Received: 12 July 2023 | Revised: 1 October 2023 | Accepted: 11 October 2023

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12278

ORIGINAL RESEARCH

AML-Net: Attention-based multi-scale lightweight model for brain tumour segmentation in internet of medical things

Muhammad Zeeshan Aslam[1] | Basit Raza[1] | Muhammad Faheem[2] | Aadil Raza[3]

[1]Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad, Pakistan

[2]School of Technology and Innovations, University of Vaasa, Vaasa, Finland

[3]Department of Physics, COMSATS University Islamabad (CUI), Islamabad, Pakistan

Correspondence

Muhammad Faheem.
Email: muhammad.faheem@uwasa.fi

Funding information

University of Vaasa, Finland, Grant/Award Number: mf19099383/225.Fi

Abstract

Brain tumour segmentation employing MRI images is important for disease diagnosis, monitoring, and treatment planning. Till now, many encoder-decoder architectures have been developed for this purpose, with U-Net being the most extensively utilised. However, these architectures require a lot of parameters to train and have a semantic gap. Some work tried to make a lightweight model and do channel pruning that made a small receptive field which compromised the accuracy. The authors propose an attention-based multi-scale lightweight model called AML-Net in Internet of Medical Things to overcome the above issues. This model consists of three small encoder-decoder architectures that are trained with different scale input images along with previously learned features to diminish the loss. Moreover, the authors designed an attention module which replaced the traditional skip connection. For the attention module, six different experiments were conducted, from which dilated convolution with spatial attention performed well. This attention module has three dilated convolutions which make a relatively large receptive field followed by spatial attention to extract global context from encoder low-level features. Then these fine features are combined with the decoder's same layer of high-level features. The authors perform the experiment on a low-grade-glioma dataset provided by the Cancer Genome Atlas which has at least Fluid-Attenuated Inversion Recovery modality. The proposed model has 1/43.4, 1/30.3, 1/28.5, 1/20.2 and 1/16.7 fewer parameters than Z-Net, U-Net, Double U-Net, BCDU-Net and CU-Net respectively. Moreover, the authors' model gives results with IoU = 0.834, $F1$-score = 0.909 and sensitivity = 0.939, which are greater than U-Net, CU-Net, RCA-IUnet and PMED-Net.

KEYWORDS

artificial intelligence, machine learning, medical image processing, medical signal processing

1 | INTRODUCTION

Cancer is the most prevalent cause of death nowadays, and among the numerous forms of cancer, cancer of the brain or brain tumour is the most dangerous [1]. Any expansion in the brain may cause problems. The tumour of the brain is benign (not cancerous) or malignant (cancerous). The pressure within the human skull may increase when a benign or malignant tumour forms. This can result in potentially fatal brain damage [2]. Among brain cancers, gliomas are the frequently occurring primary brain tumours inside adults that can damage the central nervous system. Glioma is a primary brain tumour that develops from glial cells. The World Health Organisation divides glioma into four phases based on microscopic imaging and tumour activity. Low-grade gliomas (LGGs) of Grades I and II are typically benign and grow slowly. High-grade gliomas are malignant and aggressive gliomas of grades III and IV [3].

Image segmentation is important in the detection and treatment of gliomas. A precise glioma segmentation mask, for example, may aid in operational planning, postoperative monitoring and overall survival. Segmentation is 'Differentiating the tumour part from the normal tissues such as the

white matter, grey matter and cerebrospinal fluid'. To segment brain tumours CT, PET and MRI images are being used. Among all of them, MRI is the most popular because it has high-resolution images which carry more information for treatment planning and is non-invasive in nature. T1-weighted, T2-weighted, post-contrast T1-weighted and fluid-attenuated inversion recovery (FLAIR) are the common MRI sequences or modalities that are being used to segment brain tumour [4].

