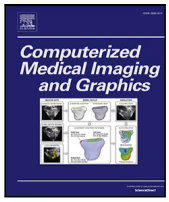
Computerized Medical Imaging and Graphics 108 (2023) 102258

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag



An amalgamation of vision transformer with convolutional neural network for automatic lung tumor segmentation

Shweta Tyagi *, Devidas T. Kushnure, Sanjay N. Talbar

Centre of Excellence in Signal and Image Processing, Department of Electronics and Telecommunication Engineering, Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, India

ARTICLE INFO

Keywords:
Lung tumor
Vision transformer
Convolutional neural network
Segmentation
Depthwise separable convolution
Atrous convolution

ABSTRACT

Lung cancer has the highest mortality rate. Its diagnosis and treatment analysis depends upon the accurate segmentation of the tumor. It becomes tedious if done manually as radiologists are overburdened with numerous medical imaging tests due to the increase in cancer patients and the COVID pandemic. Automatic segmentation techniques play an essential role in assisting medical experts. The segmentation approaches based on convolutional neural networks have provided state-of-the-art performances. However, they cannot capture long-range relations due to the region-based convolutional operator. Vision Transformers can resolve this issue by capturing global multi-contextual features. To explore this advantageous feature of the vision transformer, we propose an approach for lung tumor segmentation using an amalgamation of the vision transformer and convolutional neural network. We design the network as an encoder–decoder structure with convolution blocks deployed in the initial layers of the encoder to capture the features carrying essential information and the corresponding blocks in the final layers of the decoder. The deeper layers utilize the transformer blocks with a self-attention mechanism to capture more detailed global feature maps. We use a recently proposed unified loss function that combines cross-entropy and dice-based losses for network optimization. We trained our network on a publicly available NSCLC-Radiomics dataset and tested its generalizability on our dataset collected from a local hospital. We could achieve average dice coefficients of 0.7468 and 0.6847 and Hausdorff distances of 15.336 and 17.435 on public and local test data, respectively.

1. Introduction

Lung cancer is the deadliest type of cancer worldwide (Ferlay et al., 2021). It requires proper diagnosis for better treatment to increase the survival rate of the patients. An early-stage detection can lead to better treatment. However, it is primarily diagnosed in the later stages due to similar symptoms to common illnesses like coughing, shortness of breath, and headache. The patients go for screening when symptoms become severe and they are in their later stages of lung cancer with large and irregular-sized tumors.

Lung cancer can be diagnosed with the help of medical imaging tests, such as X-rays, Computed Tomography (CT) scans, Positron Emission Tomography (PET) scans, and so on (Cancer, 2021). Doctors generally prefer CT scans after an initial screening with X-rays; because it provides more detailed information about the size and location of the tumor. The number of lung cancer patients is rising yearly; the medical imaging tests also increase, causing a higher workload on radiologists (Mathur et al., 2020). Therefore, an automatic computer diagnosis system is required to assist medical experts. Moreover, lung tumor segmentation is the first step toward achieving this goal.

A lung tumor is defined as the uncontrolled growth of abnormal cells in the lung. A biopsy is required to check whether the tumor is cancerous or not. A small sample is taken from the tumor region and is tested under a microscope in a biopsy. There are mainly two types of lung cancer based on its microscopic view, which are small-cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), out of which NSCLC accounts for around 85% of the total cases (Siegel et al., 2021). The treatment planning depends on size, location, and cancer staging. The delineation of the tumor region can obtain the size and shape. Manual delineation is a tedious task, so the researchers develop automatic approaches.

Several researchers have presented segmentation approaches to segment the lungs and the lung tumor regions from the CT images. These studies include traditional methods like thresholding (John and Mini, 2016; Hadjileontiadis, 2005; Jamil and Butt, 2016), morphological operations (Li et al., 2007; Sahu et al., 2017), Support Vector Machines (SVMs) (Keshani et al., 2013; Naqi et al., 2018; Netto et al., 2012), and modern deep learning approaches consisting of techniques based

* Corresponding author.
E-mail address: 2019pec901@sggs.ac.in (S. Tyagi).

https://doi.org/10.1016/j.compmedimag.2023.102258
Received 9 April 2022; Received in revised form 29 May 2023; Accepted 29 May 2023
Available online 3 June 2023

on convolutional neural networks (CNNs) (Singadkar et al., 2020; Hu et al., 2020b; Dutande et al., 2021b). Furthermore, CNN-based strategies have outperformed the traditional approaches. However, they have a limitation in that they cannot capture long-range relations from the region-based convolutional operator. This limitation can be resolved by vision transformers (ViTs) (Kolesnikov et al., 2021) because they can model the long-range correlations; their performance depends on pre-training with a large amount of data. Therefore, we have implemented a network to utilize the advantages of both CNNs and ViTs.

We proposed a network based on U-Net architecture by employing a fusion of convolutional and transformer blocks. The convolutional blocks capture the initial feature maps containing the short-range inter-relations. The transformer blocks can utilize these feature maps to learn the long-range multi-contextual features, which helps to improve the segmentation accuracy. In the convolutional blocks, we implemented a combination of two convolution operations, dilated convolution to increase the receptive field and depthwise separable convolution to reduce the computation complexity of the network. The skip connections between the encoder and decoder are provided to preserve the vital information. Our proposed network performed well for lung tumor segmentation without much pre-processing.

The main contributions of our study are as follows:

- We have implemented a novel automatic lung tumor segmentation approach based on the vision transformer and convolutional neural network.
- We introduced the transformer blocks in the encoder, decoder, and bottleneck parts in the proposed network architecture. However, we employed the convolutional blocks in the initial and final layers of the network to enable high-level feature extraction.
- To reduce the computation complexity and increase the receptive field of the feature maps, we have deployed a fusion of atrous convolution and depthwise-separable convolution in the convolution blocks.
- To optimize the network, we utilized a newly proposed unified focal loss, a combination of dice and cross-entropy-based losses, which helps increase the model's generalizability by handling class-imbalance issues in the dataset.
- The proposed network is trained on a publicly available dataset NSCLC-Radiomics. The trained network is also evaluated on a local tumor dataset. Our proposed network could achieve better segmentation results.

The rest of the paper is organized as follows, related literature studies are discussed in Section 2; the proposed methodology is explained in Section 3; Section 4 presents the experiments and results with ablation study, a discussion is made in Section 5, and in Section 6, the conclusion of the study is given.

2. Related work

2.1. Convolutional neural networks

Lung tumor segmentation is vital in lung cancer diagnosis and treatment analysis. Several automatic lung tumor segmentation techniques have been proposed to assist radiologists by providing another expert opinion. Some of these techniques are based on traditional machine learning, but researchers have utilized Convolutional Neural Networks (CNNs) in contemporary methods. U-Net (Ronneberger et al., 2015) is a popular type of CNN explicitly designed for biomedical image segmentation. After that, many variants are developed for various computer vision applications, including the medical imaging field (Baid et al., 2018; Khanna et al., 2020; Baid et al., 2020; Dutande et al., 2021a). Some other networks such as the efficient version of CNNs (Hooda et al., 2018; Narayanan and Hardie, 2019), recurrent neural networks (Moitra and Mandal, 2020; Abid et al., 2021), Generative Adversarial Networks (GAN) (Pawar and Talbar, 2021; Jain et al., 2021), and Graph Neural Networks (GNN) (Dai et al., 2015) are also being explored for medical image segmentation and classification tasks.

Several research works are available for small-sized lung tumors with sizes ranging between 3 mm to 30 mm, also known as lung nodules. W. Huang et al. (Huang and Hu, 2019) propose an improved version of U-Net, which they named noisy U-Net (NU-Net). In this proposed network, they have added a spectral noise in the hidden layers to increase the sensitivity of the network to detect and segment the lung nodules more accurately. They have evaluated this approach on two datasets, LUNA16, and the Tianchi Competition dataset, and achieved a sensitivity of 97.1% to tiny nodules with 3–5 mm diameters, which is greater than the corresponding U-Net value of 90.5%. Various researchers have utilized the different versions of U-Net (Zheng et al., 2019; Singadkar et al., 2020; Wu et al., 2020)

Shi et al. (2020) introduce LIF-Nets (Leaky Integrate and Fire Networks) to detect and classify lung nodules. These networks have rich inter-frame sensing capability, which helps capture essential features from the consecutive slices with less computational cost. They achieved a sensitivity of 94.6% at 8 FPs per scan for lung nodule detection using the LUNA16 dataset.

