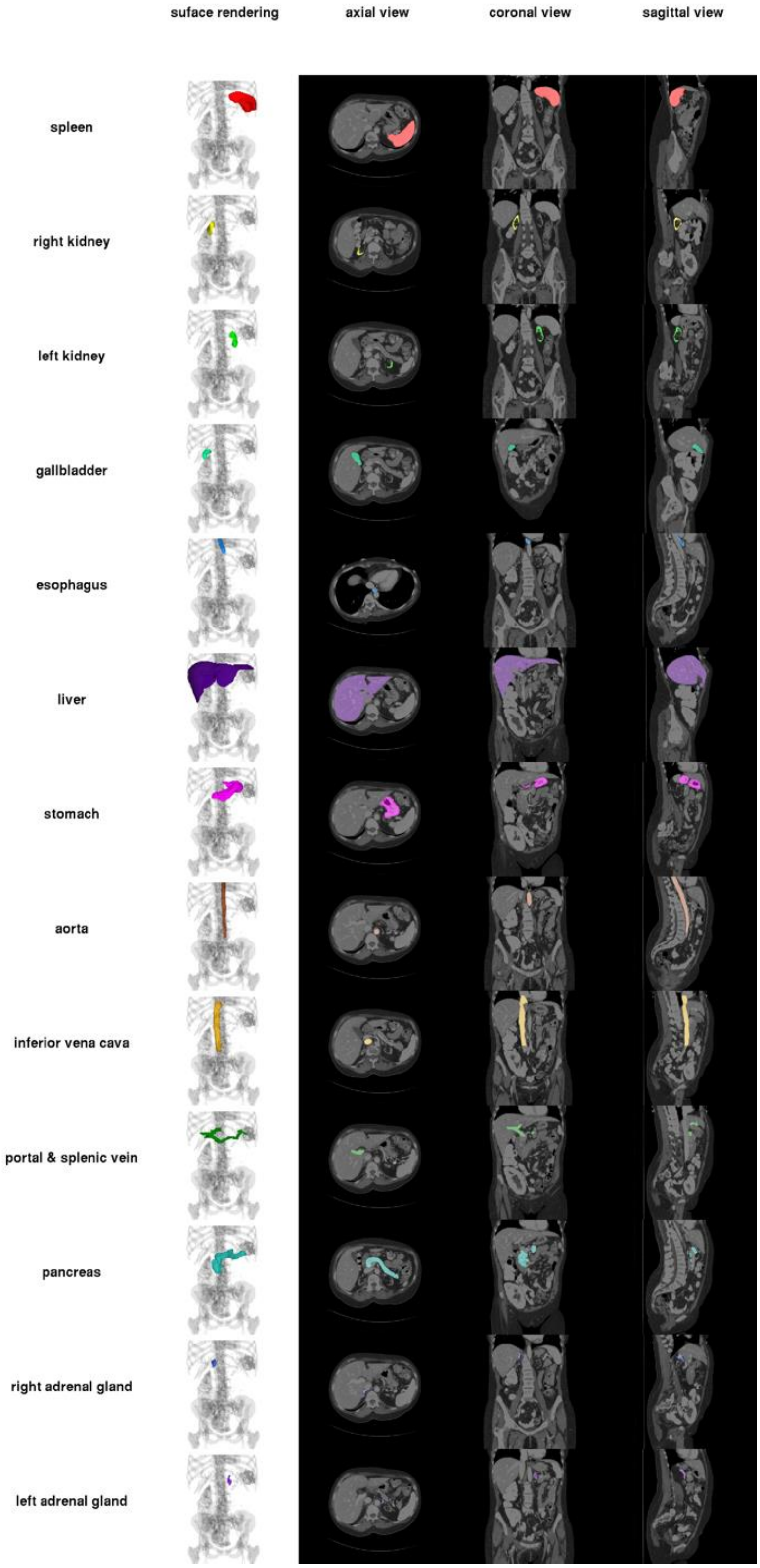


Figure 1. CT images of organs [26]

3.2. Data Preprocessing and Augmentation

Each CT image in the dataset varies in size from 512x512x85 and 512x512x195. Training with three-dimensional data requires expensive hardware resources. Therefore, resizing images can help overcome this challenge. Due to GPU limitations, the image sizes in the x and y planes were reduced by 1/4 and the number of slices of each image was set to 64. However, to compensate for the information lost during this resizing process and to diversify our training set, different slices were selected from the same image. Five different methods were used for this slice selection process:

1. The first 64 slices of each image were selected.
2. The last 64 slices of each image were selected.
3. Each image was divided into two subsets based on the first slice. Subset 1 (0,2,4…,126) and subset 2 (1,3,5…,127) contain [27].
4. Each image was divided into two subsets based on the last slice. Subset 1 (69…191,193,195) and subset 2 (68…190,192,194) contain.
5. A random start slice was determined in the depth of each image, and 64 consecutive slices was selected from the start slice.

As a result, the input image size for the network was set to 128x128x64. Additionally, a random rotation between -5 and 5 degrees was applied to the images obtained with the 5th method to increase the diversity of the data. As a result of these processes, the number of images to be trained was increased from 24 to 192. In Table 1, the dimensions of the raw data, the preprocessing steps to equalize the slice sizes, the data augmentation process, and the result data are given. 'x' represents the random starting point selected from the slices, and 'z' represents the number of slices.

