INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Segmentation of Plain CT Image of Ischemic Lesion based on Trans-Swin-UNet

Zhiqiang Luo [a,*], Tek Yong Lim [b], Xia Hua [a]

[a] School of Design, Foshan University, 18 Jiang Wan Yi Lu, Foshan, China
[b] Faculty of Computing & Informatics, Multimedia University, Cyberjaya Campus, Persiaran Multimedia, Cyberjaya, Malaysia
Corresponding author: *luozq@fosu.edu.cn

Abstract*— The present study aims to build a hybrid convolutional neural network and transformer UNet-based model, Trans-Swin-UNet, to segment ischemic lesions of the plain computed tomography (CT) image. The model architecture is built based on TransUnet and has four main improvements. First, replace the decoder of TransUNet with a Swin transformer; second, add a Max Attention module into the skip connection; third, design a comprehensive loss function; and last, speed up the segmentation performance. The present study designs two experiments to evaluate the performance of the built model using both the self-collected and public plain CT image datasets. The model optimization experiment evaluates the improvements of Trans-Swin-UNet over TransUnet. The experimental results show that each improvement of the built model can achieve a better performance than TransUNet in terms of dice similarity coefficient (DSC), Jaccard coefficient (JAC), and accuracy (ACC). The comparison experiment compares the built model with four existing UNet-based models. The experimental results show that the built model had a DSC of 0.72±0.01, a JAC of 0.78±0.04, an ACC of 0.75±0.03 using the self-collected plain CT image dataset and a DSC of 0.73±0.02, a JAC of 0.79±0.03, an ACC of 0.76±0.02 using the public plain CT image dataset, achieving the best segmentation performance among five UNet-based neural network models. The two experimental results conclude that the built model could accurately segment ischemic lesions of the plain CT image. The limitations and future work of this study are also discussed.

***Keywords*— TransUNet; medical image segmentation; ischemic lesion; swin transformer; attention gate.**

Manuscript received 5 Jun. 2024; revised 24 Aug. 2024; accepted 12 Oct. 2024. Date of publication 30 Nov. 2024.

I. INTRODUCTION

Stroke is the second leading cause of death and the third leading cause of disability worldwide [1]. Ischemic stroke due to the blockage of blood vessels in the brain causes nearly 6 million deaths worldwide every year and is the leading cause of death and disability in Chinese [2]. The medical image plays a crucial role in diagnosing ischemic stroke, and the plain computed tomography (CT) image is the primary examination method for ischemic stroke patients due to its accuracy in identifying the disease, fast imaging speed, and economical cost [3]. Precisely segmenting ischemic lesions in plain CT images is critical in diagnosing an ischemic stroke. Traditional segmentation through doctor's scanning has problems, such as being time-consuming and unstable.

In recent years, automatic segmentation methods based on deep-learning neural networks models, such as convolutional neural networks (CNN) or Transformer, have been proposed and implemented in medical image segmentation [4]. However, the segmentation of a plain CT image of an ischemic lesion still faces many challenges, given the complex and varied imaging features of ischemic lesions, which are easily affected by other brain diseases [5].

UNet has been pivotal in medical image segmentation, especially for tasks with limited data availability [6]. Since it was initiated in 2015, many UNet-based neural network models have been proposed and applied to segmenting brain tumors, organs, and lesions [7]. The following content presents the trajectory of UNet's evolution within the last ten years, focusing on three major phases: the CNN-based UNets, the Transformer-based UNets, and the hybrid CNN-Transformer UNets.

A. CNN-based UNets

The first UNet architecture [6] consists of a contracting path to capture context and a symmetric expanding path to enable precise localization. The contracting path repeats the application of two 3x3 CNNs called Encoder, and the expanding path consists of samples of the feature map followed by a 2x2 CNN called Decoder. This innovative approach quickly set a new standard for medical image segmentation.

The CNN-based UNet models subsequently introduce sophisticated features to enhance performance. For example, the UNet structure is enhanced with residual connections, facilitating deeper network training by improving feature propagation and gradient flow [8]. 3D UNet [9] enables the effective segmentation of volumetric data such as CT and MRI scans. V-Net [10] introduces a dice coefficient loss function optimized for 3D medical image data. UNet++ [11] is a revised architecture incorporating nested, dense skip pathways to improve the network's gradient flow and feature propagation, leading to more precise segmentation results.

B. *Transformer-based UNets*

When the Vision Transformer (ViT) [12] is incorporated into UNet architecture to replace CNN, the ViT UNet-based neural network models could increase cross-attention for medical image segmentation [13]. For example, UNETR [14] utilizes the transformer's strength in modeling complex dependencies across the entire volume of medical data, resulting in improved performance in segmenting 3D structures such as brain tumors and other organs.

The Swin-Unet [15] incorporates the Swin Transformer into a UNet architecture to improve the image segmentation performance. The Swin Transformer provides a shift window mechanism to connect different parts of the image, rather than the sequence parts required in ViT, which facilitates the long-range dependency capture. Thus, Swin-Unet could particularly segment delicate structures in medical images due to its ability to scale efficiently to different image resolutions.

C. *Hybrid CNN-Transformer UNets*

The hybrid CNN-Transformer UNets integrate both CNN and Transformer in UNet architectures. The goal is to leverage the local feature extraction capabilities of CNNs and the global dependency modeling of Transformers. For example, the pioneering hybrid UNet model [16] utilizes CNNs for initial feature extraction, followed by Transformers for deeper feature integration. This approach excels in various segmentation tasks, significantly improving the traditional model. Furthermore, TransBTS [17] leverages multiple CNN layers to capture local features and Transformer layers to capture the global context in the encoder. By effectively integrating the spatial relationships across different imaging modalities, TransBTS highlights the potential of Transformers in handling the complexities of multimodal medical data. The hybrid model of Xie et al. [18] could segment complex anatomical structures in 3D medical images, leveraging local and global information, while the hybrid model of Xu et al.[19] integrating a residual connection before and after the Transformer encoder reduces the information loss. Li et al. [20] developed one CNN-Transformer hybrid network that offers robust segmentation capabilities under challenging imaging conditions, such as variations in noise, contrast, and deformations.