Various region-based convolutional neural networks are designed specifically for object detection like R-CNN (Girshick et al., 2014), Fast R-CNN, (Girshick, 2015) and Faster R-CNN (Ren et al., 2015). The R-CNN is based on the principle that the information regarding the regions of interest is extracted using a selective search algorithm. An SVM classifier is used to classify the object present in that region. These networks have broad applications in medical image analysis. Ding et al. (2017) implement a deconvolutional configuration to Faster R-CNN for pulmonary nodule detection using lung CT images. Moreover, for false-positive reduction, they utilize a 3D CNN. The proposed network comprises a region proposal network (RPN) that detects nodule regions and an ROI classifier to classify the nodules. These two networks utilize the same feature extraction layers to save the computation cost of CNN training. A CNN consisting of an ROI Pooling layer and a fully-connected network is designed to classify the detected ROI as nodule or non-nodule. They experimented with their proposed architecture on the LUNA16 dataset and achieved an average FROC score of 0.893. This proposed approach achieved sensitivities of 92.2% at 1 FP and 94.4% at 4 FPs per scan.

Another network called Mask R-CNN (He et al., 2017) was explicitly proposed for image segmentation. It is designed by modifying the Faster R-CNN architecture. There are two outputs in a Faster R-CNN, one is the object label, and another is the bounding box for the object. However, Mask R-CNN adds a third output, a segmentation map of the object. Cai et al. (2020) presents an approach based on Mask R-CNN for detecting and segmenting pulmonary nodules. The proposed method is composed of three different modules, which are PrM (Pre-Processing Module), DSM (Detection and Segmentation Module), and 3DRM (3-Dimensional Reconstruction Module). In the pre-processing module, the 3D lung CT images are processed to get the 2D images, and simple image processing like mask generation and normalization are performed. In the DSM module, Mask-RCNN performs the task of detection and segmentation. For the backbone of Mask R-CNN, resnet50 is used, and a Feature Pyramid Network (FPN) is utilized to explore multi-scale feature maps. Furthermore, Region Proposal Network (RPN) proposes candidate bounding boxes. Then in the 3DRM module, the mask matrices are multiplied by the raw medical image matrices to get the sequences of predicted lung nodules and generate the 3D models of lung nodules. This approach is evaluated on the LUNA16 dataset and one other dataset from the Ali TianChi challenge and has achieved sensitivities of 88.1% at 1 FP and 88.7% at 4 FPs per scan.

Jain et al. (2021) propose Salp Shuffled Shepherd Optimization Algorithm (SSSOA) based Generative Adversarial Network for lung nodule segmentation. They have combined SSA (Salp Swarm Algorithm) (Mirjalili et al., 2017) and SSOA (Shuffled Shepherd Optimization Algorithm) (Kaveh and Zaerreza, 2020) to develop SSSOA. In their

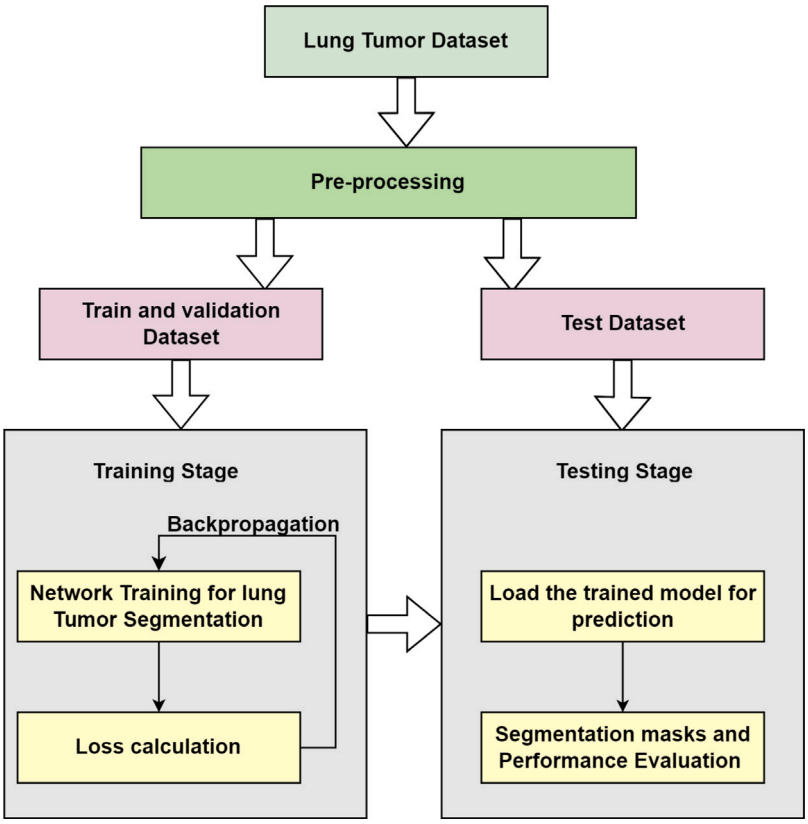


Fig. 1. Pipeline of proposed approach.

study, first, the data is preprocessed using Gaussian filtering, and then lung lobes are segmented. They evaluated their proposed approach on the LIDC-IDRI dataset and achieved 0.9387 accuracies, 0.7986 dice coefficient, and 0.8026 Jaccard similarity.

Many more such studies are present for lung nodule analysis, but the literature still needs to be updated for later-stage large tumors. The researchers in Uzelaltinbulat and Ugur (2017) have proposed a lung tumor segmentation algorithm based on simple image processing techniques. Anthimopoulos et al. (2018) implemented a convolutional neural network-based approach based on dilated convolutions for lung tumor segmentation and achieved a dice similarity coefficient 0.6267. They named this network as 2D-LungNet. A 3D version of LungNet has been implemented by Hossain et al. (2019) for lung tumor segmentation and achieved a dice index of 0.6577. In Kamal et al. (2020), the authors designed a lung tumor segmentation approach based on 3D recurrent DenseUnet by employing Convolutional Long Short-Term Memory LSTM modules in the network; They achieved a dice similarity coefficient of 0.7228. Although these approaches can perform well, there is still some scope to further improve lung tumor segmentation performance by implementing an efficient methodology.

2.2. Attention based networks

Various attention-based modules are implemented to improve the performance of the CNNs further. The self-attention technique has been proposed in Wang et al. (2018), which is proved to be very significant in enhancing the accuracy of CNNs. The attention mechanism was deployed in U-Net architecture by Oktay et al. (2018) to perform medical image segmentation. In Schlemper et al. (2019), the authors have utilized skip connections with attention mechanisms in encoder–decoder architecture for medical image segmentation. For channel-wise feature-recalibration, the squeeze and Excitation (SE) networks are proposed by Hu et al. (2018). This SE module can improve performance in any convolutional neural network architecture. The squeeze and excitation modules for spatial feature recalibration and a hybrid version are implemented by Roy et al. (2018). These modules can also enhance the accuracy of a CNN architecture. A multi-scale attention network (Fan et al., 2020) was proposed using a hybrid deep attention-aware network to segment the liver and tumor regions. Hu et al. (2020a) present a lung tumor segmentation approach based on a hybrid attention mechanism using deformable convolutions.

2.3. Transformer based networks

Vision Transformers have recently gained popularity in the computer vision field (Kolesnikov et al., 2021). Initially, Transformers were designed for Natural Language Processing (NLP) applications. However, researchers started experimenting with vision transformers either in the encoder structures of the segmentation network to make a stronger encoder (Chen et al., 2021) or in both encoder and decoder (Gao et al., 2021). Other versions of vision transformer-based networks are also available in medical image analysis (Valanarasu et al., 2021; Lin et al., 2021; Cao et al., 2021).