Table 1. Data obtained as a result of preprocessing and data augmentation of raw data

	Steps	Width	Height	Slice Number Range	Number of Data	Selected Slices	Rotation
Raw Data		512	512	85-195	24	-	-
Preprocessing Steps	1	128	128	64	24	(0….64)	-
	2	128	128	64	24	(z-63 …z-2, z-1, z)	-
	3	128	128	64	24	z>125 (0,2,4…,126) z<125 (x,x+1,x+2…x+63)	-
		128	128	64	24	z>126 (1,3,5…,127) z<126 (x, x+1, x+2…x+63)	-
	4	128	128	64	24	(69…191,193,195)	-
		128	128	64	24	(68…190,192,194)	-
	5	128	128	64	24	(x, x+1, x+2…x+63)	-
Data Augmentation	6	128	128	64	24	(x, x+1, x+2…x+63)	-5,5
Result Data		128	128	64	192		

The model used in this study has two stages. In the first stage, the gray images in the dataset were converted to images with HSV (Hue, Saturation, Value) [28] color space with the colormap function in the Simple ITK library. With this function, single channel images are normalized between 0 and 1 and a color map is used to assign colors to pixels in the image, pixels with a value of 1 are assigned the first color in the color map. The result is three-channel images with RGB (Red-Green-Blue) [28] color space. Images converted to RGB color space are converted to desired color space (HSV). In Eq. (1), 'Δ' represents the difference between the maximum (Cmax) and minimum (Cmin) values of the R, G, B components. To convert RGB images to the HSV color space, the maximum and minimum values of the R, G and B components are found, and the difference between them is calculated. In Eq. (2), 'H' represents the Hue, which is calculated based on difference between color components (Δ) and the maximum component (Cmax). In Eq. (3), 'S' represents the Saturation, which is calculated based on the maximum component (Cmax). In Eq. (4), 'V' represents the Value or Brightness, which directly corresponds to the value of the maximum component.

$$\begin{aligned} Cmax &= max\,(\,R\,,G\,,B\,) \\ Cmin &= min\,(\,R\,,G\,,B\,) \\ \Delta &= Cmax - Cmin \end{aligned} \tag{1}$$

$$H = 0\,,\ \Delta = 0$$

$$\mathrm{H} = \begin{cases} 60x\left(\frac{G-B}{\Delta}Mod6\right) & ,\quad Cmax = R\,' \\ 60x\left(\frac{B-\mathrm{R}}{\Delta}+2\right) & ,\quad Cmax = G\,' \\ 60x\left(\frac{\mathrm{R}-\mathrm{G}}{\Delta}+4\right) & ,\quad Cmax = B' \end{cases} \tag{2}$$

$$S = \begin{cases} 0 & ,Cmax = 0 \\ \frac{\Delta}{Cmax} & ,Cmax \neq 0 \end{cases} \tag{3}$$

$$V = Cmax \tag{4}$$

3.3. A 3D U-Net based on Early Fusion Model

The model used in this study is based on the two-stage 3D U-Net with early fusion approach using different color spaces proposed by Kayhan [12]. The 3D U-Net model with early fusion approach uses 3D U-Net with the same layers in both stages. The model uses two different color maps (Grayscale and HSV) of CT images. In both stages, the proposed 3D U-Net network is trained with HSV images, and an output of 15 channels is obtained by combining the output of the first stage (13 organs and one background) with the grayscale image. The image obtained is determined as the input to the 2nd stage. This merging process is called early fusion. In this way, it is aimed to make the features of the organs more evident. These combined images are again trained with the proposed 3D U-Net architecture, and predictive segmentation results are obtained. The general structure of this model is given in Figure 2 [12].

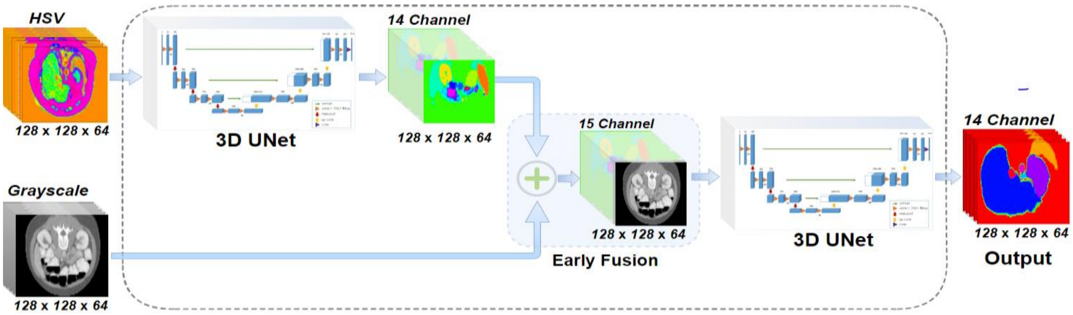


Figure 2. Fusion model [12]

Kayhan [12], proposed a 3D U-Net model consisting of encoder and decoder stages. . In the first step of this model, images are fed into the encoder network. At each level, two 3x3x3 convolution operations are performed on the input images. Batch Normalization and ReLU activation function are used after each convolution operation. Maximum pooling and two-step 2x2x2 convolution are applied to the feature map obtained while transitioning from one level to another. The outputs obtained as a result of these processes are merged. In the decoding network, upsampling is performed with a two-step 2x2x2 transpose convolution until the input image size is obtained. Batch Normalization and ReLU activation function are implemented after each transpose convolution. The feature map at each level in the decoding network and the feature map obtained from the corresponding encoder section are combined. Then, two 3x3x3 convolutions are applied to this combined feature map. In the last layer, because of the 1x1x1 convolution operation, a 128x128x48 size feature map with 14 channels is obtained. Three-dimensional multi-organ segmentation was performed by applying the softmax activation function to the output feature map.

In this study, the 3D U-Net model proposed by Kayhan [12] was fine-tuned to improve multi-organ segmentation performance. The two-step 2x2x2 convolution layer used for downsampling in the encoder network was removed, and a dropout layer was added after each inter-level transition in both the encoder and decoder networks. These fine-tuning operations were implemented to prevent overfitting of the model. Additionally, the number of slices was increased from 48 to 64 so that this model could learn more features from images and better capture context information. Figure 3 shows the 3D U-Net based model used in this study, and Table 2 shows the layers of this 3D U-Net model and the filter, input and output dimensions used in these layers.