The present study pays attention to the hybrid CNN-Transformer UNet, TransUNet [21]. The architecture of TransUNet begins with a CNN-based feature extractor that reduces the spatial dimensions of the input image while increasing the feature dimensions. The resultant feature maps are then flattened and fed into Transformer blocks, which further process these features through self-attention mechanisms. Specifically, the Transformer used in TransUNet follows the encoder design from the Vision Transformer (ViT), which processes sequences of linear embeddings of patches of feature maps. Then, feature maps from various stages of the transformer encoder are passed to the corresponding layers in the decoder. The decoder consists of several up-sampling layers with applications of the 3x3 CNN that gradually reconstructs the spatial resolution of the feature maps.

Applying more advanced Transformer models, such as those incorporating newer attention mechanisms or graph-based approaches, could more accurately enhance understanding and segmenting of complex anatomical structures. Therefore, the present study aims to build and evaluate a UNet-based neural network model, Trans-Swin-UNet. Its architecture adopts the above hybrid CNN-Transformer UNets, TransUNet, but has four main improvements. The paper’s remaining contents first introduce the detailed architecture of the built Trans-Swin-UNet model, present the preprocessing methods, and the two experiments designed to test the built model. Second, the paper presents the details of the results of the experiments. Last, the conclusion summarizes the main work of this study and indicates the feature work.

II. Materials and Method

The section first presents the details of the built Trans-Swin-UNet model, which has four main improvements on the existing TransUNet model: using the Swin Transformer for the decoder, designing one new max-attention model, building a comprehensive loss function, and speeding up the segmentation. Then, the three methods of preprocessing plain CT images are presented. Last, the two experiments test the advantages of each improvement of Trans-Swin-UNet over TransUNet and the super performance of Trans-Swin-UNet over four existing UNet-based models. The CT image datasets for the test experiment include self-collected data from a local hospital and public CT images from the Kaggle website.

A. *Trans-Swin-UNet model*

Fig.1 illustrates the structure of the proposed Trans-Swin-UNet model, which adopts an encoder-decoder UNet structure. The encoder consists of four down-sampling CNNs and Transformer modules, and the decoder comprises three Swin Transformer modules and three patch-expanding operations. The feature map size is doubled, and the Max Attention module concatenates with the original CNN segmented images. Finally, the predicted lesion mask is output, and the loss function is calculated to iterate the neural network model.

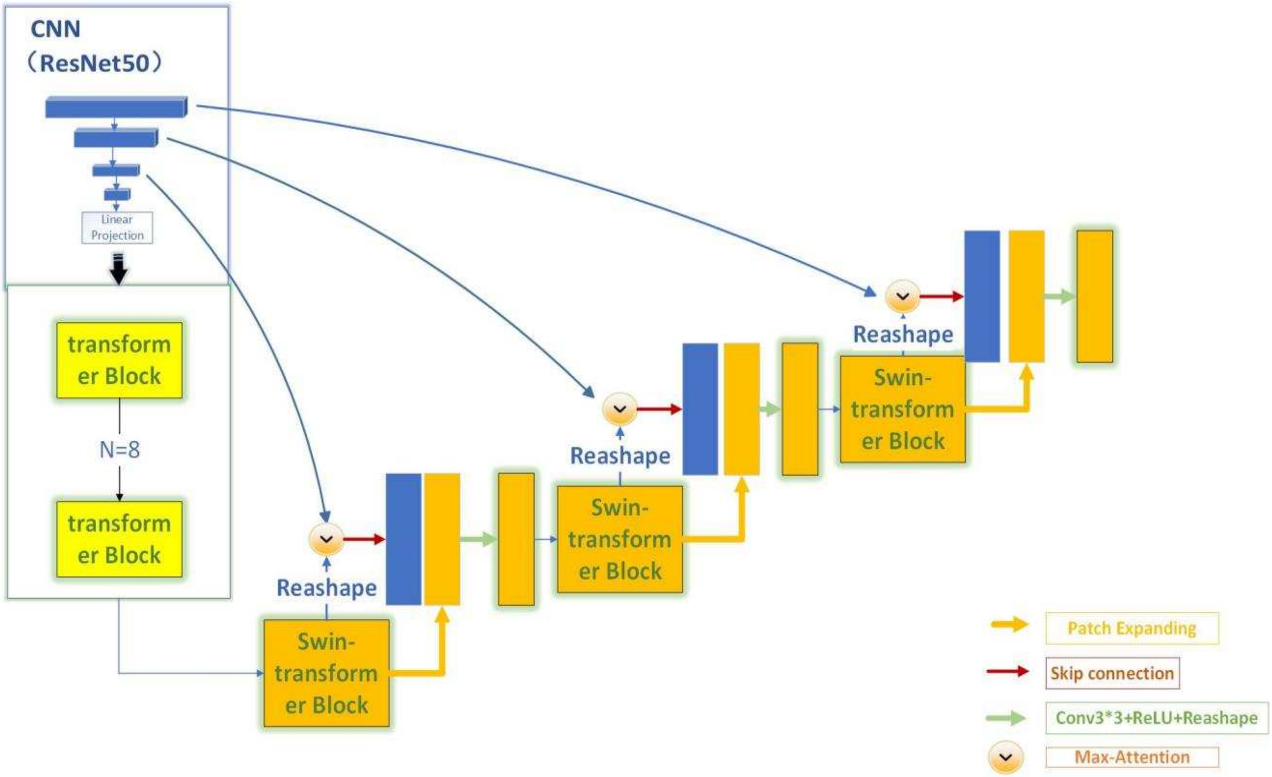

Fig. 1 The structure of the Trans-Swin-UNet model

1) Swin Transformer Decoder: The Swin transformer has four modules: Layer Norm (LN) layer, W-MSA and SW-MSA, and 2-layer MLP layer [15], as shown in Fig. 2. The Swin transformer decoder in the improved model replaces the original TransUNet decoder, which has the following two advantages:

- *Expanding global information acquisition*

Exchanging information between windows at different levels ensures the model can capture global contextual details in the image, which is crucial for segmenting early cerebral infarction lesions. The Shifted Windows Multi Head Self Attention (SW-MSA) module is introduced, an improved version of W-MSA with offset processing.