The vision transformer-based networks for medical image analysis are primarily based on pre-trained networks. As far as we know, a transformer-based network has yet to be explored for lung tumor segmentation tasks. We have implemented transformer modules in the encoder–decoder structure of our proposed network for lung tumor segmentation. Due to dense layers in vision transformer blocks, the number of network parameters increases. Therefore, we utilized depth-wise separable convolutions in the convolutional blocks to reduce the computation complexity of the network to some extent. The dilation rate of 2 is used for this convolution, which does not affect the computations but increases the receptive field of the feature maps, which helps achieve good feature maps in the initial layers passed to the transformer blocks for further processing to get the final segmentation maps. The convolutional blocks with up-sampling layers are used in the decoder part.

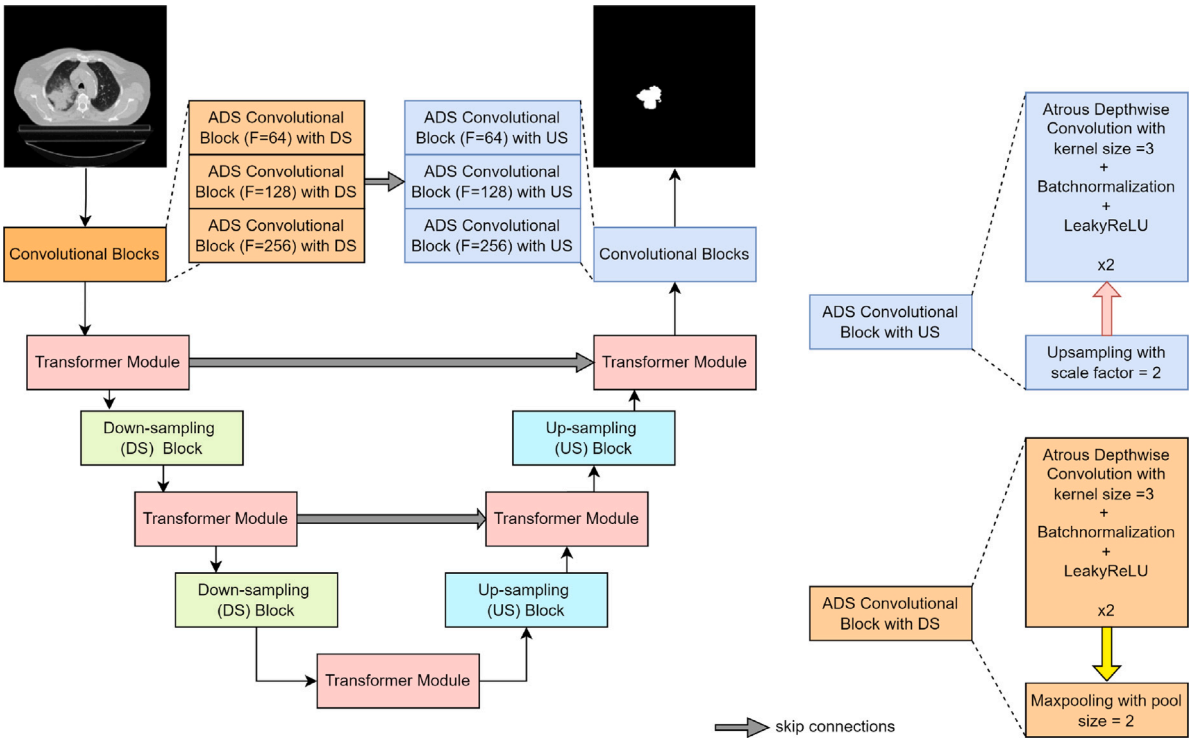


Fig. 2. Architecture of proposed network.

3. Proposed methodology

We have proposed an automatic lung tumor segmentation approach based on a vision transformer and a combination of atrous (Chen et al., 2017) and depthwise separable convolutions (Chollet, 2017). The overview of the processing pipeline is shown in Fig. 1. First, the two lung tumor datasets are acquired, NSCLC-Radiomics data from the online database and another from the local hospital. This data is pre-processed and cleaned to get better segmentation results. All images were original 512 × 512 in size, which was resized to 256 × 256 to reduce the computational complexity. Then the data is split into training, validation, and test data. Training and validation data are used to train and validate the network, and the loss function is used for back-propagation to improve the training accuracy. Once the network is trained, the test dataset is used to test the generalizability of the network.

3.1. Architecture

The proposed network is designed based on the encoder–decoder structure. In the initial and final stages of the network, we employed convolutional blocks to preserve the predominant local features during the initial phase and get the final segmentation maps more precisely. We implemented the transformer module in the deeper layers, along with the down-sampling block in the encoder and an up-sampling block in the decoder part of the network. The transformer module is implemented in the bottleneck part to focus on the deep feature enhancement. The architecture of the proposed network is presented in Fig. 2. The skip connections between the encoder and decoder blocks are provided to preserve the feature maps.

We have utilized a hybrid convolution in the convolutional blocks, a fusion of atrous and depthwise separable convolutions. Both these types of convolutions are detailed in Fig. 3. The receptive field is a crucial factor that needs to be considered for extracting prominent dense features. It is defined as the region's size in the input feature map, used to get the section of the output feature map. The standard convolution has a receptive field equal to its kernel size. So, to get a higher receptive field using regular convolutions, a high kernel size is required, which leads to more computations, and a more powerful GPU will be required to train such a network. The solution to this issue is provided in Chen et al. (2017) by introducing atrous convolution, also known as dilated convolution. It helps to increase the receptive field with the same kernel size by inflating the kernel with skip points and no extra computations. The length of skip points is defined by the dilation rate, which is equal to one for standard convolution, the visualization of which is shown in Fig. 3(a). Dilated convolution can be performed by increasing the value of the dilation rate; the example of a dilation rate equal to two can be visualized in Fig. 3(b). The dilated convolution can extract more fine-grained features by looking at the more significant portion of the image without affecting its spatial resolution. It is also computationally efficient, as with the same number of parameters used by standard convolution, atrous convolution can perform better by capturing dense features.

Another type of convolution we used is depthwise separable convolution to reduce the number of computations in the network. It is performed by dividing the standard convolution into two parts. First depthwise convolution is completed, followed by the pointwise convolution. A depthwise separable convolution can perform the regular convolution operation with kernel shape $(F \times F \times C)$ with less number of computations by using a depthwise convolution with kernel size $(F \times F \times 1)$ followed by pointwise convolution having $(1 \times 1 \times C)$ kernel size, the visualization of the same is given in Fig. 3(c).

This can be understood with the help of computation costs of all types of convolutions. So let us consider the image with size $(H \times W)$ and M number of channels, which is convolved with N number of convolution filters with size $(F \times F)$. For a standard convolution, the computation cost (C_s) can be defined as given in Eq. (1)

$$C_s = H \times W \times F \times F \times M \times N \tag{1}$$

For dilated convolution, there will be no change in computation complexity. However, for depthwise separable convolution, computation cost will be reduced. Depthwise separable convolution is a