Table 2. 3D U-Net based model layers, input and output values

Layers	Input Size		Output Size	Encoder	Layers	Input Size	Output Size	Decoder
	HSV Image	Gray Image						
Convolution	128x128x64x3	128x128x64x1	128x128x64x32	3x3x3 conv padding 1	Deconvolution	16x16x8x256	32x32x16x128	2x2x2 Transposed conv
	128x128x64x32		128x128x64x32		Concatenate Dropout	32x32x16x128 32x32x16x128	32x32x16x256	50%
Pooling	128x128x64x32		64x64x32x32	2x2x2 max pooling	Convolution	32x32x16x256	32x32x16x128	3x3x3 conv padding 1
Dropout	64x64x32x32		64x64x32x32	50%		32x32x16x128	32x32x16x128	
Convolution	64x64x32x32		64x64x32x64	3x3x3 conv padding 1	Deconvolution	32x32x16x128	64x64x32x64	2x2x2 Transposed conv
	64x64x32x64		64x64x32x64		Concatenate Dropout	64x64x32x64 64x64x32x64	64x64x32x128	50%
Pooling	64x64x32x64		32x32x16x64	2x2x2 max pooling	Convolution	64x64x32x128	64x64x32x64	3x3x3 conv padding 1
Dropout	32x32x16x64		32x32x16x64	50%		64x64x32x64	64x64x32x64	
Convolution	32x32x16x64		32x32x16x128	3x3x3 conv padding 1	Deconvolution	64x64x32x64	128x128x64x32	2x2x2 Transposed conv
	32x32x16x128		32x32x16x128		Concatenate Dropout	128x128x64x32 128x128x64x32	128x128x64x64	50%
Pooling	32x32x16x128		16x16x8x128	2x2x2 max pooling	Convolution	128x128x64x64	128x128x64x32	3x3x3 conv padding 1
Dropout	16x16x8x128		16x16x8x128	50%		128x128x64x32	128x128x64x32	
Convolution	16x16x8x128		16x16x8x256	3x3x3 conv padding 1	Convolution	128x128x64x32	128x128x64x14	1x1x1 conv
	16x16x8x256		16x16x8x256					

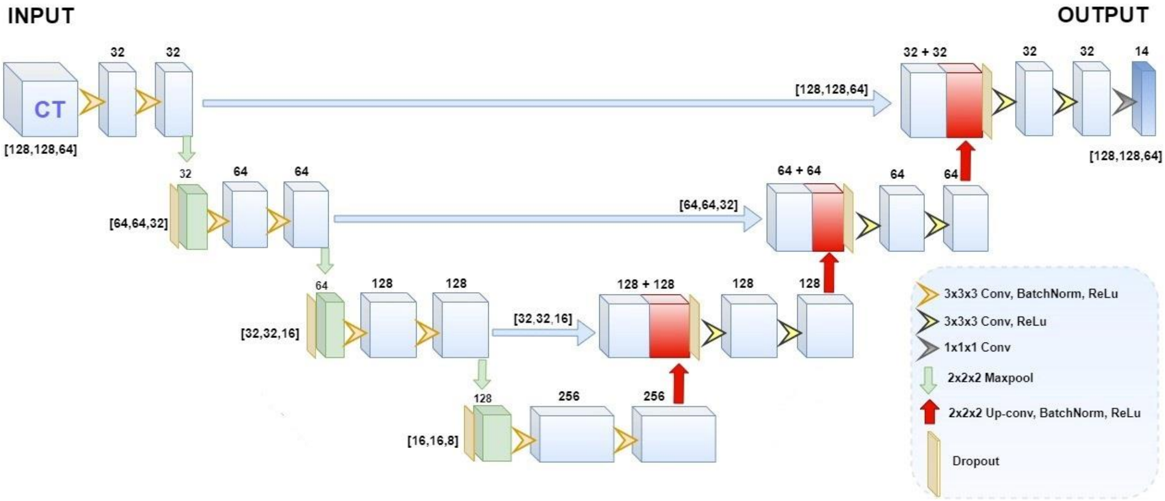


Figure 3. 3D U-Net based model

3.4. Hyperparameter Optimization

Hyperparameter optimization was carried out to improve the performance of the fusion model used in this study. To optimize the fusion model, training was conducted using various parameter sets. The data was split into 80% for training and 20% for testing. The dice score was used as the evaluation metric. In Table 3, the dice score results of different parameter sets on the test data set are given. When Table 3 is examined, batch size two was used in all samples, and the Adam optimization algorithm was used. Different learning rate values, dropout rates, activation functions and epoch numbers used with these parameters were compared. The dropout layer had a positive effect on the dice score result. The learning rate that gives the highest dice score is 1e-3, the activation function is ReLU, the dropout rate is 0.5, and the epoch number is 200. As a result of this optimization, the parameter set giving the best result was determined.

3.5. Performance Evaluation Metric

The segmentation process performed in this study was evaluated using the Dice similarity coefficient. This metric evaluates the level of similarity in two images by measuring the number of matching pixels. Dice similarity coefficient formula is given in Eq. (5). The meaning of the symbols used in this equation is explained below [29].

- **Y :** Actual labels
- Ý : Predicted labels
- $\bar{y}_{ij.}$: Elements in Ý
- y_{ij} : Elements in Y
- n : Row elements
- m: Column elements

$$DC = \frac{2|Ý \cap Y|}{|Ý| + |Y|} \tag{5}$$

$$\frac{2\sum_{i=1}^{n}\sum_{j=1}^{m} \bar{y}_{ij}.y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} \bar{y}_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}$$