W-MSA restricts the attention calculation within a fixed window, while SW-MSA enables information interaction and transmission between adjacent windows through the shifted window approach. It helps the model capture richer contextual information and improve its performance.

- *Improving computational efficiency*

Swin transformers introduce a shifted window mechanism to limit self-attention computation within local windows. This mechanism allows the model to maintain efficient computational speed even when processing large-scale images. Thus, it could significantly reduce computational complexity.

The W-MSA module first divides the input feature map into predetermined window sizes M, with each window having a width and height of M. This way, the original feature map with height h and width w will be divided into h/M x w/M windows. Subsequently, each window independently applies a multi-head attention module for self-attention calculation. Just calculate the computational cost of the feature map with height h, width w, and depth c as follows: $4hwc^2 + 2\ (hw)\ ^2c$. Here, the height of each window is *M,* and the width is *M*, then the computational cost of the future map is $4(Mc)^2 + 2\ M^4c$.

The W-MSA module achieves local self-attention calculation through window partitioning, improving computational efficiency. This improvement is significant for processing high-resolution images or large datasets, as it can significantly reduce the use of computing resources and memory while maintaining model performance.

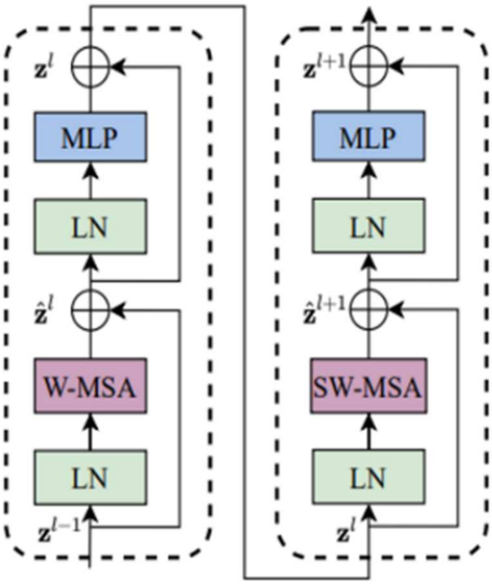

Fig. 2 The structure of the Swin transformer module

2) Max Attention Model: The Max Attention module consists of an Attention Gate mechanism [16], [22] and a Pyramid Pooling module (PPM) [23]. Attention Gate is a mechanism used to fuse features of different scales in deep learning models. It can effectively reduce the loss of spatial information during the down-sampling process, especially when dealing with small and scattered features such as acute cerebral infarction lesions.

3) Max Attention Model: The Max Attention module consists of an Attention Gate mechanism [16], [22] and a Pyramid Pooling module (PPM) [23]. Attention Gate is a mechanism used to fuse features of different scales in deep learning models. It can effectively reduce the loss of spatial information during the down-sampling process, especially when dealing with small and scattered features such as acute cerebral infarction lesions.

Fig. 3 shows the inside operation of the Max Attention module. First, g and x are operated in parallel, where g comes from the Decoder part of the model, and x comes from the CNN part. The size g is 1/2 of x, so x is down-sampled, or g is up-sampled. Enable A and B to perform point-by-point "+" operations. G obtains A through Wg. X obtains B through Wx, followed by the operation of A+B to obtain C. Second, C performs the ReLU operation to obtain D. D performs ψ Operation yields E. E performs the Swish operation to obtain F. F obtains the attention coefficient through resampling α (Attention coefficient is the attention weight). Final attention coefficient α multiplied by x to obtain the final result.

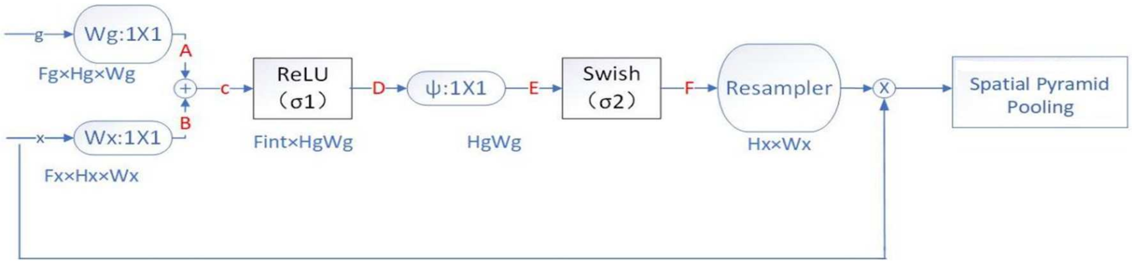

Fig. 3 The structure of the Max Attention module

The Swish function helps optimize the training process, making converging easier. Its design combines the advantages of Sigmoid and ReLU functions to provide better performance in deep learning. The output of the Swish function tends to approach 0 in negative regions, while it has a larger output in positive regions. This nonlinear characteristic enables the Swish function to fit complex nonlinear relationships better, thereby improving the model's expressive power.

The smoothness characteristics of the Swish function also help accelerate the training convergence process. Since the Swish function is differentiable across the entire range of real numbers, it can avoid the problem of vanishing or exploding gradients during the training process.

PPM is a technique used to capture multi-scale contextual information [19]. Its core idea is introducing pooling operations in convolutional neural networks at different scales to fuse multi-scale feature information. Fig. 4 shows the structure of the PPM module, which includes three processes. First, the pyramid pooling module performs pooling operations on the input feature maps at different scales, generating multiple feature maps of different sizes. By pooling at various scales (e.g., 1x1, 2x2, 3x3, 6x6), feature information of different resolutions can be captured, thereby enriching the model's ability to understand the context.

Second, the pyramid pooling module up-samples feature maps of different scales to maintain their size consistent with the original feature map. The up-sampled feature map is concatenated with the original feature map in the channel dimension to form a composite feature map containing multi-scale information.

Finally, the pyramid pooling module uses a convolutional layer to compress the channel dimension of the concatenated feature map, which could maintain the number of channels in the output feature map consistent with the original feature map. This convolutional layer can be a 1x1 convolution used to reduce the number of channels and fuse feature information of different scales.