A precise segmentation is required to detect the accurate size of the brain tumour and its location which will help in medical diagnosis, surgery and treatment planning which increases the survival chances. However, segmenting brain tumours is challenging due to their different size and shapes that vary from patient to patient. Moreover, tumours can occur in different locations of the brain and have unclear boundaries within the background and tumour portion. Brain tumour segmentation using the conventional method is time-consuming and computationally expensive. Besides, the accuracy of brain tumour segmentation relies on experience for its degree of subjectivity. There are some regions where hospitals are available and have limited equipment. Also, expert radiologists are not available at every hospital [5]. The Internet of Medical Things (IoMT) is deployed in many applications in the healthcare system. The IoMT is providing monitoring, tracking and remote diagnosis systems [6]. Till now, many deep learning (DL)-based solutions have been provided for the segmentation task. One of them is U-Net which is popular for segmentation due to its encoder-decoder structure but it has some limitations [7]. U-Net used simple skip connection to combine features of encoder and the decoder block. When we combine low level features of the encoder block with high-level features of the decoder block then there is a semantic gap between them [8, 9]. To overcome this we proposed a model attention-based multi-scale lightweight (AML-Net) and the main contributions of our work are given below:

- We proposed an attention-based multi-scale lightweight model AML-Net that extracts information at various scales and can applied on IoMT devices.
- We designed an attention module which has three consecutive dilated convolutions followed by spatial attention to get a large receptive field and capture contextual information from encoder low-level features.
- For the attention module, we conducted six different experiments by using channel and spatial attention along with different numbers of kernels in dilated convolution. Moreover, 21 numbers of kernels in dilated convolution with spatial attention are proposed that perform well as compared to others.

The remainder of the thesis is organised as follows: Related work is presented in Section 2. The proposed model AML-Net is demonstrated in Section 3. Experimental details and results are given in Section 4. Result comparison is given in Section 5. Ablation study is given in Section 6. Finally, the conclusion is given in Section 7.

2 | RELATED WORK

Segmentation is important in medical image analysis for finding tumour positions, computer-aided surgery, organ anatomical studies and other applications. Computer-assisted segmentation approaches produce more reliable findings and assist physicians in making more accurate diagnoses. In recent years, DL segmentation algorithms have become more robust over classical and machine learning approaches due to their superior feature extraction capabilities [10]. Several DL approaches for brain tumour segmentation have yielded encouraging results. DL needs massive amounts of data to accelerate the training process. Many researchers have recently begun implementing DL in medical imaging domains such as retinal image analysis, breast image analysis, brain image analysis and others. The widespread availability of graphical processing units (GPU) and open-source software packages contribute to DL's rapid rise. Convolutional neural networks (CNN) are commonly utilised in image segmentation and classification (Appendices A1–A5).

2.1 | DL for medical image segmentation

In recent years, U-Net [7] and its different variants U-Net++ [11], U-Net 3+ [12], and Cascaded U-Net (CU-Net) [13] are proposed that provide better results for segmentation. Despite U-Net, another architecture fully convolutional network (FCN)-8s was proposed that does not have fully connected layers and input of any size can be fed to this network. The decoder of this network does not use all the information from pooling layers which makes it ineffective [14]. Later, a lightweight architecture, PMED-Net for medical imaging segmentation was proposed that has only three layers in their encoder and decoder parts. Also, this has a small number of filters which makes a small receptive field. To obtain contextual information with high accuracy, a large receptive field is required [15]. U-Net [16] and PMED-Net [17] use skip connections to mitigate this loss. However, they still have flaws. In the first skip connection, the first layer of an encoder which has low-level features is connected to the last layer of a decoder, which has higher-level features. Hence, a semantic gap between two sets of features being merged is observed.

CNNs have been scaled up to improve their accuracy in various studies [18, 19]. A FCN has a lot of architectural representations for classification. For the first time, it is used in segmentation by improving the power of classification models like AlexNet, VGG and GoogleNet. It also presents different stride rates like FCN-8s, FCN-16s and FCN-32s for accurate predictions [14]. An encoder-decoder-based architecture SegNet for segmentation was proposed which outperforms FCN, DeconvNet and DeepLab-LargeFOV. The encoder part of this architecture is like the VGG16 that has 13 convolutional layers in it. The decoder part utilises pooling indices created in the max pooling operation of the encoder and achieves non-line upsampling [16]. A lightweight network based on SegNet is proposed for the segmentation of different regions of the eye called

ORED-Net. They used a non-identical residual connection from the encoder block to the decoder to overcome information loss and pass rich information throughout the model [20].