Table 3. Test dice results of different parameter sets

Batch size	Optimization algorithm	Learning rate	Activation function	Epoch	Dropout	Spleen	Right kidney	Left kidney	Esophagus	Gallbladder	Liver	Stomach	Aorta	Inferior vena cava	Portal and splenic vein	Pancreas	Right adrenal gland	Left adrenal gland	Average
2	Adam	1e-3	ReLU	100	-	0.934	0.944	0.937	0.619	0.665	0.958	0.869	0.851	0.814	0.709	0.754	0	0.583	0.741
					0.1	0.93	0.924	0.924	0.675	0.941	0.941	0.857	0.864	0.788	0.72	0.734	0.579	0.559	0.779
					0.2	0.914	0.944	0.91	0.754	0.663	0.958	0.861	0.846	0.78	0.732	0.745	0.609	0.588	0.793
					0.3	0.890	0.937	0.905	0.738	0.744	0.955	0.858	0.864	0.808	0.742	0.689	0.63	0.626	0.799
					0.4	0.943	0.942	0.925	0.511	0.70	0.95	0.866	0.844	0.788	0.711	0.683	0.584	0.635	0.775
					0.5	0.922	0.942	0.924	0.711	0.656	0.953	0.869	0.872	0.807	0.751	0.755	0.641	0.639	0.803
				200		**0.954**	**0.948**	**0.949**	**0.743**	**0.719**	**0.96**	**0.885**	**0.881**	**0.80**	**0.764**	**0.779**	**0.641**	**0.632**	**0.82**
				300		0.947	0.95	0.949	0.723	0.713	0.964	0.882	0.882	0.82	0.744	0.765	0.641	0.629	0.816
			PReLU	200		0.96	0.951	0.948	0.718	0.707	0.964	0.876	0.872	0.809	0.742	0.787	0.591	0.545	0.805
		5e-3	ReLU			0.938	0.94	0.91	0.724	0.698	0.96	0.87	0.876	0.813	0.74	0.745	0.619	0.621	0.804
		5e-4				0.943	0.95	0.946	0.636	0	0.961	0.872	0.879	0.819	0.744	0.741	0.627	0.634	0.75

4. RESULTS

The fusion model used in this study was evaluated on the MICCIA 2015 dataset. Train and test performance results of the best parameter set determined because of hyperparameter optimization are given in Table 4. In multi-organ segmentation, the training and test dice results are 0.932 and 0.82, respectively. The curves of the results obtained during the training and test set are given in Figure 4, and the final test set results for each organ are given in Figure 5.

Table 4. Training and test dice results of each organ

	Dice Coefficient													
SET	**Spleen**	**Right kidney**	**Left kidney**	**Esophagus**	**Gallbladder**	**Liver**	**Stomach**	**Aorta**	**Inferior vena cava**	**Portal and splenic vein**	**Pancreas**	**Right adrenal gland**	**Left adrenal gland**	**Average**
Train set	0.962	0.951	0.95	0.935	0.922	0.968	0,945	0,932	0.914	0.881	0.887	0.939	0.931	0.932
Test set	0.954	0.948	0.949	0.743	0.719	0.96	0.885	0.881	0.80	0.764	0.779	0.641	0.632	0.82

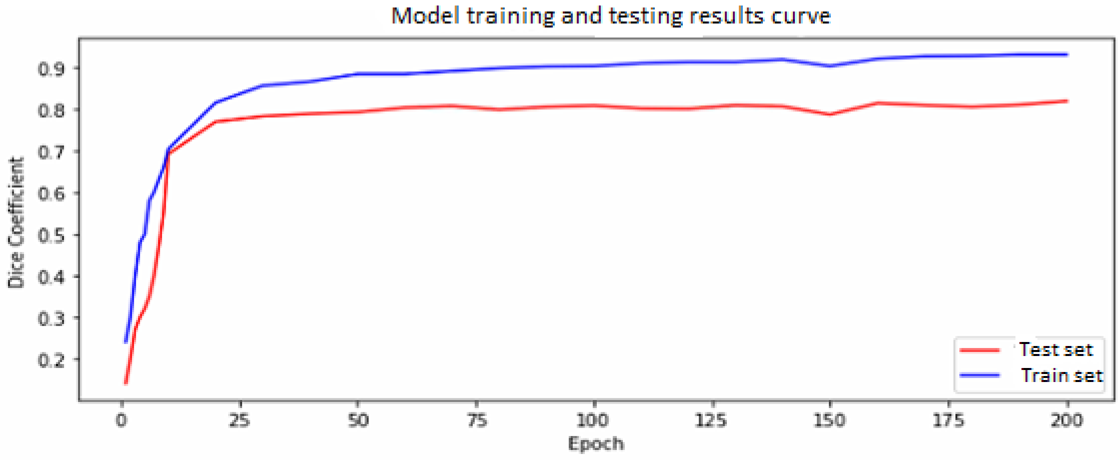


Figure 4. Multi-organ segmentation training and test result curves

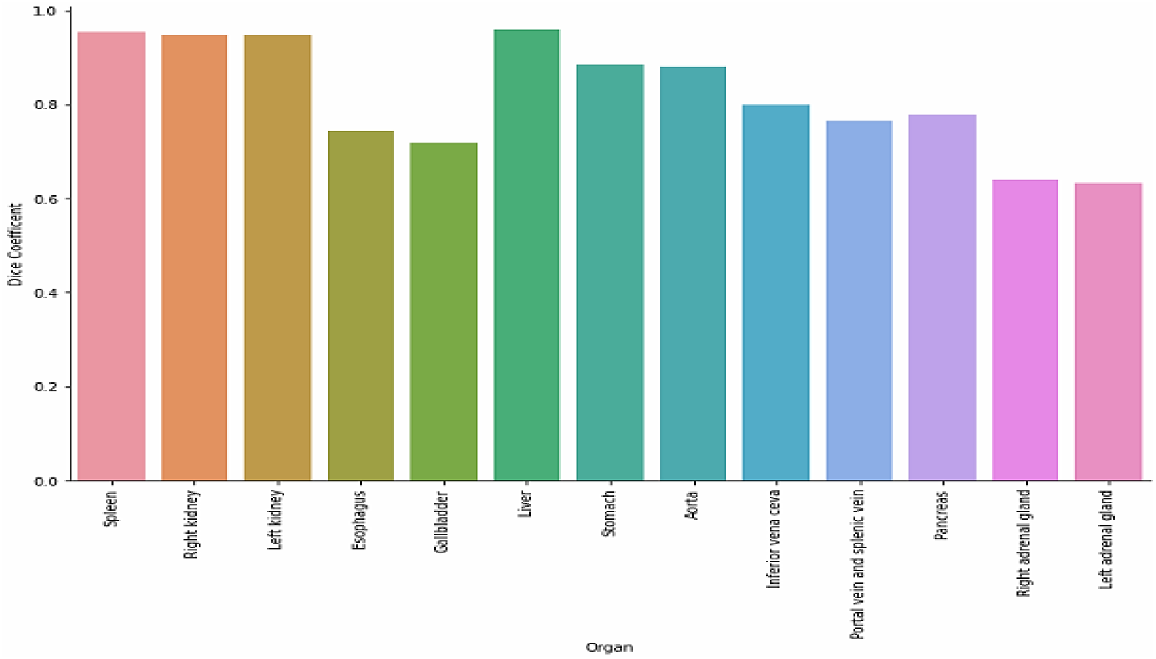


Figure 5. Test dice results of each organ

Figure 6 shows the predicted segmentation mask images and actual mask images of the final test results. These images are slices of a CT image. In addition, each organ is numbered. The confusion matrix of the fusion model is given in Figure 7. The confusion matrix shows the number of correct and incorrect pixels of each organ. Precision recall, f1 score and accuracy results of the fusion model are given in Table 5. These metrics were calculated for each organ using the pixel counts from the confusion matrix.