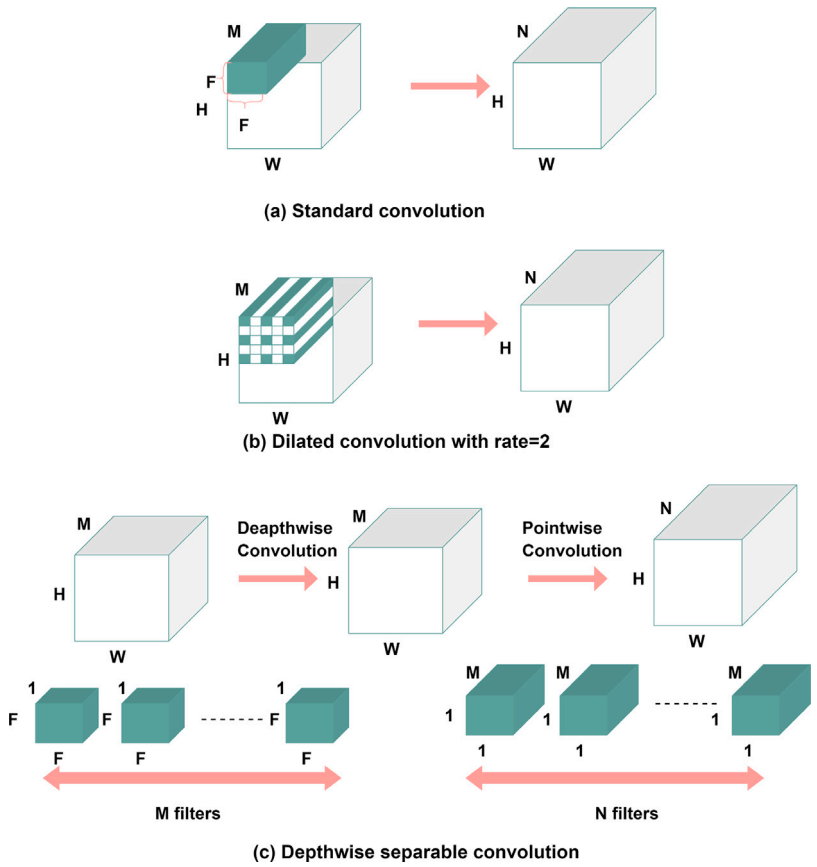


Fig. 3. Illustration of different convolutions, (a) standard convolution with kernel shape $(F \times F)$ and dilation rate = 1, (b) atrous convolution with kernel shape (3×3) and dilation rate = 2, and (c) depthwise separable convolution.

combination of depthwise and pointwise convolutions. The cost of depthwise (C_d) and pointwise (C_p) convolution can be calculated as presented in Eq. (2) and Eq. (3), respectively.

$$C_d = H \times W \times F \times F \times M \tag{2}$$

$$C_p = H \times W \times M \times N \tag{3}$$

The combined computation cost for depthwise separable convolution (C_{dw}) can be measured by adding the individual costs C_d and C_p, as given in Eq. (4)

$$C_{dw} = H \times W \times F \times F \times M + H \times W \times M \times N \tag{4}$$

We can compare the computation costs of standard and depthwise separable convolution by dividing Eq. (4) and Eq. (1), as given in Eq. (5).

$$\frac{C_{dw}}{C_s} = \frac{H \times W \times F \times F \times M + H \times W \times M \times N}{H \times W \times F \times F \times M \times N} = \frac{1}{N} + \frac{1}{F^2} \tag{5}$$

From Eq. (5), it can be observed that for a filter size of 3, the computation cost will decrease by 8–9 times, with less impact on performance.

We have employed the double convolutional layers with atrous depthwise separable convolution with a dilation rate equal to two in the atrous depthwise separable (ADS) convolution blocks used in the upper layers of the network. With the help of atrous convolution with dilation rate, the network can capture more unique features and structure maps, which are further utilized by the transformer modules embedded in the deep layers to capture the short-range relationships in the feature maps. Each ADS convolutional block uses a batch normalization layer and a LeakyReLU activation. Batch normalization helps regularize the network and reduce the internal covariant shift. The LeakyReLU function has the advantage that it can prevent the vanishing gradient issue. It is an improved version of the ReLU activation function in which there is non-zero output when the input is negative as opposed to ReLU, where the outputs are zero for negative inputs, which causes dead neuron issues in the network and vanishing gradient problems. The LeakyReLU activation helps to solve the dead ReLU issue and is helpful for better training.

3.2. Transformer module

Transformers were initially designed for natural language processing tasks, in which multi-head self-attention was employed for extracting high-level features. The schematic view of the transformer module used in the proposed network is presented in Fig. 4. First, the patches are extracted from the feature maps extracted by the previous blocks; these patches are passed through the linear projection layer to get the flattened sequence with the embedded positional encoding of the patches. The positional encoding is done to preserve the arrangement of the patches in the output feature map. Otherwise, crucial locational information can be lost. The output of positional encoding in the original vision transformer (Kolesnikov et al., 2021) is given in Eq. (6).

$$z_0 = [x_{class}; {x_p}^1 E; {x_p}^2 E; ..; {x_p}^N E] + E_{pos} \tag{6}$$

where $({x_p}^1, {x_p}^2, \ldots, {x_p}^N)$ are the patches extracted from the input feature map ans N denotes the number of patches created. $E \in R^{P^2 C}$ is the projection of patch embedding, where R is the spatial resolution of the image with height H, width W and number of channels C, P^2 is the patch size $(P \times P)$, and D is the dimension of embedding space. $(E_{pos} \in R^{N \times D})$ is the positional embedding. x_{class} is the class label.

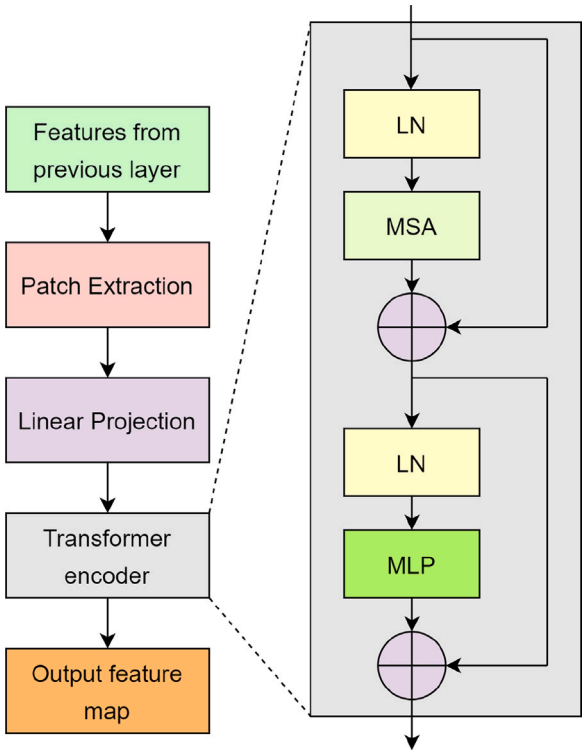


Fig. 4. Schematic view of transformer module.

We did not add any class embeddings as our task is segmentation, and we want the output features map in the form of pixel-level classification. So we modified Eq. (6) and the modified version is defined by Eq. (7)

$$z_0 = [x_p{}^1 E; x_p{}^2 E; ..; x_p{}^N E] + E_{pos} \tag{7}$$

After this, transformer encoder (Kolesnikov et al., 2021) is applied to get the output feature maps. The transformer encoder uses multi-head self-attention (MSA) and Multi-layer perceptron (MLP) layers. Layer normalization (LN) is applied to make the gradient smoother by normalizing layer distributions. In the MSA layer, the input feature map is split into multiple heads, which helps to learn the distinct levels of self-attention by each head. The single self-attention for a set of queries, keys, and values (Q, K, V) can be defined as given in Eq. (8)

$$SA(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{8}$$

where d is a scaling factor, softmax represents the softmax activation function. The final output of the MSA layer is obtained by concatenating the results of all SA blocks and is defined in Eq. (9)

$$MSA(Q, K, V) = Concat(SA_1, SA_2, \dots, SA_n)W \tag{9}$$

where W is a learnable weight matrix, this output of the MSA layer is applied to the MLP layer. The layer normalization is applied before both MSA and MLP layers, and residual connections are used after both.

MLP layer consists of two linear layers with Gaussian Error Linear Units (GELU) (Hendrycks and Gimpel, 2016). The operations in MLP layers can be defined as given in Eqs. (10) and (11) by considering z_{l-1} as the previous layer feature maps.

$$z_l{}^{MSA} = MSA(LN(z_{l-1})) + z_{l-1} \tag{10}$$

$$z_l{}^{MLP} = MLP(LN(z_l{}^{MSA})) + z_l{}^{MSA} \tag{11}$$

Table 1
Implementation details.