The most popular network for medical image segmentation is U-Net. U-Net is made up of two paths: contraction and expansion. Contraction paths can aid in the extraction of more advanced features, but they also decrease the size of feature maps. To recover the size of the segmentation map, the path is expanded. However, the preceding process decreases the 'where' while increasing the 'what'. That is, we gain advanced features but lose the localisation information [7].

To improve segmentation performance, a lot of architectures are proposed based on U-Net. CU-Net was proposed to outperform U-Net in brain tumour segmentation tasks. This network used two U-Net and deep supervision making it extremely large and slow [21]. Another U-Net-based architecture, the BCDU-Net was present that has bi-directional ConvLSTM (BConvLSTM) and dense convolutions. Skip connections of U-Net are replaced by BConvLSTM that combine features extracted by encoder layers to the corresponding decoder layers in a non-linear fashion. They use densely connected convolutions to strengthen feature propagation and support feature reusability [22]. Another extension of U-Net is proposed to call MultiResUNet which uses inception blocks instead of convolutional pairs in U-Net. By using inception, they increase the power of spatial feature reusability.

A modified form of U-Net has been proposed named U-Net++. U-Net++ introduced dense skip connection and deep supervision. Deep supervision provides inference time pruning and achieves comparable performance by using one loss layer [11]. Another modified form of U-Net is Swin-UNet which is a Transformer-like U-Net proposed to obtain global and long-range semantic information to segment medical images [23]. U-Net 3+ provides accurate segmentation by using full-scale skip connections and deep supervision. Full-scale skip connection joins small and same-level features with large-level features for obtaining fine-grained details in full scale while deep supervision acquires representations from the full-scale aggregated feature map. It also has a smaller number of parameters than U-Net and U-Net++ [12].

Some researchers used different techniques to extract features at different stages. In a research, the researchers used median filters to remove irrelevant information and then fed them to fuzzy sets for clustering. This step minimises the uncertainty issues on the edge of the brain tumour. Finally, a FCN is applied to optimise the refined features [24]. In another study, the researchers used Handcraft features and deep features extracted by using VGG-Unet. These two types of features are extracted with a firefly algorithm and combined. In the VGG-Unet, the encoder part is made of VGG16 and the decoder part is the same as the traditional U-Net. At the end, features are concatenated and classified [25].

Some works take advantage of transfer learning for the detection and segmentation of brain tumour. In a study, the researchers proposed a framework that is implemented in stages. In the first stage, a fusion-based technique is applied to enhance the brain tumour contrast. In the second stage, a saliency-based segmentation is applied for tumour segmentation. In the third stage, a pre-trained model EfficientNetB0 is applied on enhanced samples images and tumour segmented images. In the end, a dragonfly optimisation is applied for feature selection [26]. In another study, the researchers proposed a transfer learning-based CNN which transfers trained AlexNet features to the CNN model which extracts important features. Some features are reduced by using max-pooling and others are carried out using dense convolution layers. This modified CNN provides better accuracy for the detection and classification of brain tumour. Although this model provides better results, it requires 91.76 million parameters to train which is computationally very expensive [27].

2.2 | Attention for medical image segmentation

Convolution operation within CNNs is the basic operation which extracts information by using spatial and channel information. Most of the studies focus on the spatial dimensions to increase the representation of the model. On the other hand, SE-Net incorporates inter-dependencies of the channels for feature re-calibration which focuses on important features and suppresses less important features [28]. In another study, the researchers extended the SE block with a dense skip connection within U-Net and named it WRAU-Net. They use adaptive pooling instead of global averaging pooling in the squeeze phase which down-sample channel dimension to $32 \times 32 \times n$ [29].