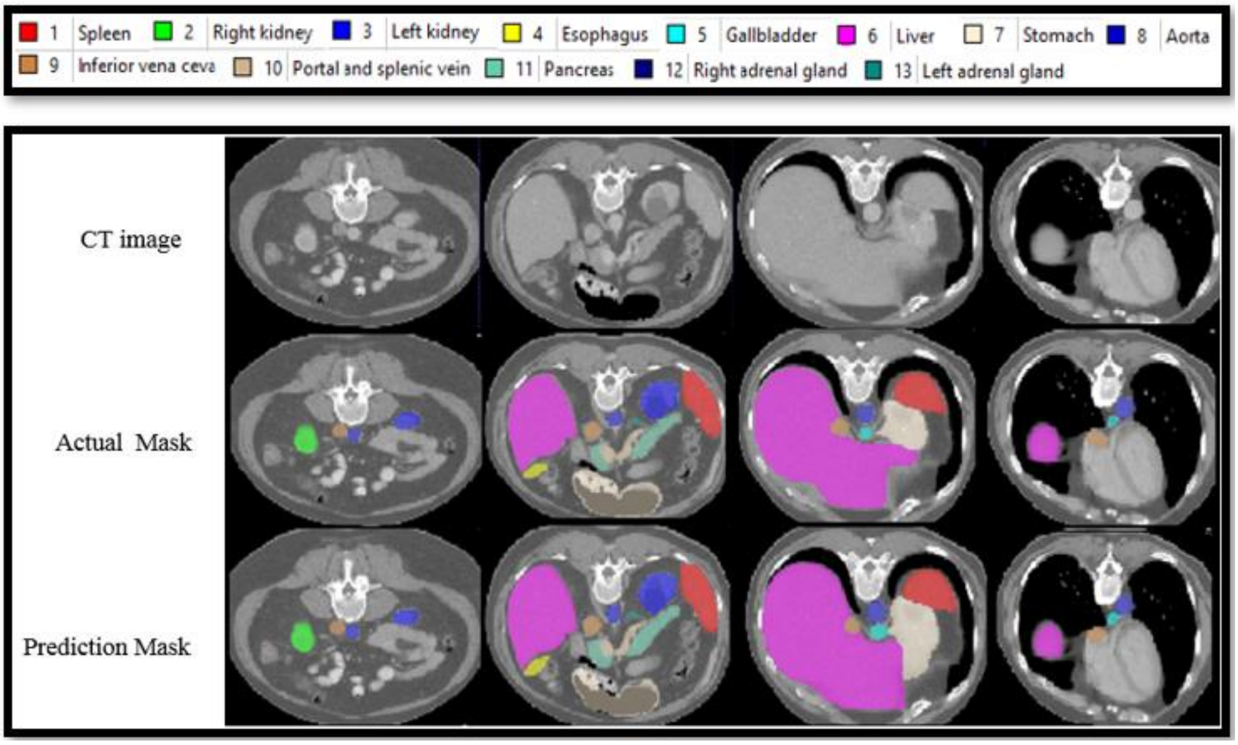
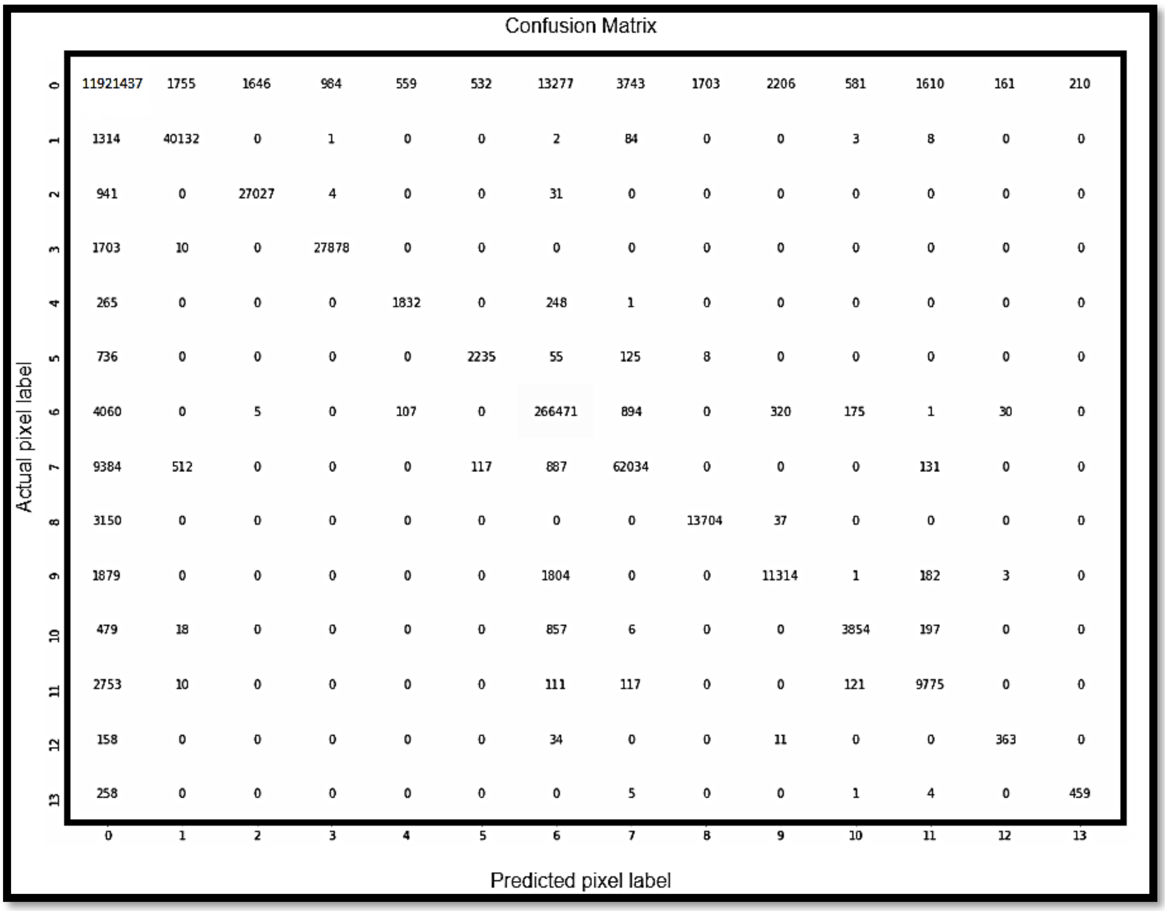