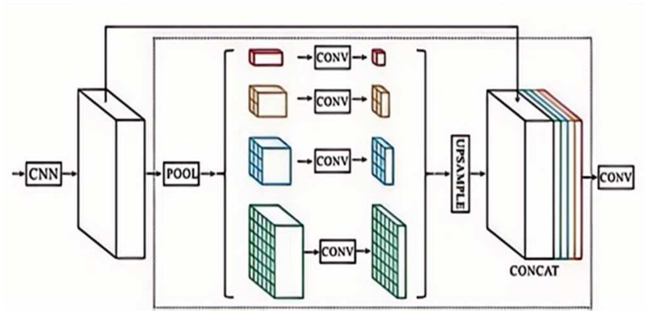

Fig. 4 The structure of the Pyramid Pooling module

The model can capture global and local feature information through the pyramid pooling module, improving its recognition ability for targets of different scales. This is particularly important in image segmentation, as segmenting the lesions of a plain CT image requires the model to understand and process lesions of various sizes and shapes.

4) Loss Function: The present study builds a comprehensive loss function, the Total Loss function, which integrates two loss functions, Cross Entropy Loss and Dice Loss [24], shown in equation (1).

$$Total\ Loss = 0.7 \times Dice + CrossEntropy\ Loss \quad (1)$$

Cross entropy loss measures the difference between the predicted probability distribution and the true probability distribution. When there is an imbalance between categories, the standard cross entropy loss may not be sufficient to segment medical images accurately. Its definition is shown in equation (2):

$$L = \frac{1}{N}\sum_i -[y_i \log(p_i) + (1-y_i) \times \log(1-p_i)] \quad (2)$$

where y_i is the sample label, p_i indicates the probability of the i-th sample.

Dice loss is a variation of the Dice coefficient used for loss calculation in optimization processes. The Dice coefficient measures the degree of overlap between predicted and true segmentation, as shown in equation (3). Dice loss is particularly effective in dealing with a class imbalance in medical image segmentation:

$$Dice = \frac{2\times(pred\cap true)}{pred\cup true} = \frac{2\times intersection}{union} \quad (3)$$

where pred represents the set of predicted values, true means the set of true values. The numerator calculates the intersection size between the sets of pred and true. Still, due to the repeated calculation of these common elements by the denominator during the calculation process, it needs to be multiplied by 2 for correction.

Although Dice loss is the preferred loss function for dealing with imbalanced classes, it may cause instability and oscillation during model training when the target area in the medical image is tiny (i.e., there are very few foreground pixels). Once a small portion of foreground pixels are misclassified, the Dice loss value will significantly change, leading to an unstable gradient.

5) Speed optimization of segmentation. The present study uses the first four modules of ResNet50 (STAGE0-STAGE3) to replace the first three modules of ResNet50 of TransUnet (STAGE0-STAGE2) to increase the receptive field of the CNN module [25], [31]. CNN needs to set the receptive field reasonably based on the target size and content of the medical image. By adjusting the convolutional layer step size, convolutional kernel size, and pooling layer settings, the receptive field size can be controlled, improving the model's adaptability to targets of different sizes and the accuracy of image segmentation. The present study also reduces the number of Transformer modules of the TransUnet model from 12 to 8 to reduce computational complexity, accelerate training and inference speed, and minimize the risk of overfitting.

B. Preprocessing plain CT images

The plain CT image must be processed before being sent to the Trans-Swin-UNet model for training and prediction. The following three preprocessing steps are usually implemented, and sometimes, the noises in the CT image may need to be removed [26].

1) Histogram Equalization: Histogram equalization involves statistical analysis of the frequency of each grayscale level in an image, then remapping each pixel's grayscale values based on this frequency information [27], [28]. Thus, pixels with similar grayscale values will be stretched apart after transformation, increasing the local contrast of the image. Meanwhile, as the transformation is global, it enhances the contrast of the entire image. This makes the details and contours in the image clearer and more visible. Fig. 5 compares plain CT images before and after histogram equalization. Through this operation, the contrast of plain CT images is significantly enhanced, the clarity is improved, and the structural details are richer.

2) Skull Removal: The skull removal method based on contour evolution is a segmentation technique for plain CT images designed to extract brain tissue from brain images [29]. This method utilizes the continuity and consistency of brain tissue between adjacent image slices and brain tissue's strength probability density functions (PDFs) to guide the segmentation process. Fig. 6 compares plain CT images before and after skull removal.

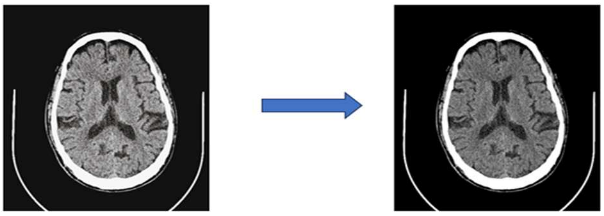

Fig. 5 The plain CT images before (left) and after (right) histogram equalization.

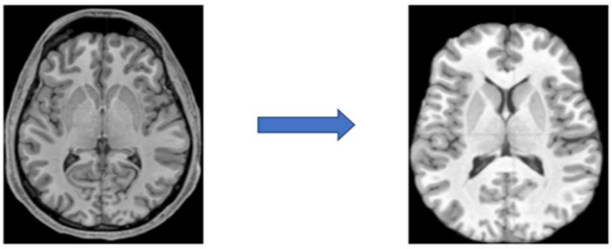

Fig. 6 The plain CT images before (left) and after (right) skull removal.

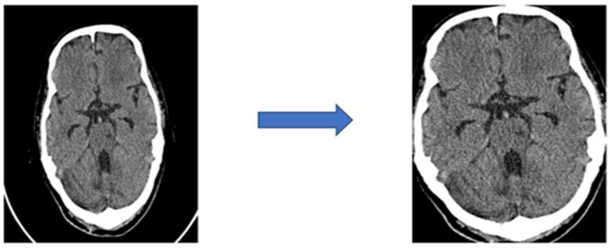

Fig. 7 The plain CT images before (left) and after (right) central region cropping.

3) Central Region Cropping: Central region cropping is an image processing technique commonly used in deep learning models to preprocess images [29]. The core idea of this technology is to crop out the central region from the original image and use it as input for the model. It could improve the performance of deep learning models since, in some cases, the central region in the image may contain the most important information, while the edge regions may contain noise or irrelevant information. By cropping out these edge regions, the model can focus more on crucial details in the image, thereby improving its accuracy. Fig. 7 compares plain CT images before and after central region cropping.