A lightweight convolutional block attention module (CBAM) was proposed that can be easily integrated with any CNN. In this block, spatial attention applies along the spatial dimension of the feature map which focuses on 'where' an informative part is. The channel attention applies along the channel dimension and focuses on 'what' is meaningful given an input [30]. Another CBAM-like architecture was proposed as bottleneck attention module, is applied to low-level features to de-noise it such as by ignoring background and gradually focusing on high-level semantical details [31]. In another article, researchers proposed a linear attention module (LAM) and used it with a channel attention module to join low-level features of the encoder with high-level features of the decoder part. As the computational complexity of the attention mechanism within the transformer is quadratic, with the LAM this complexity is reduced efficiently [32].

Previous studies show receptive field, as well as spatial information, is important to obtain higher accuracy. A bilateral segmentation network (BiSeNet) was proposed consisting of a spatial path and a context path. They also proposed an attention refinement module to refine features and a features fusion module to fuse features of spatial and contextual paths efficiently [33]. By using BiSeNet researchers proposed a novel attentive bilateral contextual network consisting of a spatial path and a context path. Spatial has 3 layers with a large number of channels that focus on low-level features and extract rich spatial information. Context path with LAM which provides adequate receptive field and extracts high-level global information with

cost efficiency. They also introduced the feature aggregation model to combine features of spatial and contextual paths [15].

In another approach, researchers proposed an efficient deep CNN based on ResNet and DenseNet. They add a connection between the first and last convolution of the ResNet block to avoid the vanishing gradient problem. They used 1 × 1 convolution in the decoder to reduce the number of channels instead of cropping channels [34]. A variation of U-Net for thermoscopic segmentation was proposed that used group normalisation in the encoder, attention gates (AG) within skip connection and a bottleneck section between the encoder and decoder. AG suppresses noisy features and focuses on important contextual features within skip connection [35].

During the last few years, U-Net [7] and its different variants have provided remarkable results in segmentation. However, these have some flaws. BCDU-Net [22] has BConvLSTM and the model must look at both forward and backward paths which are computationally expensive. The CU-Net has two U-Nets due to which the model becomes large and slow [13]. MultiResNet has generalising issues and does not perform well on datasets that have fewer instances [8]. The FCN-8s decoder does not use all the information from pooling layers which makes it ineffective [16]. U-Net++ uses dense skip connections while it does not delve far enough into full-scale information [11]. U-Net 3+ has full-scale skip connections while it does not deal with varying shapes and sizes of MRI images [36]. U-Net [7] and PMED-Net [17] use simple skip connections and have a semantic gap between the encoder and decoder. A summary of the literature review is given in Table 1.

3 | AML-NET FRAMEWORK

The proposed architecture AML-Net design is made up of three small encoder-decoder networks. Input images of different scales are given to each of the smaller networks to extract features at different scales. Each of them gives crude results which are improved on the upcoming level. Prediction produced at a kth level small network (N_k) is up-sampled to match the dimension of N_{k+1} level input. This up-sampled prediction is concatenated with N_{k+1} level input and fed to the next level small network. With this cascaded technique, the network iteratively reuses input and extracts information at different scales.

3.1 | Pyramid levels

There are three pyramid levels denoted by N_k where $k = (1, 2, 3)$, as shown in Figure 1. In each pyramid level, an individual encoder-decoder model is applied which is shown in Figure 2. Our model has a lower number of parameters which is strengthened by this pyramid structure that allows us to extract information at various stages. If the input image is $H \times W$, the corresponding input and their mask sizes are $2^{k-1} \times H \times 2^{k-1} \times W$, where $k = (1, 2, 3)$. The output of the k networks is up-sample to match the $k+1$ level dimension which is double that of the previous level networks. This up-sampled output is then concatenated to the $k+1$ level input to feed it to the next level network. This pyramid strategy boosts the network's ability to extract features or information from images of different scales of smaller regions of interest. In the first stage, the input size is 48 × 48 × 3 which provides rough results. In the second stage, the output of the previous pyramid level is up-sampled and concatenated with the original input and makes a size of 96 × 96 × 3 which provides us comparatively better results. Finlay at the third stage the previous pyramid level is up-sampled and concatenated with the original input and makes a size of 192 × 192 × 3 which provides remarkable results.