Figure 6. CT Image, actual mask and prediction mask



Confusion Matrix

Actual pixel label

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	11921437	1755	1646	984	559	532	13277	3743	1703	2206	581	1610	161	210
1	1314	40132	0	1	0	0	2	84	0	0	3	8	0	0
2	941	0	27027	4	0	0	31	0	0	0	0	0	0	0
3	1703	10	0	27878	0	0	0	0	0	0	0	0	0	0
4	265	0	0	0	1832	0	248	1	0	0	0	0	0	0
5	736	0	0	0	0	2235	55	125	8	0	0	0	0	0
6	4060	0	5	0	107	0	266471	894	0	320	175	1	30	0
7	9384	512	0	0	0	117	887	62034	0	0	0	131	0	0
8	3150	0	0	0	0	0	0	0	13704	37	0	0	0	0
9	1879	0	0	0	0	0	1804	0	0	11314	1	182	3	0
10	479	18	0	0	0	0	857	6	0	0	3854	197	0	0
11	2753	10	0	0	0	0	111	117	0	0	121	9775	0	0
12	158	0	0	0	0	0	34	0	0	11	0	0	363	0
13	258	0	0	0	0	0	0	5	0	0	1	4	0	459

Predicted pixel label

Figure 7. Confusion matrix for multi-organ segmentation

Table 5. Precision, recall, f1-score, accuracy results of each organ

Evaluation metrics	Spleen	Right kidney	Left kidney	Esophagus	Gallbladder	Liver	Stomach	Aorta	Inferior vena cava	Portal and splenic vein	Pancreas	Right adrenal gland	Left adrenal gland	Average
Precision	0.966	0.965	0.942	0.78	0.707	0.979	0.849	0.811	0.745	0.712	0.758	0.641	0.631	0.806
Recall	0.945	0.942	0.965	0.733	0.774	0.939	0.981	0.889	0.814	0.813	0.82	0.65	0.686	0.842
F1 -Score	0.955	0.953	0.953	0.755	0.738	0.958	0.91	0.848	0.777	0.759	0.787	0.645	0.657	0.823
Accuracy	0.954	0.948	0.949	0.743	0.719	0.96	0.885	0.881	0.80	0.764	0.779	0.641	0.632	0.82

5. DISCUSSION

The dataset was initially trained using a single-stage 3D U-Net based model. This model was trained separately on both grayscale images and HSV images. Finally, the dataset was trained with a fusion model combining different color spaces (HSV and grayscale). In Table 6, the test results of the 3D U-Net with HSV, 3D U-Net with Grayscale, and the fusion model are compared.

Table 6. Comparison of 3D U-Net with grayscale, 3D U-Net with HSV, and fusion model

	Dice Coefficient													
Model	Spleen	Right kidney	Left kidney	Esophagus	Gallbladder	Liver	Stomach	Aorta	Inferior vena cava	Portal and splenic vein	Pancreas	Right adrenal gland	Left adrenal gland	Average
3D U-Net with Grayscale	0.92	0.936	0.895	0.587	0.709	0.956	0.857	0.87	**0.82**	0.735	0.718	**0.642**	0.588	0.787
3D U-Net with HSV	0.938	0.932	0.927	0.71	**0.72**	0.956	0.832	**0.896**	0.817	0.703	0.644	0.583	0.554	0.786
Fusion Model	**0.954**	**0.948**	**0.949**	**0.743**	0.719	**0.96**	**0.885**	0.881	0.80	**0.764**	**0.779**	0.641	**0.632**	**0.82**

In Table 7, the fusion model is compared with the results presented in Larsson et al. [15] and with the results of the model proposed by Kayhan[12]. The fusion model used in this study is a fine-tuned version of the 3D U-Net model proposed by Kayhan. The model results presented in the study of Larsson et al. and Kayhan were obtained from the MICCIA 2015 data set used in this study. The CNN and FCN architectures in Table 7 were developed by Larsson et al. [15]. IMI and CLS models are the two models that gave the best results in the "Multi-Atlas Abdomen Labeling Challenge" competition. The fusion model outperformed other models in terms of segmentation accuracy for all organs except two (inferior vena cava and right adrenal gland). The IMI model had the highest correct prediction rate for the inferior vena cava, while the FCN model had the highest correct prediction rate for the right adrenal gland. The fusion model ranks 2nd in accuracy of the Inferior vena cava and right adrenal gland organs. In addition, the fusion model gave the highest segmentation result in the mean of all organs, and it was observed that the fine-tuning made to the 3D U-Net model proposed by Kayhan increased the performance.