C. Experiment

1) Experiment Environment: The software and hardware used for the test experiment are shown in Table 1, which are installed on a server.

2) *Experiment Dataset*: The experiment used two plain CT image datasets. The first dataset, consisting of 189 sets, was self-collected from one local rehabilitation hospital. Each set contained clear lesion labels annotated by two experienced radiologists. The second dataset, consisting of 500 plain CT images with labeled lesions, was downloaded from the public Kaggle website.

TABLE I
THE SETTING FOR THE TEST ENVIRONMENT

Name	Configuration Parameters
Processor	CPU 3.2GHz (Intel Xeon E5-2680)
Memory	256GB
Graphics card	NVIDIA RTX A6000
OS	ubuntu 20.04
Software	Python v3.8, Pytorch v 1.6.0

4) *Experiment Design*: The present study designs the first experiment to explore the optimization performance of Trans-Swin-UNet over TransUnet, which tests each of four improvements: the Swin Transformer decoder, the max-attention module, the new loss function, and the optimized CNN layers. Then, the study designs the second experiment to compare the Trans-Swin-UNet with the current four UNet-based segmentation models, where each model's performance is evaluated through five-fold cross-validation. These two experiments deploy the same evaluation metrics [30], including the Dice similarity coefficient (DSC), Jaccard coefficient (JAC), and Accuracy (ACC).

III. RESULTS AND DISCUSSION

The section first presents the experimental results of the model optimization based on the four improvements. Then, it describes the experimental results of comparing five UNet-based segmentation models for the plain CT image.

A. Results of Model Optimization Experiments

1) Experiment with Swin Transformer Decoder: Table 2 and Table 3 show the experimental results comparing the performances between Trans-Swin-UNet and TransUnet, where the Swin Transformer decoder replaces the original Decoder. Although the Swin Transformer decoder increases computational complexity, the segmentation results are improved on both plain CT image datasets. The built model achieves a DSC value of 0.71±0.01, a JAC value of 0.75±0.05, and an Accuracy of 0.75±0.03 in the first plain CT image dataset and a DSC value of 0.72± 0.03, a JAC value of 0.76± 0.01, and an Accuracy of 0.76±0.02 in the second plain CT image dataset.

2) Experiment with Max Attention Module: The present study designs the new Max Attention module and adds the module at the skip connection of TransUNet. The new module is supposed to help the model extract more critical information from the input plain CT image. Tables 4 and 5 show the experimental results comparing the performances of four Trans-Swin-UNet models. The Max Attention module numbers used in the Trans-Swin-UNet are varied from 0 to 3. The results show that when the Max Attention module is added to all three skip connections, the model (Trans-Swin-UNet_D3) performs best on both plain CT image datasets. The built model with three Max Attention modules achieves a DSC value of 0.72±0.01, a JAC value of 0.71±0.03, and an Accuracy of 0.73±0.03 in the first plain CT image dataset and a DSC value of 0.73± 0.03, a JAC value of 0.73± 0.02, and an Accuracy of 0.74±0.02 in the second plain CT image dataset.

TABLE II
THE COMPARISON BETWEEN TRANS-SWIN-UNET AND TRANSUNET AFTER REPLACING THE SWIN TRANSFORMER DECODER - FIRST DATASET

Model	No. of variables (M)	Computing Quantity (GFlops)	Evaluation criteria (Mean ± SD)		
			DSC	JAC	ACC
TransUNet	82.58	78.32	0.68 ± 0.03	0.69 ±0.03	0.70 ±0.05
Trans-Swin-Unet	**120.32**	**116.15**	**0.71 ±0.02**	**0.75 ±0.05**	**0.75 ±0.03**

Note: DSC is the Dice Similarity Coefficient, JAC is the Jaccard Coefficient, and ACC is the Accuracy.

TABLE III
THE COMPARISON BETWEEN TRANS-SWIN-UNET AND TRANSUNET AFTER REPLACING THE SWIN TRANSFORMER DECODER - SECOND DATASET

Model	No. of variables (M)	Computing Quantity (GFlops)	Evaluation criteria (Mean ± SD)		
			DSC	JAC	ACC
TransUNet	75.28	68.42	0.67 ±0.06	0.68 ±0.04	0.70 ±0.03
Trans-Swin-Unet	**110.92**	**103.23**	**0.72 ±0.03**	**0.76 ±0.01**	**0.76 ±0.02**

Note: DSC is the Dice Similarity Coefficient, JAC is the Jaccard Coefficient, and ACC is the Accuracy.

TABLE IV
THE COMPARISONS AFTER DEPLOYING THE MAX ATTENTION MODULE - FIRST DATASET

Model	No. of variables (M)	Computing Quantity (GFlops)	Evaluation criteria (Mean ± SD)		
			DSC	JAC	ACC
Trans-Swin-UNet_D0	84.23	90.23	0.70 ± 0.03	0.70 ±0.04	0.71 ±0.02
Trans-Swin-UNet_D1	85.74	92.36	0.70 ±0.02	0.70 ±0.03	0.72 ±0.03
Trans-Swin-UNe_D2	86.21	95.52	0.71 ±0.03	0.70 ±0.02	0.72 ±0.03
Trans-Swin-UNet_D3	**86.79**	**97.47**	**0.72 ±0.01**	**0.71 ±0.03**	**0.73 ±0.03**

Note: DSC is the Dice Similarity Coefficient, JAC is the Jaccard Coefficient, and ACC is the Accuracy, and D0, D1, D2, D3 represent the number of Max-Attention module deployed in the Trans-Swin-UNet.