3.2 | Encoder-decoder network

The small encoder-decoder architectures are trained all alone to minimise the loss. Single encoder-decoder architecture is shown in Figure 2. This encoder-decoder network has three parts: an encoder, decoder and an attention module. The encoder part of the AML-Net is shown in Figure 2a. The encoder part has three layers and each layer has two consecutive 3 × 3 convolutions along with rectified linear unit (ReLU) which is then down-sampled with 2 × 2 max-pooling of stride two. The first block of the encoder part takes input x_{ij} and performs a convolution operation on it. In convolution operation, input x_{ij} is multiplied with W_{ij} and summed up. This total sum c_{ij} is fed to the ReLU activation function, which discards negative values in each convolution feature map as shown in Equation (3). The result 2 of activating function a_{ij} is the input for the next encoder block that performs the same operation as the previous encoder block. Convolution operations are shown in Equations (1) and (2).

$$Z = X * F \tag{1}$$

$$z(i,j) = \sum_{u=-k}^{k} \sum_{v=-k}^{k} x[u,v]F[i-u, j-v] \tag{2}$$

$$f(z) = \max(0, z) \tag{3}$$

where x denotes input, F denotes filters and z denotes the convolution feature map.

The decoder part also has three layers and each layer has a 2 × 2 up-sample followed by 2 × 2 convolution to increase the dimension of the feature map and then concatenated with the previous same-level feature map. After this two 3 × 3 convolution layers along with ReLU are applied to decode the latent code.

3.3 | Attention module

To achieve a relatively large receptive field and overcome the semantic gap between the encoder and decoder we use the attention module within the skip connection. Our proposed attention module consists of three consecutive dilated

TABLE 1 Summary of literature review.

Methods	Datasets	Strengths	Weakness
U-Net [7]	Electron microscopic (EM) images	Consists of an encoder and decoder module which extract advanced features and localisation information	Have a semantic gap between encoder and decoder
MultiResUNet [8]	SBVPI and UBIRIS.v2 datasets	Spatial features are iteratively reused due to inception blocks across different scales	Have generalising issues and did not perform well on datasets with few instances
U-Net 3+ [12]	Liver and spleen datasets	It uses full-scale skip connections, deep supervision and hybrid loss to get full-scale information and fine segmentation boundaries	Does not deal with varying shapes and sizes of MRI images
CU-Net [13]	BraTS 2017	Skip connection along with auxiliary and branch supervision provides information from low to high layers. The class imbalance problem is solved by implementing loss-weighted samplings	The model becomes slow and large as it uses two U-Nets along with branch and auxiliary supervision
PMED-Net [17]	Low-grade glioma, ISIC 2018, Nuclei and X-ray	A multi-scale lightweight model to extract semantic information from different scales	Contextual dependencies for all image regions are homogeneous and non-adaptive There is a semantic gap
SegNet [16]	SUN RGB-D indoor and road scenes	Segmentation performance is improved by using max-pooling indices between encoder and decoder	Memory is needed to keep index information during max-pooling operation. For sparse datasets it is not an effective solution as decoders create sparse feature maps
BCDU-Net [22]	DRIVE, ISIC 2018, and lung segmentation datasets	Features are reused finer due to the presence of bi-directional skip connection in ConvLSTM and convolutions which are densely connected with the encoder part	For diligent training due to the presence of Bi-directional ConvLSTM, the model must be seen in both forward and backward paths
Swin transformer [37]	ImageNet-1K, ImageNet-22K, COCO 2017 and ADE20K datasets	Swin transformer is a hierarchical transformer which uses shifted windows to get feature representation	Shifted windows do not provide self-attention to the non-overlapping local windows to get higher accuracy
U-Net++ [38]	Liver, cell nuclei, colon polyp and lung nodule datasets	Use nested and dense skip connection that overcomes the semantic gap between encoder and decoder and enhances the gradient flow	Does not delve far enough into full-scale information
Znet [39]	Low-grade glioma	A deep encoder-decoder-based architecture that provides a better dice coefficient	Did not utilise long-skip connections that are necessary for better feature reusability in symmetrical networks