Configuration	Value/operations used
Data augmentation	Random rotation, Random flipping, Random zooming, Random intensity shifting
Optimizer	Adam
Initial learning rate	0.01
Loss function	Unified focal loss
Evaluation metrics	Dice Similarity Coefficient, Hausdorff Distance

The GELU activation is based on the Gaussian distribution function and is defined in Eq. (12).

$$GELU(z) = zP(Z \leq z) = z.\frac{1}{2}\left[1 + erf(z/\sqrt{2})\right] \tag{12}$$

where $P(Z \leq z) = \frac{1}{2}\left[1 + erf(z/\sqrt{2})\right]$ is the Gaussian distribution function which is differentiable for all values of z, and weights the input according to there values. GELU activation is a smoother version of ReLU and helps mix feature maps in the transformer module better.

4. Experiments and results

4.1. Implementation details

The network is trained on a 16 GB NVIDIA P100 GPU with CUDA 10.0, cuDNN 7.4, and 128 GB RAM. PyTorch 1.8.1 designs the network and implements network training and testing. The size of input images is set to 256 × 256. The training data is augmented using random rotation, intensity shifting, flipping, and zooming. For optimization, the Adam optimizer is used with an initial learning rate of 0.001, and a step decay is used for the patience of 5 epochs with a step size of 0.1. Table 1 depicts the details regarding the experimental set-up.

4.2. Dataset description and pre-processing

For this study, two datasets are utilized. The first is a publicly available dataset, NSCLC-Radiomics Dataset (Aerts et al., 2019), comprised of 422 CT scans of patients with Non-Small Cell Lung Cancer (NSCLC). The data is in RTSTRUCT (DICOM Radiotherapy Structure Sets) files, and annotation masks are provided for the gross tumor volume (GTV) in DICOM Segmentation files containing a manual delineation of the 3D volume. This dataset includes 52073 2D images, of which only around 14% of the total images contain the lung tumor. We have pre-processed the data and extracted those 2D images containing the lung tumor. The description of the data is given in Table 2. Three hundred twenty scans consisting of 5568 2D images containing lung tumors are used for training the network, and 62 CT scans are used for validation, in which there are 900 images. A test set containing 50 scans is kept aside to test the trained model.

Another CT-image dataset used to test the generalizability of the proposed network has been collected from a local hospital, and two expert radiologists verified the tumor annotations. There are 50 CT scans in this dataset, and they are in DICOM format. The segmentation masks are in NIFTI format. All these scans are taken before the treatment to diagnose the lung cancer patient. This data is used only for testing purposes.

The training dataset needs to be pre-processed before training the deep learning model so that the trained model can generalize better on a different dataset as well. As in medical imaging, the data varies depending on various factors. For example, in our case, CT images from both datasets vary based on multiple factors like CT image acquiring machine specifications, slice thickness, and geographical factors (as patients are in different locations). To pre-process the data, we first

Table 2
Data description.

Dataset split	Dataset used	Number of scans	Number of images
Training	NSCLC-Radiomics	320	5568
Validation	NSCLC-Radiomics	62	900
Test	NSCLC-Radiomics	50	900
	Local dataset	50	1000

clipped the data in the interesting intensity range for the lung tumor segmentation task, (−1000 to 1000) HU (Hounsfield Unit). There are different-sized tumors in the data, and the images corresponding to small-sized tumors are less than large-sized ones. To deal with this class imbalance issue, the data is augmented using data augmentation techniques such as random rotation, random flipping, random zooming, and random intensity shift.

4.3. Loss function and evaluation metrics

During the training of a neural network, gradient descent optimization is used, which minimizes the error gap between predicted and actual labels. This error is also known as the loss function. During back-propagation, the optimization algorithm minimizes this loss function and updates the weight parameters accordingly to improve the overall accuracy. The most popular loss functions for the binary segmentation task are binary cross-entropy (BCE), dice loss, focal loss, or combination. BCE and focal loss are distribution-based, whereas dice loss is region-based. The dice loss has been observed to work better than BCE for class-imbalanced data (Jadon, 2020). However, the dice loss has the limitation of the unstable gradient. Yeung et al. (2022) introduced a unified focal loss for better generalization in imbalanced data to deal with this limitation. This loss combines modified versions of focal loss and focal Tversky loss. It deals with the issues of larger hyper-parameter search space and the convergence problem associated with the traditional focal loss functions. Dice loss (L_{DSC}) is defined based on the dice similarity coefficient (DSC) and is given by Eq. (13).

$$L_{DSC} = 1 - DSC = 1 - \frac{2TP}{2TP + FP + FN} \tag{13}$$

where TP, FP, and FN denote the true positives, false positives, and false negatives, respectively.

Focal loss (L_F) is derived from the binary cross entropy (BCE) loss (L_{BCE}). The BCE loss is defined in Eq. (14), and its simplified version is given in Eq. (15).

$$L_{BCE}(y, \hat{y}) = -(y \log \hat{y}) + (1 - y) \log 1 - \hat{y} \tag{14}$$

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1 - p) & \text{if } y = 0 \end{cases} \tag{15}$$

where y refers to the ground truth segmentation mask and $\hat{y}$ refers to the predicted mask. BCE loss works on the pixel level and tries to minimize the pixel-wise error. But if there is a class imbalance, it can over-represent the larger class resulting in poor segmentation results. Focal loss tries to solve this issue by modifying the BCE loss. The probability (p_g) of ground truth masks can be defined, as given in Eq. (16), and BCE loss can be rewritten as provided in Eq. (17).

$$p_g = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{if } y = 0 \end{cases} \tag{16}$$

$$L_{BCE}(p, y) = CE(p_g) = -\log p_g \tag{17}$$

The focal loss modifies this BCE loss by adding a modulating factor as defined in Eq. (18).

$$L_F(p_g) = \alpha(1 - p_g)^{\gamma}.L_{BCE}(p, y) \tag{18}$$

where α and γ are the hyper-parameters of focal loss, which controls the class weights and weight decay for pixel-wise classification; for $\gamma = 0$, focal loss works as BCE loss. Due to controlled class weights, it can perform well for class imbalance data.

Unified focal loss is a further improvement over the focal loss and is calculated by adding the modified focal loss and modified focal Tversky (FT) loss (L_{FT}). Tversky loss (L_T) is defined for the Tversky index (TI). TI is similar to DSC except for the weights α and β assigned to FPs and FNs, respectively. TL is defined in Eq. (19).

$$L_T = 1 - TI = 1 - \frac{TP}{TP + \alpha FP + \beta FN} \tag{19}$$

FT loss can be expressed in terms of the Tversky index as given in Eq. (20).

$$L_{FT} = \sum_{c=1}^{C} (1 - TI)^{\frac{1}{\gamma}} \tag{20}$$

If we combine focal and FT loss, there are so many hyper-parameters to fine-tune. Therefore in the modified focal loss (L_{mF}) and modified FT loss (L_{mFT}), the parameters α and β are replaced by a common parameter δ, and also γ is reformulated as represented in Eq. (21) and Eq. (22).

$$L_{mF} = \delta(1 - p_g)^{1-\gamma}.L_{BCE}(p, y) \tag{21}$$

$$L_{mFT} = \sum_{c=1}^{C} (1 - mTI)^{\gamma} \tag{22}$$

where mTI is the modified Tversky index in which α and β are replaced by δ. The unified focal loss L_{UF} is obtained by adding the two losses given in Eq. (21) and Eq. (22) by adding a control parameter λ, which is used to determine the relative weights of two losses. L_{UF} is defined in Eq. (23).

$$L_{UF} = \lambda L_{mF} + (1 - \lambda) L_{mFT} \tag{23}$$

In the unified focal loss function, parameters are used, γ controls the rare class enhancement and background class suppression, δ controls the weights of positive and negative samples, and λ controls the relative weights of two losses. It can outperform with better results.

We utilized two evaluation metrics to evaluate the algorithm: Dice Coefficient Similarity (DSC) Index and Hausdorff Distance (HD). DSC measures the relative overlap between ground truth masks and predicted masks and is expressed as given in Eq. (24).