Table 7. Comparison of the results of the fusion model with the results of other models

Dice Coefficient															
Model	**Spleen**	**Right kidney**	**Left kidney**	**Esophagus**	**Gallbladder**	**Liver**	**Stomach**	**Aorta**	**Inferior vena cava**	**Portal and splenic vein**	**Pancreas**	**Right adrenal gland**	**Left adrenal gland**	**Average**	**Number of successful organs**
CNN[15]	0.93	0.866	0.911	0.624	0.662	0.946	0.775	0.860	0.776	0.567	0.602	0.631	0.583	0.75	-
FCN[15]	0.936	0.897	0.911	0.613	0.588	0.949	0.764	0.87	0.72	0.758	0.715	**0.646**	0.63	0.767	1
CLS (MICCAI 2015)	0.911	0.893	0.901	0.375	0.607	0.940	0.704	0.811	0.76	0.649	0.643	0.557	0.582	0.723	-
IMI (MICCAI 2015)	0.919	0.901	0.914	0.604	0.692	0.948	0.805	0.857	**0.828**	0.754	0.74	0.615	0.623	0.790	1
Kayhan's Model[12]	0.94	0.934	0.937	0.698	0.703	0.951	0.847	0.873	0.816	0.698	0.774	0.611	0.558	0.796	-
Fusion Model	**0.954**	**0.948**	**0.949**	**0.743**	**0.719**	**0.96**	**0.885**	**0.881**	0.80	**0.764**	**0.779**	0.641	**0.632**	**0.82**	**11**

The fusion model results in Table 8 are compared with the results of the state-of-the-art models (Swin-Unet [30], TransUNet [31], LeViT-UNet [32], MISSFormer [33], CoTr [34], nnFormer [35], nnU-Net [36], UNETR [37], Swin UNETR [38]) on the MICCIA 2015 dataset. When the results are examined, it is seen that the fusion model is at a level to compete with state-of-the-art models.

Table 8. Comparison of the results of fusion model with the results of state-of-the-art models

Model	**Spleen**	**Right kidney**	**Left kidney**	**Gallbladder**	**Liver**	**Stomach**	**Aorta**	**Pancreas**
Swin-Unet	0.906	0.796	0.832	0.665	0.942	0.766	0.854	0.565
TransUNet	0.936	0.77	0.818	0.631	0.94	0.764	0.872	0.558
LeViT-UNet	0.888	0.802	0.846	0.622	0.931	0.727	0.873	0.59
MISSFormer	0.919	0.82	0.852	0.686	0.944	0.808	0.869	0.656
CoTr	0.922	0.864	0.853	**0.814**	0.968	0.76	0.921	0.802
nnFormer	0.898	0.87	0.875	0.781	0.954	0.825	0.89	**0.819**
nnU-Net	0.923	0.897	0.848	0.806	0.971	0.823	0.928	0.82
UNETR	0.861	0.797	0.813	0.698	0.942	0.762	0.889	0.589
Swin UNETR	0.887	0.891	0.852	0.765	**0.969**	0.797	**0.927**	0.772
Fusion Model	**0.954**	**0.948**	**0.949**	0.719	0.96	**0.885**	0.881	0.779

Multi organ segmentation was performed in this study. However, this study has limitations. The small size of the dataset may restrict the model's ability to accurately detect certain organ. Additionally, the fusion model, which combines different color spaces, may increase computational costs. However, this approach has provided a unique perspective in the literature by allowing for more comprehensive image analysis and improved detection of organ boundaries. This approach could be a roadmap for similar applications in the future.

6. CONCLUSION

In this study, a model with early fusion approach is used to automatically perform multi-organ segmentation on CT images. Experimental studies and fine tuning were carried out to determine the model that gives better results. Firstly, a single-stage 3D U-Net model was trained for multi-organ segmentation with only Grayscale and only HSV images with the selected parameter set. The performance of the 3D U-Net model with grayscale, the 3D U-Net model with HSV, and the fusion model were compared. The 3D U-Net test set accuracy rate with grayscale is 0.787, the 3D U-Net test set accuracy rate with HSV is 0.786, and the test set accuracy rate of the fusion model is 0.82. In the fusion model, segmentation accuracy of the spleen, right kidney, left kidney, and liver is 90%, stomach, aorta, and inferior vena cava segmentation accuracy is 80%, esophagus, gallbladder, portal, and spleen vein, and pancreas segmentation accuracy is over 70%. The right and left adrenal glands, which give the lowest segmentation result among the organs, are over 60%. The fusion model achieved a high segmentation success rate in large-volume organs. It has been observed that the success of segmentation is low in small volume organs (right adrenal gland and left adrenal gland).

When this study is evaluated in general, the segmentation of organs is the first step to examining the internal structure of the organs. As a result of segmentation, various diseases can be diagnosed. However, the segmentation and classification of organs by radiologists is difficult and time-consuming because the shapes of the organs vary. In addition, since it requires knowledge and experience, the rate of making mistakes is high. To overcome these difficulties, a fusion model based on 3D U-Net combining different color spaces was used for automatic multi-organ segmentation on CT images. In addition, this fusion model was compared with state-of-the-art models made in this field. As a result, successful and promising results were obtained.

For future work, increasing the diversity of data used in multi-organ segmentation and incorporating attention mechanisms may improve the performance for small-sized organs. Additionally, using computers with high hardware capabilities to increase the resolution and number of slices in the images may also lead to better segmentation results.

Declaration of Ethical Standards

The authors conducted this study in accordance with all ethical standards.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Data Availability

The datasets analyzed during the current study are available in the [syn3193805] repository, [https://www.synapse.org/#!Synapse:syn3193805/wiki/217760].

Acknowledgement

This manuscript has been prepared from Beyza Kayhan's master's thesis supervised by Sait Ali UYMAZ and was approved by the Graduate Education Institute of Konya Technical University on January 7, 2022. Beyza KAYHAN designed the model and the computational framework and carried out the experiment and wrote the manuscript with support from Sait Ali UYMAZ. Sait Ali UYMAZ designed

and supervised the project.

7. REFERENCES

[1] N. Shen *et al.*, "Multi-organ segmentation network for abdominal CT images based on spatial attention and deformable convolution," *Expert Systems with Applications,* vol. 211, p. 118625, 2023.