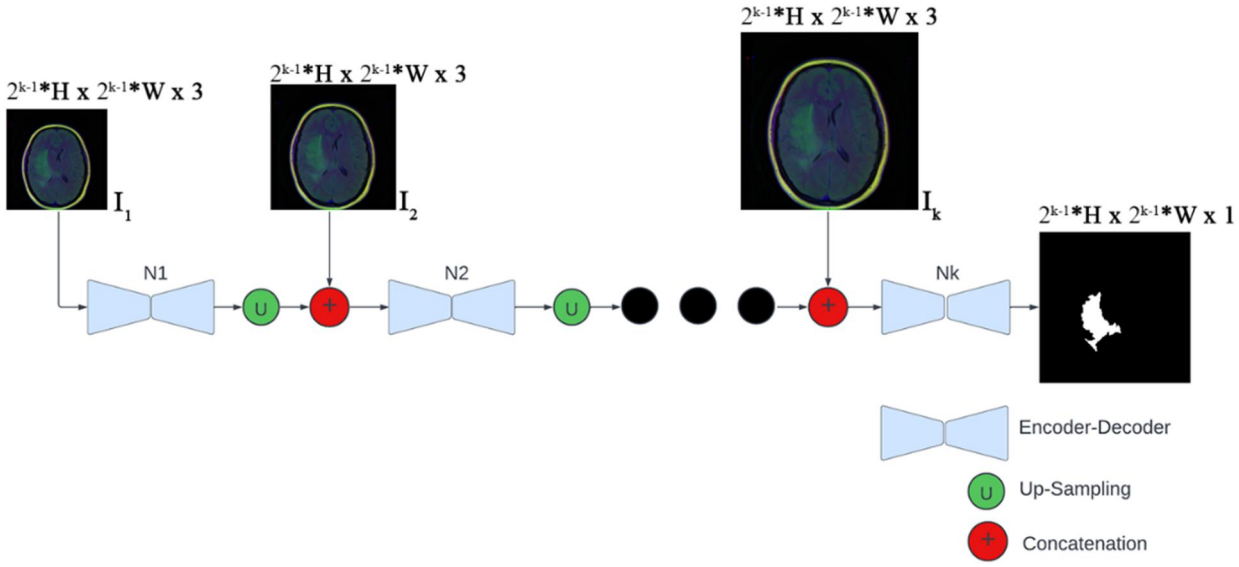

FIGURE 1 The attention-based encoder-decoder architecture. *Nk* is the number of pyramid levels where *Nk* =(1, 2, 3). At each level, features are extracted and refined at the next level. Features of different scales are extracted at four different levels of pyramids.

convolution layers followed by spatial attention. The filter size and number of filters for each dilated convolution are 3 × 3 and 21 respectively. The first dilated convolution is applied to the features map of the encoder part and the output of this is provided to the next dilated convolution. Dilated convolution is worked through the function given in Equation (4).

$$s(t) = (f \times lg)(t) = \sum_{u=-\infty}^{\infty} f(\tau)g(t - l\tau) \tag{4}$$

These three dilated convolutions dilate the features map of the encoder part and increase the receptive field which has more information than the original feature map of the encoder. This large receptive field is fed to the spatial attention block which captures the spatial attention map by aggregating features along spatial dimensions. Before computing spatial attention, max-pooling and averaging pooling are applied to capture spatial features along the channel axis. These features are concatenated and convolution is convolved on it. This convolution layer provides us with information about the part of the features that are to be highlighted or compressed.