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{24}$$

Hausdorff Distance is defined as the maximum distance of a set to the closest point in another set. In our case, the two sets are ground truth labels and predicted labels. For two sets X and Y, it is given by Eq. (25).

$$D_H(X, Y) = max(D_{XY}, D_{YX}) \tag{25}$$

4.4. Segmentation results

We tested the trained network on the test dataset images and obtained the segmentation results. The output segmentation results are shown in Fig. 5. The original images are shown in the first column; ground truth masks are presented in the second column, and the predicted masks are depicted in the third column. We have included different CT images with lung tumors varying in location, size, and density in the results. Our proposed network could capture various-size tumors with significant segmentation performance. It is visible through the segmentation results that the predicted masks are close to the ground truth annotations.

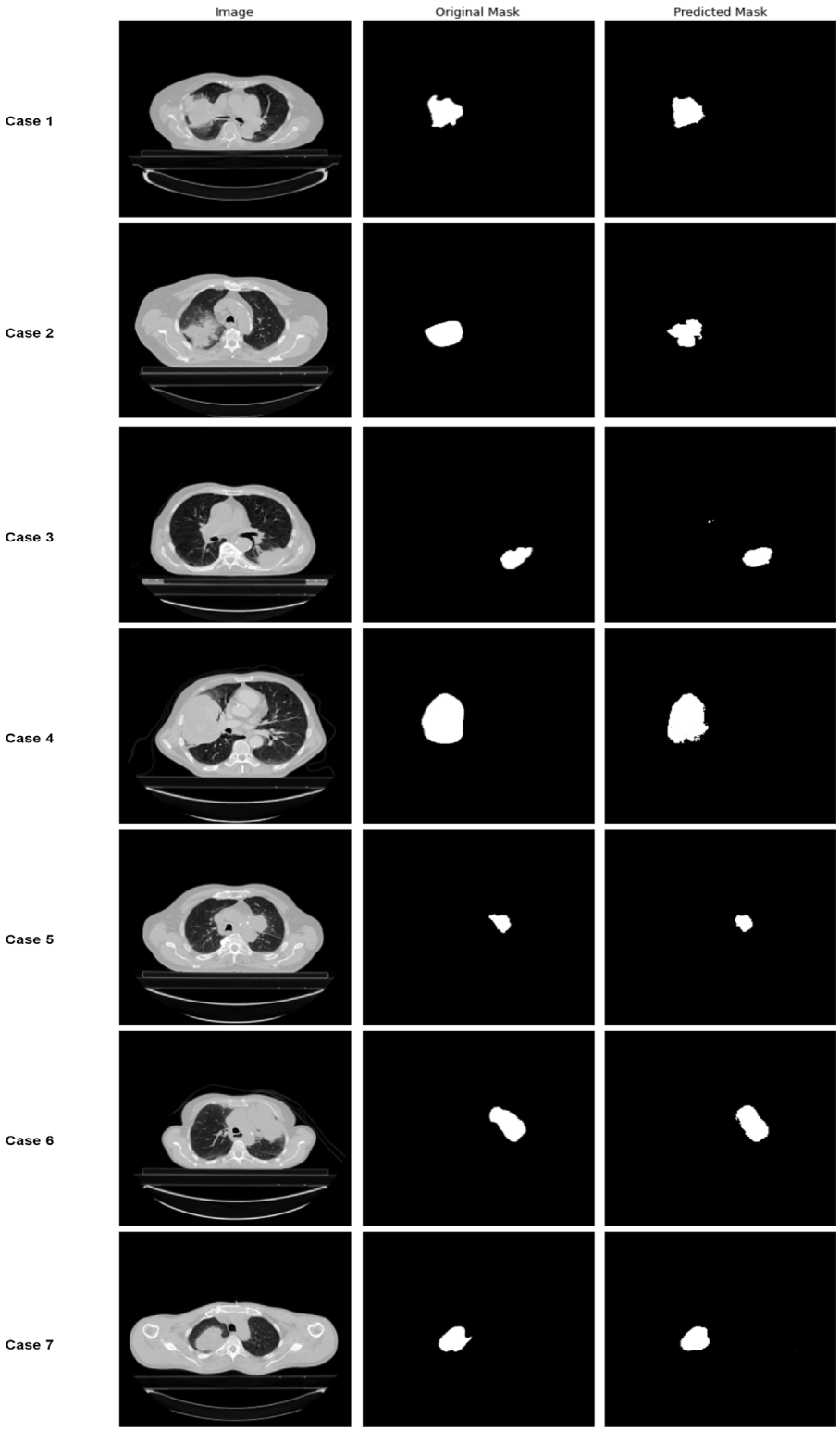


Fig. 5. Segmentation results of eight test images, the first column shows the input test images, ground truth segmentation masks are provided in the second column, and the third column depicts prediction masks.

4.5. Ablation study

We have performed an ablation study based on different parameters. First, we used different network configurations depicted in Table 3. First, we implemented UNet with standard convolution with a dilation rate of one ($r = 1$) and achieved a dice similarity coefficient of 0.6365 and a Hausdorff distance of 18.452 for NSCLC-Radiomics test data and a dice coefficient of 0.5842 and a Hausdorff distance of 23.326 is achieved for local test data. In the second network configuration, the dilation rate of two ($r = 2$) is used in the convolutional blocks in the UNet network, and the dice coefficients are increased to 0.6414 and 0.5996. Hausdorff distances are decreased to 18.132 and 22.927 for NSCLC-Radiomics data and the local dataset, respectively. We implemented a recurrent, residual UNet network (R2U-Net) (Alom et al., 2019) in our

Table 3
Evaluation metrics using different network combinations.

Networks	Number of parameters (million)	p-value	NSCLC test data		Local test data	
			DSC	HD	DSC	HD
UNet ($r = 1$)	31	0.047	0.6365	18.452	0.5842	23.326
UNet ($r = 2$)	31	0.045	0.6414	18.132	0.5996	22.927
R2U-Net	1	0.036	0.6586	16.838	0.6024	22.318
Attention UNet	34	0.032	0.6878	16.224	0.6316	21.874
Proposed network (with standard convolutional blocks)	48.6	0.028	0.7394	15.873	0.6772	18.721
Proposed network (with ADS convolutional blocks)	43	0.026	0.7468	15.336	0.6847	17.435

third network configuration and achieved improved dice coefficients and Hausdorff distances. The performance is further enhanced when we experimented with attention UNet (Oktay et al., 2018). Using attention UNet, we achieved dice coefficients of 0.6878 and 0.6316 for NSCLC and local data, respectively, and Hausdorff distances are also improved for both datasets. Then we experimented with our proposed network, first, using standard convolutional blocks in the initial and final layers of the network and second, ADS convolutional blocks. With traditional convolutional blocks, the proposed network could achieve dice coefficients of 0.7394 and 0.6772 and Hausdorff distances of 15.873 and 18.721 for NSCLC and local data, respectively. Using our final configuration of the proposed network, we achieved dice coefficients of 0.7468 and 0.6847 and Hausdorff distances of 15.336 and 17.435 for NSCLC-Radimics and local test data, respectively. We also computed the p-value for each model using the method discussed in Tanizaki et al. (2020). The p-value is a quantitative parameter that defines how correctly the segmentation is done. A p-value of less than 0.05 indicates the suitability of the segmentation algorithm, and the segmentation correctness is inversely proportional to the p-value. Our proposed network has a p-value of 0.026 which is better than other experimented models. It can be observed from Table 3 that deploying ADS convolutional blocks increased the proposed network's performance and reduced the number of parameters.

Second, we experimented with the proposed network with different loss functions, and these results are presented in Table 4. With binary cross-entropy (BCE) loss, the network could achieve dice similarity coefficients of 0.7139 and 0.6423 and Hausdorff distances of 19.367 and 20.138 for NSCLC data and local data, respectively. BCE loss deals with pixel-wise probabilities and tries to reduce pixel-wise errors, which may lead to the over-representation of more significant class or larger objects in the case of class-imbalance data. Therefore, measuring the adequate overlap between actual and predicted segmentation maps is required, done by dice similarity coefficient with dice loss. When the proposed network has experimented with dice loss, it could perform better with dice coefficients of 0.7212 and 0.6531 on public and local test sets, respectively. A focal loss is combined with dice loss, and the network performed slightly better than using dice loss alone. We set the hyper-parameter values for focal loss, which are optimal in Lin et al. (2017), $\alpha = 0.25$, and $\gamma = 2$. Finally, we experimented with unified focal loss. For our experiment with unified loss function, we set the hyper-parameters values, which are optimum values according to the original study (Yeung et al., 2022). We set $\delta = 0.6$, $\lambda = 0.5$ and $\gamma = 0.5$. We could achieve dice similarity coefficients of 0.7468 and 0.6847 and Hausdorff distances of 15.336 and 17.435 on NSCLC and local datasets, respectively.