[2] Y. Wang, Y. Zhou, W. Shen, S. Park, E. K. Fishman, and A. L. Yuille, "Abdominal multi-organ segmentation with organ-attention networks and statistical fusion," *Medical image analysis,* vol. 55, pp. 88-102, 2019.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature,* vol. 521, no. 7553, pp. 436-444, 2015.

[4] A. Şeker, B. Diri, and H. H. Balık, "A review about deep learning methods and applications," *Gazi J Eng Sci,* vol. 3, no. 3, pp. 47-64, 2017.

[5] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Computer vision and image understanding,* vol. 189, p. 102805, 2019.

[6] H.-S. Bae, H.-J. Lee, and S.-G. Lee, "Voice recognition based on adaptive MFCC and deep learning," in *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, 2016: IEEE, pp. 1542-1546.

[7] S. Caldera, A. Rassau, and D. Chai, "Review of deep learning methods in robotic grasp detection," *Multimodal Technologies and Interaction,* vol. 2, no. 3, p. 57, 2018.

[8] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis,* vol. 42, pp. 60-88, 2017.

[9] M. Toğaçar and B. Ergen, "Biyomedikal Görüntülerde Derin Öğrenme ile Mevcut Yöntemlerin Kıyaslanması," *Fırat Üniversitesi Mühendislik Bilimleri Dergisi,* vol. 31, no. 1, pp. 109-121, 2019.

[10] H. R. Roth *et al.*, "A multi-scale pyramid of 3D fully convolutional networks for abdominal multi-organ segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*, 2018: Springer, pp. 417-425.

[11] T. Ghosh, L. Li, and J. Chakareski, "Effective deep learning for semantic segmentation based bleeding zone detection in capsule endoscopy images," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018: IEEE, pp. 3034-3038.

[12] B. Kayhan, "Deep learning based multiple organ segmentation in computed tomography images," Master's thesis, Konya Technical University, 2022.

[13] B. Kayhan and S. A. Uymaz, "Multi Organ Segmentation in Medical Image.," *Current Studies in Healthcare and Technology* pp. 59-72, 2023.

[14] R. Trullo, C. Petitjean, D. Nie, D. Shen, and S. Ruan, "Joint segmentation of multiple thoracic organs in CT images with two collaborative deep architectures," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, 2017: Springer, pp. 21-29.

[15] M. Larsson, Y. Zhang, and F. Kahl, "Robust abdominal organ segmentation using regional convolutional neural networks," *Applied Soft Computing,* vol. 70, pp. 465-471, 2018.

[16] H. R. Roth *et al.*, "Deep learning and its application to medical image segmentation," *Medical Imaging Technology,* vol. 36, no. 2, pp. 63-71, 2018.

[17] C. Shen *et al.*, "On the influence of Dice loss function in multi-class organ segmentation of abdominal CT using 3D fully convolutional networks," *arXiv preprint arXiv:1801.05912,* 2018.

[18] H. Kakeya, T. Okada, and Y. Oshiro, "3D U-JAPA-Net: mixture of convolutional networks for abdominal multi-organ CT segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*, 2018: Springer, pp. 426-433.

[19] S. Vesal, N. Ravikumar, and A. Maier, "A 2D dilated residual U-Net for multi-organ segmentation in thoracic CT," *arXiv preprint arXiv:1905.07710,* 2019.

[20] O. Mietzner and A. Mastmeyer, "Automatic multi-object organ detection and segmentation in abdominal CT-data," *medRxiv,* p. 2020.03. 17.20036053, 2020.

[21] B. Rister, D. Yi, K. Shivakumar, T. Nobashi, and D. L. Rubin, "CT-ORG, a new dataset for multiple organ segmentation in computed tomography," *Scientific Data,* vol. 7, no. 1, p. 381, 2020.

[22] Y. Liu *et al.*, "CT-based multi-organ segmentation using a 3D self-attention U-net network for pancreatic radiotherapy," *Medical physics,* vol. 47, no. 9, pp. 4316-4324, 2020.

[23] X. Fang and P. Yan, "Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction," *IEEE Transactions on Medical Imaging,* vol. 39, no. 11, pp. 3619-3629, 2020.

[24] F. Zhang, Y. Wang, and H. Yang, "Efficient context-aware network for abdominal multi-organ segmentation," *arXiv preprint arXiv:2109.10601,* 2021.

[25] H. Kaur, N. Kaur, and N. Neeru, "Evolution of multiorgan segmentation techniques from traditional to deep learning in abdominal CT images–A systematic review," *Displays,* vol. 73, p. 102223, 2022.

[26] Z. Xu. "Multi-atlas labeling beyond the cranial vault - workshop and challenge." https://www.synapse.org/#!Synapse:syn3193805/wiki/217760. (accessed 6.3.2021).

[27] X. Gao, Y. Qian, and A. Gao, "COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models," *arXiv preprint arXiv:2107.01682,* 2021.

[28] V. Chernov, J. Alander, and V. Bochko, "Integer-based accurate conversion between RGB and HSV color spaces," *Computers & Electrical Engineering,* vol. 46, pp. 328-337, 2015.

[29] L. Wang, C. Wang, Z. Sun, and S. Chen, "An improved dice loss for pneumothorax segmentation by mining the information of negative areas," *IEEE Access,* vol. 8, pp. 167939-167949, 2020.

[30] H. Cao *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*, 2022: Springer, pp. 205-218.

[31] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306,* 2021.

[32] G. Xu, X. Wu, X. Zhang, and X. He, "Levit-unet: Make faster encoders with transformer for medical image segmentation," *arXiv preprint arXiv:2107.08623,* 2021.

[33] X. Huang, Z. Deng, D. Li, and X. Yuan, "Missformer: An effective medical image segmentation transformer," *arXiv preprint arXiv:2109.07162,* 2021.

[34] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, 2021: Springer, pp. 171-180.

[35] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nnformer: Interleaved transformer for volumetric segmentation," *arXiv preprint arXiv:2109.03201,* 2021.

[36] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods,* vol. 18, no. 2, pp. 203-211, 2021.

[37] A. Hatamizadeh *et al.*, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574-584.

[38] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*, 2021: Springer, pp. 272-284.