TABLE V
THE COMPARISONS AFTER DEPLOYING THE MAX ATTENTION MODULE - SECOND DATASET

Model	No. of variables (M)	Computing Quantity (GFlops)	Evaluation criteria (Mean ± SD)		
			DSC	JAC	ACC
Trans-Swin-UNet_D0	73.56	80.36	0.71 ±0. 03	0.71 ±0.05	0.72 ±0.01
Trans-Swin-UNet_D1	73.98	81.45	0.71 ±0.05	0.71 ±0.02	0.73 ±0.03
Trans-Swin-UNe_D2	75.35	85.63	0.71±0.02	0.71 ±0.01	0.73 ±0.02
Trans-Swin-UNet_D3	**78.80**	**87.21**	**0.73 ±0.03**	**0.73 ±0.02**	**0.74 ±0.02**

Note: DSC is the Dice Similarity Coefficient, JAC is the Jaccard Coefficient, and ACC is the Accuracy, and D0, D1, D2, D3 represent the number of Max-Attention module deployed in the Trans-Swin-UNet.

3) *Experiment with Total Loss Function*: Tables 6 and 7 compare the new Total Loss function and the conventional loss functions for the plain CT image segmentation, including Dice loss, Boundary loss, and Focal loss. The results indicate the Total Loss function performs best among all loss functions in segmenting CT images of cerebral infarction. The built model adopting the Total Loss function achieves a DSC value of 0.58±0.01, a JAC value of 0.37±0.04, and an Accuracy of 0.71±0.03 in the first plain CT image dataset and a DSC value of 0.60± 0.02, a JAC value of 0.40± 0.04, and an Accuracy of 0.73±0.03 in the second plain CT image dataset.

TABLE VI
THE COMPARISONS AMONG FOUR LOSS FUNCTIONS - FIRST DATASET

Model	**Evaluation criteria (Mean ± SD)**		
	DSC	**JAC**	**ACC**
Dice	0.56±0.02	0.32±0.01	0.66±0.03
Boundary Loss	0.56±0.04	0.34±0.05	0.67±0.02
Focal Loss	0.57±0.03	0.36±0.01	0.66±0.03
Total Loss	**0.58±0.01**	**0.37±0.04**	**0.71±0.03**

Note: DSC is the Dice Similarity Coefficient, JAC is the Jaccard Coefficient, and ACC is the Accuracy.

TABLE VII
THE COMPARISONS AMONG FOUR LOSS FUNCTIONS– SECOND DATASET

Model	**Evaluation criteria (Mean ± SD)**		
	DSC	**JAC**	**ACC**
Dice	0.57±0.03	0.33±0.02	0.66±0.02
Boundary Loss	0.58±0.03	0.34±0.03	0.65±0.01
Focal Loss	0.58±0.01	0.35±0.01	0.65±0.02
Total Loss	**0.60±0.02**	**0.40±0.04**	**0.73±0.03**

Note: DSC is the Dice Similarity Coefficient, JAC is the Jaccard Coefficient, and ACC is the Accuracy.

4) *Experiment with CNN Layers*: A horizontal comparison was made between the results of TransUnet using ResNet101 and ResNet150 models after three and four down-sampling, respectively, to achieve the best segmentation effect: (1) ResNet50-D3 model, three down-sampling; (2) ResNet50-D4 model, four down-sampling; (3) ResNet101-D3 model, three down sampling; (4) ResNet101-D4 model, four down-sampling; (5) ResNet150-D3 model, three down-sampling; (6) ResNet150-D4 model, four down-sampling. Tables 8 and 9 show the data results of ResNet50, ResNet101, and ResNet152 after three and four down-samplings, respectively. The built model adopting ResNet50-D4 achieves a DSC value of 0.65±0.03, a JAC value of 0.37±0.03, and an Accuracy of 0.70±0.03 in the first plain CT image dataset and a DSC value of 0.65± 0.02, a JAC value of 0.37± 0.02, and an Accuracy of 0.71±0.03 in the second plain CT image dataset.

B. Results of Comparing Multiple Segmentation Models

This study compares multiple UNet-based segmentation models for plain CT images. Each model's performance is evaluated through five-fold cross-validation. Tables 10 and 11 show that the experiment results indicate that Tran-Swin-UNet performs best among all comparison models. Specifically, compared to UNet [6], Attention UNet [22], ResNet [25], and Swin UNet [15], Tran-Swin-UNet has the best segmentation performances in terms of the Dice similarity coefficient, Jaccard coefficient, and Accuracy. Specifically, the results of the proposed Trans-Swin-UNet model achieve a DSC value of 0.72±0.01, a JAC value of 0.78±0.04, and an Accuracy of 0.75±0.03 in the first (self-collected) plain CT image dataset and a DSC value of 0.73± 0.02, a JAC value of 0.79± 0.03, and an Accuracy of 0.76±0.02 in the second (public) plain CT image dataset.

TABLE VIII
THE COMPARISON AMONG THREE CNN LAYERS AND TWO DOWN SAMPLINGS - FIRST DATASET

Model	**No. of variables (M)**	**Computing Quantity (GFlops)**	**Evaluation criteria (Mean ± SD)**		
			DSC	**JAC**	**ACC**
ResNet50_D3	10.52	1.52	0.63 ±0.05	0.36 ±0.05	0.68 ±0.02
ResNet50_D4	**17.14**	**2.74**	**0.65 ±0.03**	**0.37 ±0.03**	**0.70 ±0.03**
ResNet101_D3	31.23	12.23	0.63 ±0.06	0.38 ±0.02	0.66 ±0.05
ResNet101_D4	53.47	18.16	0.65 ±0.02	0.38 ±0.03	0.69 ±0.06
ResNet150_D3	70.54	25.32	0.63 ±0.05	0.39 ±0.04	0.69 ±0.05
ResNet150_D4	121.36	32.26	0.65 ±0.04	0.38 ± 0.02	0.69 ±0.03

Note: DSC is the Dice Similarity Coefficient, JAC is the Jaccard Coefficient, and ACC is the Accuracy.

TABLE IX
THE COMPARISON AMONG THREE CNN LAYERS AND TWO DOWN SAMPLINGS - SECOND DATASET

Model	**No. of variables (M)**	**Computing Quantity (GFlops)**	**Evaluation criteria (Mean ± SD)**		
			DSC	**JAC**	**ACC**
ResNet50_D3	8.36	1.15	0.62 ±0.04	0.35±0.01	0.67±0.01
ResNet50_D4	**11.34**	**1.23**	**0.65±0.02**	**0.37±0.02**	**0.71±0.03**
ResNet101_D3	25.67	8.22	0.61±0.03	0.33±0.04	0.65±0.04
ResNet101_D4	32.47	17.54	0.64±0.01	0.32±0.05	0.66±0.07
ResNet150_D3	50.13	21.38	0.63±0.04	0.34±0.03	0.65±0.04
ResNet150_D4	80.26	26.56	0.62±0.05	0.34±0.01	0.63±0.02

Note: DSC is the Dice Similarity Coefficient, JAC is the Jaccard Coefficient, and ACC is the Accuracy.