Table 4
Evaluation metrics using different loss functions.

Networks	NSCLC test data		Local test data	
	DSC	HD	DSC	HD
Network with L_{BCE}	0.7139	19.367	0.6423	20.138
Network with L_{DSC}	0.7212	17.247	0.6531	19.764
Network with L_F and L_{DSC}	0.7325	16.783	0.6626	18.365
Network with L_{UF}	0.7468	15.336	0.6847	17.435

4.6. Comparative study

We have presented a comparative analysis using different studies conducted for lung tumor segmentation, given in Table 5. Some of these studies have experimented with only one dataset, and one evaluation metric is used to evaluate them (Anthimopoulos et al., 2018; Hossain et al., 2019; Kamal et al., 2020). They achieved good results but with more computational complexity. Data robustness is also a suitable parameter for learning the quality of the model. It is defined as the quality of data collected. Most researchers clean the training data before model training. However, if the data is from two different sources and the model can still perform well, then the robustness is high. One study by Dutande et al. (2021a) proposed a deep residual separable convolutional neural network (DRS-CNN) that utilized two different datasets for training and testing. They were able to achieve better segmentation results with less computation complexity. Our proposed network is trained on a public dataset and evaluated on a public and a local dataset, so the data robustness of our approach is high. It is observed that it can generalize well on a new dataset entirely different from the training dataset. It can be observed from Table 5 that the network's complexity increased due to transformer blocks. However, we could achieve better results than the previous studies.

5. Discussion

Deep learning techniques have accomplished significant evaluation results for medical image segmentation. Convolutional neural networks have outperformed traditional approaches. Various researchers are still exploring convolutional neural networks for computer vision applications. However, the convolutional networks cannot extract the long-range relational feature maps due to regional operations performed by convolution operators. Recently transformers embedded with multi-head self-attention (MHA) have achieved satisfactory natural language processing results. They can solve the issue of CNNs and capture the long-range relationships among feature maps. Vision transformers are developed for image analysis.

Lung cancer needs a proper diagnosis for better treatment and survival of the patient. For this purpose, automatic lung tumor segmentation plays an important role. Various convolutional neural network-based approaches have provided significant results for lung tumor segmentation. However, CNN cannot capture the long-range relations among the feature maps, affecting their segmentation performance. Although convolutional neural networks can learn the feature maps more accurately when pre-trained on a vast image dataset, data are scarce in the medical imaging field. Moreover, the available data differs by image-acquiring machines, image resolution, and slice thickness. Another issue is highly imbalanced medical image data due to a big

Table 5
Comparison of lung tumor segmentation performance using different networks.

Networks	Number of parameters (million)	Data robustness	Dataset used	DSC	HD
2D LungNet (Anthimopoulos et al., 2018)	1.3	LOW	NSCLC-Radimics (Aerts et al., 2019)	0.6267	–
3D LungNet (Hossain et al., 2019)	4	LOW	NSCLC-Radimics (Aerts et al., 2019)	0.6577	–
3D recurrent denseunet (Kamal et al., 2020)	19	LOW	NSCLC-Radimics (Aerts et al., 2019)	0.7228	–
DRS-CNN (Dutande et al., 2021a)	11.89	HIGH	MSD (Simpson et al., 2019)	0.6411	12.4461
			StructSeg (Anon, 2019)	0.6539	24.0874
Proposed Network	43	HIGH	NSCLC-Radiomics (Aerts et al., 2019)	0.7468	15.336
			Local dataset	0.6847	17.435

difference between positive and negative classes. Therefore, there is a need to develop an algorithm to solve these issues.

We implemented a segmentation network with transformer blocks employed in the deeper layers of the network. The convolution blocks are utilized in the network's initial and final layers. The transformer module helped achieve excellent segmentation results with the help of patch embedding and a multi-head self-attention mechanism. Moreover, we utilized a custom loss function that generalizes the dice-based loss using modified hyperparameter tuning.

The proposed network is trained and tested on a publicly available dataset. A local dataset is also utilized to check the robustness and generalizability of the proposed model. As visible through the results, our network can also achieve satisfactory results on local data. Although there is a difference in the evaluation metrics, the segmentation results are still significantly better. We implemented a custom loss function consisting of dice and cross entropy-based losses and achieved better results, as given in Table 3.

6. Conclusion

We implemented a novel approach and designed a vision transformer-based network with atrous depthwise separable convolutional blocks for lung tumor segmentation. Our network achieved better results than previous lung tumor segmentation approaches. Also, we checked the robustness of our network using a local dataset and observed that it could perform well on a different dataset. Furthermore, the proposed network can be tuned for other medical image modalities.

There are still some improvements that can be made to improve the segmentation accuracy further. Therefore, our future study will explore more robust architectures for lung tumor segmentation. Furthermore, a combined study using CT and histological images can efficiently and effectively analyze the tumor. Another work that can be done is to analyze cancer before and after the treatment to check the treatment response of patients.

CRediT authorship contribution statement

Shweta Tyagi: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Devidas T. Kushnure:** Investigation, Validation, Reviewing. **Sanjay N. Talbar:** Investigation, Reviewing, Supervision.

Declaration of competing interest

Declarations of interest: The authors declare that there are no conflicts of interest.

Data availability

The authors do not have permission to share data.

Acknowledgment

The authors are grateful for the support provided by Dr. Abhishek Mahajan from Tata Memorial Hospital, India, for data collection and annotation verification.

References

Abid, M.M.N., Zia, T., Ghafoor, M., Windridge, D., 2021. Multi-view convolutional recurrent neural networks for lung cancer nodule identification. Neurocomputing 453, 299–311.

Aerts, H., Wee, L., Rios Velazquez, E., Leijenaar, R., Parmar, C., Grossmann, P., Lambin, P., 2019. Data from NSCLC-radiomics [data set]. The cancer imaging archive.

Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K., 2019. Recurrent residual U-Net for medical image segmentation. J. Med. Imaging 6 (1), 014006.

Anon, 2019. Structseg challenge. URL: https://structseg2019.grand-challenge.org/Dataset/.

Anthimopoulos, M., Christodoulidis, S., Ebner, L., Geiser, T., Christe, A., Mougiakakou, S., 2018. Semantic segmentation of pathological lung tissue with dilated fully convolutional networks. IEEE J. Biomed. Health Inf. 23 (2), 714–722.

Baid, U., Talbar, S., Rane, S., Gupta, S., Thakur, M.H., Moiyadi, A., Sable, N., Akolkar, M., Mahajan, A., 2020. A novel approach for fully automatic intra-tumor segmentation with 3D U-Net architecture for gliomas. Front. Comput. Neurosci. 14, 10.

Baid, U., Talbar, S., Rane, S., Gupta, S., Thakur, M.H., Moiyadi, A., Thakur, S., Mahajan, A., 2018. Deep learning radiomics algorithm for gliomas (drag) model: a novel approach using 3d unet based deep convolutional neural network for predicting survival in gliomas. In: International MICCAI Brainlesion Workshop. Springer, pp. 369–379.

Cai, L., Long, T., Dai, Y., Huang, Y., 2020. Mask R-CNN-based detection and segmentation for pulmonary nodule 3D visualization diagnosis. IEEE Access 8, 44400–44409.

Cancer, 2021. About lung cancer. URL: https://www.cancer.org/cancer/lung-cancer/about/what-is.html.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2021. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1251–1258.

Dai, S., Lu, K., Dong, J., Zhang, Y., Chen, Y., 2015. A novel approach of lung segmentation on chest CT images using graph cuts. Neurocomputing 168, 799–807.

Ding, J., Li, A., Hu, Z., Wang, L., 2017. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 559–567.

Dutande, P., Baid, U., Talbar, S., 2021a. Deep residual separable convolutional neural network for lung tumor segmentation. Comput. Biol. Med. 105161.

Dutande, P., Baid, U., Talbar, S., 2021b. LNCDS: A 2D-3D cascaded CNN approach for lung nodule classification, detection and segmentation. Biomed. Signal Process. Control 67, 102527.

Fan, T., Wang, G., Li, Y., Wang, H., 2020. Ma-net: A multi-scale attention network for liver and tumor segmentation. IEEE Access 8, 179656–179665.

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D.M., Piñeros, M., Znaor, A., Bray, F., 2021. Cancer statistics for the year 2020: An overview. Int. J. Cancer 149 (4), 778–789.

Gao, Y., Zhou, M., Metaxas, D.N., 2021. UTNet: a hybrid transformer architecture for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 61–71.

Girshick, R., 2015. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587.

Hadjileontiadis, L.J., 2005. Wavelet-based enhancement of lung and bowel sounds using fractal dimension thresholding-Part I: Methodology. IEEE Trans. Biomed. Eng. 52 (6), 1143–1148.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

Hendrycks, D., Gimpel, K., 2016. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415.

Hooda, R., Mittal, A., Sofat, S., 2018. An efficient variant of fully-convolutional network for segmenting lung fields from chest radiographs. Wirel. Pers. Commun. 101 (3), 1559–1579.

Hossain, S., Najeeb, S., Shahriyar, A., Abdullah, Z.R., Haque, M.A., 2019. A pipeline for lung tumor detection and segmentation from CT scans using dilated convolutional neural networks. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1348–1352.

Hu, H., Li, Q., Zhao, Y., Zhang, Y., 2020a. Parallel deep learning algorithms with hybrid attention mechanism for image segmentation of lung tumors. IEEE Trans. Ind. Inform. 17 (4), 2880–2889.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Hu, Q., Souza, L.F.d.F., Holanda, G.B., Alves, S.S., Silva, F.H.d.S., Han, T., Reboucas Filho, P.P., 2020b. An effective approach for CT lung segmentation using mask region-based convolutional neural networks. Artif. Intell. Med. 103, 101792.

Huang, W., Hu, L., 2019. Using a noisy U-Net for detecting lung nodule candidates. IEEE Access 7, 67905–67915.

Jadon, S., 2020. A survey of loss functions for semantic segmentation. In: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology. CIBCB, IEEE, pp. 1–7.

Jain, S., Indora, S., Atal, D.K., 2021. Lung nodule segmentation using salp shuffled shepherd optimization algorithm-based generative adversarial network. Comput. Biol. Med. 137, 104811.

Jamil, I., Butt, S.I., 2016. Adaptive thresholding technique for segmentation and juxtapleural nodules inclusion in lung segments. Int. J. Bio-Sci. Bio-Technol. 8 (5), 105–114.

John, J., Mini, M., 2016. Multilevel thresholding based segmentation and feature extraction for pulmonary nodule detection. Proc. Technol. 24, 957–963.

Kamal, U., Rafi, A.M., Hoque, R., Wu, J., Hasan, M., et al., 2020. Lung cancer tumor region segmentation using recurrent 3d-denseunet. In: International Workshop on Thoracic Image Analysis. Springer, pp. 36–47.

Kaveh, A., Zaerreza, A., 2020. Shuffled shepherd optimization method: a new meta-heuristic algorithm. Eng. Comput.

Keshani, M., Azimifar, Z., Tajeripour, F., Boostani, R., 2013. Lung nodule segmentation and recognition using SVM classifier and active contour modeling: A complete intelligent system. Comput. Biol. Med. 43 (4), 287–300.

Khanna, A., Londhe, N.D., Gupta, S., Semwal, A., 2020. A deep residual U-Net convolutional neural network for automated lung segmentation in computed tomography images. Biocybern. Biomed. Eng. 40 (3), 1314–1327.

Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S., et al., 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale.

Li, W., Nie, S., Cheng, J., 2007. A fast automatic method of lung segmentation in CT images using mathematical morphology. In: World Congress on Medical Physics and Biomedical Engineering 2006. Springer, pp. 2419–2422.

Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., 2021. DS-TransUNet: Dual swin transformer U-Net for medical image segmentation. arXiv preprint arXiv:2106.06716.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Mathur, P., Sathishkumar, K., Chaturvedi, M., Das, P., Sudarshan, K.L., Santhappan, S., Nallasamy, V., John, A., Narasimhan, S., Roselind, F.S., et al., 2020. Cancer statistics, 2020: report from national cancer registry programme, India. JCO Glob. Oncol. 6, 1063–1075.

Mirjalili, S., Gandomi, A.H., Mirjalili, S.Z., Saremi, S., Faris, H., Mirjalili, S.M., 2017. Salp swarm algorithm: A bio-inspired optimizer for engineering design problems. Adv. Eng. Softw. 114, 163–191.

Moitra, D., Mandal, R.K., 2020. Prediction of non-small cell lung cancer histology by a deep ensemble of convolutional and bidirectional recurrent neural network. J. Digit. Imaging 33 (4), 895–902.

Naqi, S.M., Sharif, M., Yasmin, M., 2018. Multistage segmentation model and SVM-ensemble for precise lung nodule detection. Int. J. Comput. Assist. Radiol. Surg. 13 (7), 1083–1095.

Narayanan, B.N., Hardie, R.C., 2019. A computationally efficient U-Net architecture for lung segmentation in chest radiographs. In: 2019 IEEE National Aerospace and Electronics Conference. NAECON, IEEE, pp. 279–284.

Netto, S.M.B., Silva, A.C., Nunes, R.A., Gattass, M., 2012. Automatic segmentation of lung nodules with growing neural gas and support vector machine. Comput. Biol. Med. 42 (11), 1110–1121.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention U-Net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Pawar, S.P., Talbar, S.N., 2021. LungSeg-Net: Lung field segmentation using generative adversarial network. Biomed. Signal Process. Control 64, 102296.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Roy, A.G., Navab, N., Wachinger, C., 2018. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 421–429.

Sahu, S.P., Agrawal, P., Londhe, N.D., et al., 2017. A new hybrid approach using fuzzy clustering and morphological operations for lung segmentation in thoracic CT images. Biomed. Pharmacol. J. 10 (4), 1949–1961.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: Learning to leverage salient regions in medical images. Med. Image Anal. 53, 197–207.

Shi, Y., Li, H., Zhang, H., Wu, Z., Ren, S., 2020. Accurate and efficient LIF-nets for 3D detection and recognition. IEEE Access 8, 98562–98571.

Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A., 2021. Cancer statistics, 2021. CA: Cancer J. Clin. 71 (1), 7–33.

Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063.

Singadkar, G., Mahajan, A., Thakur, M., Talbar, S., 2020. Deep deconvolutional residual network based automatic lung nodule segmentation. J. Digit. Imaging 33 (3), 678–684.

Tanizaki, K., Hashimoto, N., Inatsu, Y., Hontani, H., Takeuchi, I., 2020. Computing valid p-values for image segmentation by selective inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9553–9562.

Uzelaltinbulat, S., Ugur, B., 2017. Lung tumor segmentation algorithm. Procedia Comput. Sci. 120, 140–147.

Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M., 2021. Medical transformer: Gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 36–46.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803.

Wu, W., Gao, L., Duan, H., Huang, G., Ye, X., Nie, S., 2020. Segmentation of pulmonary nodules in CT images based on 3D-UNET combined with three-dimensional conditional random field optimization. Med. Phys. 47 (9), 4054–4063.

Yeung, M., Sala, E., Schönlieb, C.-B., Rundo, L., 2022. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Comput. Med. Imaging Graph. 95, 102026.

Zheng, S., Guo, J., Cui, X., Veldhuis, R.N., Oudkerk, M., Van Ooijen, P.M., 2019. Automatic pulmonary nodule detection in CT scans using convolutional neural networks based on maximum intensity projection. IEEE Trans. Med. Imaging 39 (3), 797–805.