Fig. 8 shows the segmentation results of each model based on the self-collected plain CT image. The UNet model significantly differs from the gold standard and can only segment larger lesion areas. In contrast, the surrounding small lesion areas need to be recognized. Attention Unet has improved the segmentation accuracy of small lesion areas compared to UNet due to the introduction of the attention mechanism. Swin Unet has improved the segmentation accuracy of small lesion areas compared to UNet due to the use of the Swin Transformer module. ResNet slightly improves segmentation performance compared to UNet due to the increased down-sampling times. However, the above four UNet-based models have a significant gap in segmentation performance compared to the Trans-Swin-UNet model (see Table X) since the built model integrates the main features used in each of the four UNet-based models. Such integration supports the built model in effectively and accurately detecting the small lesion areas. In summary, the segmentation performance of the Trans-Swin-UNet model is best aligned with the gold standard.

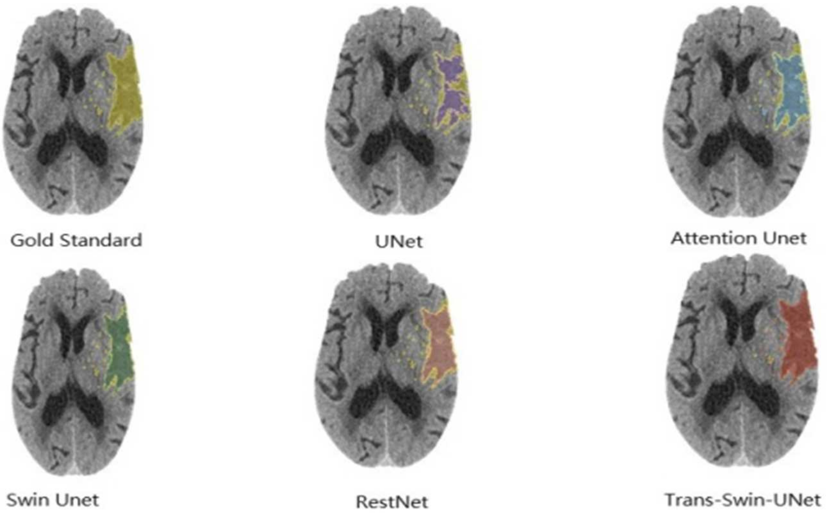

Fig. 8 The comparisons of segmentation results

TABLE X
THE COMPARISON BETWEEN TRANS-SWIN-UNET AND FOUR EXISTING MODELS FOR CT IMAGE SEGMENTATION - FIRST DATASET

Model	**No. of variables (M)**	**Computing Quantity (GFlops)**	**Evaluation criteria (Mean ± SD)**		
			DSC	**JAC**	**ACC**
Unet	20.36	21.52	0.70±0.05	0.73±0.01	0.65±0.02
Attention UNet	50.14	90.32	0.69 ±0.03	0.78±0.03	0.72 ±0.05
Swin UNet	100.23	103.21	0.68 ±0.06	0.72 ±0.04	0.73 ±0.04
ResNet	95.23	105.32	0.69 ±0.02	0.72 ±0.05	0.73 ±0.05
Trans-Swin-UNet	**120.21**	**130.12**	**0.72 ±0.01**	**0.78 ±0.04**	**0.75 ±0.03**

Note: DSC is the Dice Similarity Coefficient, JAC is the Jaccard Coefficient, and ACC is the Accuracy.

TABLE XI
THE COMPARISON BETWEEN TRANS-SWIN-UNET AND FOUR EXISTING MODELS FOR CT IMAGE SEGMENTATION – SECOND DATASET

Model	**No. of variables (M)**	**Computing Quantity (GFlops)**	**Evaluation criteria (Mean ± SD)**		
			DSC	**JAC**	**ACC**
Unet	18.25	20.33	0.68 ±0.03	0.74 ±0.03	0.68 ±0.02
Attention UNet	48.14	83.26	0.70 ±0.02	0.77 ±0.02	0.73 ±0.01
Swin UNet	95.23	100.15	0.67 ±0.07	0.73 ±0.01	0.73 ±0.05
ResNet	90.15	108.23	0.68 ±0.02	0.73 ±0.02	0.74 ±0.03
Trans-Swin-UNet	**115.29**	**125.37**	**0.73 ±0.02**	**0.79 ±0.03**	**0.76 ±0.02**

Note: DSC is the Dice Similarity Coefficient, JAC is the Jaccard Coefficient, and ACC is the Accuracy.

IV. CONCLUSION

The present study builds a UNet-based neural network model, Trans-Swin-UNet, for segmenting ischemic lesions of the plain CT image. The neural network model is built based on the four main improvements on TransUNet. The optimization experiment shows that each improvement could perform better than the original setup. The comparison experiment finds that Trans-Swin-UNet could achieve the best segmentation performance among the five UNet-based neural network models regarding the three evaluation indicators: Dice similarity coefficient, Jaccard coefficient, and Accuracy.

The present study had limited plain CT image datasets trained and tested in the experiments. Future work could apply the proposed model to segment more public CT image datasets. Moreover, research into model compression and efficient training techniques will play a vital role in the practical deployment of the hybrid CNN-Transformer UNet model of segmenting medical images in clinic settings.

ACKNOWLEDGMENT

This research was funded by a matching grant between Foshan University, China, and Multimedia University, Malaysia (Grant No. MMUE/230034).

REFERENCES

[1] V. L. Feigin et al., "Pragmatic solutions to reduce the global burden of stroke: a World Stroke Organization–Lancet Neurology Commission," Dec. 01, 2023, Elsevier Ltd. 10.1016/S1474-4422(23)00277-6.

[2] M. Zhou et al., "Mortality, morbidity, and risk factors in China and its provinces, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017," The Lancet, vol. 394, no. 10204, pp. 1145–1158, Sep. 2019, 10.1016/S0140-6736(19)30427-1.

[3] J. Vymazal, A. M. Rulseh, J. Keller, and L. Janouskova, "Comparison of CT and MR imaging in ischemic stroke," Dec. 01, 2012, Springer Verlag. 10.1007/s13244-012-0185-9.

[4] H. Abbasi, M. Orouskhani, S. Asgari, and S. S. Zadeh, "Automatic brain ischemic stroke segmentation with deep learning: A review," Neuroscience Informatics, vol. 3, no. 4, p. 100145, Dec. 2023, 10.1016/j.neuri.2023.100145.

[5] M. Soltanpour, R. Greiner, P. Boulanger, and B. Buck, "Improvement of automatic ischemic stroke lesion segmentation in CT perfusion maps using a learned deep neural network," Comput Biol Med, vol. 137, Oct. 2021, 10.1016/j.compbiomed. 2021.104849.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2015, pp. 234–241. 10.1007/978-3-319-24574-4_28.

[7] M. Malik, B. Chong, J. Fernandez, V. Shim, N. K. Kasabov, and A. Wang, "Stroke Lesion Segmentation and Deep Learning: A Comprehensive Review," Bioengineering, vol. 11, no. 1, Jan. 2024, 10.3390/bioengineering11010086.

[8] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The Importance of Skip Connections in Biomedical Image Segmentation," Aug. 2016, [Online]. Available: http://arxiv.org/abs/1608.04117

[9] A. Annotation¨ozgün, C. ¸ Içek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric

Segmentation from Sparse Annotation¨Ozgün," 2016. [Online]. Available: http://lmb.informatik.uni-freiburg.de/resources/opensource/unet.en.html

[10] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in 2016 Fourth International Conference on 3D Vision (3DV), IEEE, Oct. 2016, pp. 565–571. 10.1109/ 3DV.2016.79.

[11] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," 2018, pp. 3–11. 10.1007/978-3-030-00889-5_1.

[12] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[13] O. Petit, N. Thome, C. Rambour, and L. Soler, "U-Net Transformer: Self and Cross Attention for Medical Image Segmentation," in Lecture Notes in Computer Science, vol. 12966, 2021. 10.1007/978-3-030-87589-3_28.

[14] A. Hatamizadeh et al., "UNETR: Transformers for 3D Medical Image Segmentation," in 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Mar. 2022, pp. 1748–1758. 10.1109/WACV51458.2022.00181.

[15] H. Cao et al., "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," in Lecture Notes in Computer Science, vol. 13803, 2021, pp. 205–218. 10.1007/978-3-031-25066-8_9.

[16] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical Transformer: Gated Axial-Attention for Medical Image Segmentation," in Lecture Notes in Computer Science, vol. 12901, 2021. 0.1007/978-3-030-87193-2_4.

[17] W. Wang, C. Chen, M. Ding, J. Li, H. Yu, and S. Zha, "TransBTS: Multimodal Brain Tumor Segmentation Using Transformer," Mar. 2021, 10.1007/978-3-030-87193-2_11.

[18] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation," in Lecture Notes in Computer Science, vol. 12903, 2021, pp. 171–180. 10.1007/978-3-030-87199-4_16.

[19] Z. Xu and C. Ding, "Combining convolutional attention mechanism and residual deformable Transformer for infarct segmentation from CT scans of acute ischemic stroke patients," Front Neurol, vol. 14, 2023, 10.3389/fneur.2023.1178637.

[20] Z. Li et al., "TFCNs: A CNN-Transformer Hybrid Network for Medical Image Segmentation," in Artificial Neural Networks and Machine Learning – ICANN 2022, 2022, pp. 781–792. 10.1007/978-3-031-15937-4_65.

[21] J. Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," 2021. [Online]. Available: https://github.com/Beckschen/

[22] O. Oktay et al., "Attention U-Net: Learning Where to Look for the Pancreas," Apr. 2018, [Online]. Available: http://arxiv.org/abs/1804.03999

[23] Z. Zhang, H. Tian, Z. Xu, Y. Bian, and J. Wu, "Application of a pyramid pooling Unet model with integrated attention mechanism and Inception module in pancreatic tumor segmentation," J Appl Clin Med Phys, vol. 24, no. 12, Dec. 2023, 10.1002/acm2.14204.

[24] M. Yeung, E. Sala, C. B. Schönlieb, and L. Rundo, "Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation," Computerized Medical Imaging and Graphics, vol. 95, Jan. 2022, 10.1016/j.compmedimag.2021.102026.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Dec. 2015, pp. 770–778. 10.1109/CVPR.2016.90.

[26] K. L. Lew, C. Y. Kew, K. S. Sim, and S. C. Tan, "Adaptive Gaussian Wiener Filter for CT-Scan Images with Gaussian Noise Variance," Journal of Informatics and Web Engineering, vol. 3, no. 1, pp. 169–181, Feb. 2024, 10.33093/jiwe.2024.3.1.11.

[27] W. T. Chan, "Conditional Noise Filter for MRI Images with Revised Theory on Second-order Histograms," International Journal on Robotics, Automation and Sciences, vol. 3, pp. 25–32, Nov. 2021, 10.33093/ijoras.2021.3.5.

[28] H. Çiğ, M. T. Güllüoğlu, M. B. Er, U. Kuran, and E. C. Kuran, "Enhanced Disease Detection Using Contrast Limited Adaptive Histogram Equalization and Multi-Objective Cuckoo Search in Deep Learning," Traitement du Signal, vol. 40, no. 3, pp. 915–925, Jun. 2023, 10.18280/ts.400308.

[29] J. Muschelli, "Recommendations for Processing Head CT Data," Front Neuroinform, vol. 13, Sep. 2019, 10.3389/fninf.2019.00061.

[30] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," Dec. 01, 2022, BioMed Central Ltd. 10.1186/s13104-022-06096-y.

[31] C. C. Chai, W. H. Khoh, Y. H. Pang, and H. Y. Yap, "A Lung Cancer Detection with Pre-Trained CNN Models," Journal of Informatics and Web Engineering, vol. 3, no. 1, pp. 41–54, Feb. 2024, doi: 10.33093/jiwe.2024.3.1.3.