This attention module is applied to intermediate feature maps to pay attention to important features and suppress irrelevant features. Spatial attention applies along the spatial dimension of the feature map which focuses on 'where' an informative part is [30]. At first, the Average pooling and Maximum Pooling are applied to the intermediate feature map. Average pooling provides an average of the features while Maximum polling provides prominent features. These two pooling techniques aggregate feature map channel information using two pooling techniques, resulting in two 2D maps: $F^s_{\text{avg}} \in R^{1xHxW}$ and $F^m_{\text{axs}} \in R^{1xHxW}$. Each represents average-pooled and maximum-pooled characteristics over the channel. A 7 × 7 convolution layer concatenates and convolves them to produce the 2D spatial attention output. Calculations of spatial attention are shown in Equations (5) and (6).

$$M_s(F) = \sigma\left(f^{7x7}([\text{AveragePool}(F); \text{MaximumPool}(F)])\right) \tag{5}$$

$$M_s(F) = \sigma\left(f^{7x7}\left(F^s_{\text{avg}}; F^s_{\text{max}}\right)\right) \tag{6}$$

where σ indicates the sigmoid function and f7 × 7 depicts a convolution operation of kernel size 7 × 7. The attention module of the proposed model is given in Figure 2c. The pseudocode for the implementation of the proposed model AML-Net is given in Algorithm 1.

Algorithm 1 AML-Net Pseudo Code.

Input:
Images, Masks: Brain MRI images along with their actual masks.
Output:
Weights and biases of the model.
Pseudo Code:
1. Discard brain MRI images and their masks that don't have tumour.
2. Split total of the 1373 images and their masking into train, validate and test datasets.
3. For each of the 961 training images:
   i. Resize image into $2k-1 \times H \times 2k-1 \times W \times 3$, where k is the pyramid level.
   ii. Rescale RGB image such that largest dimension is 250 pixels.
   iii. Normalise values of image between 0 and 1.

4. For pyramid level, where $k = 3$:
   i. For epoch 50:
      a. Train encoder-decoder architecture to 480 iterations selected from 961 training image.
      b. For each of the 480 images:
         i. Do forward pass and predict the segmentation of brain MRI image.
         ii. Do backward pass and use Adam optimiser to optimise loss.
   ii. For each test data set image, predict the segmentation result of k pyramid level.
5. For each test data set image, predict the segmentation result of overall model.

4 | EXPERIMENTS AND RESULTS

In this section, details of the experimental setup, loss function, performance metrics and datasets are presented along with the obtained results.

4.1 | Dataset

The Brain MRI dataset is from the cancer imaging archive. This dataset comprises 110 patients that are from five different institutions. Institutions and their contribution to the dataset are given in Table 2. All the patients suffer from LGG, and instances have manual masking of FLAIR modality. Among 110 patients, 101 patients have all the MRI sequences, six patients do not have pre-contrast sequences and nine patients do not have post-contrast sequences. The dataset contains 3929 sample images of which 1373 have tumours while 2556 have no tumour in it [40].

4.2 | Loss function

Class labels in the field of brain tumour segmentation are imbalanced which reduces the effectiveness of semantic segmentation. For most loss functions, training error-free segmentation of tiny tumour parts is difficult and time-consuming [41]. The dice are the closeness between predicted and real class labels computed using Coefficient Loss. Equation (7) shows the DCL function for binary class segmentation.

$$\text{Loss}(\text{Actual}, \text{Predicted}) = 1 - \left(2 * \frac{(\text{Actual} \cap \text{Predicted})}{(\text{Actual} \cup \text{Predicted})}\right) \tag{7}$$

4.3 | Performance metrics

To evaluate the effectiveness of our proposed model, we employ three performance metrics. We then compute the confusion matrix based on actual and predicted outputs, which gives us true positives (TP), false positives, true negatives and false negatives (FN). These numbers are then used to compute the three performance measures shown below.

4.3.1 | IoU score

'The intersection of the actual segmentation label with the predicted segmentation divided by the union between the actual and the predicted segmentation'. This is used to find the intersection between two bounding boxes [42]. The formula for calculating the IoU is given in Equation (8).

$$\text{IoU} = \frac{(\text{Actual} \cap \text{Pedicted})}{(\text{Actual} \cup \text{Predicted})} \tag{8}$$

4.3.2 | $F1$ score

The $F1$ score is calculated using 'the harmonic means of recall and accuracy'. This is sometimes referred to as a dice score. $F1$ score values lie between 1 and 0, 1 means good precision and recall and 0 means either precision or recall is not good [43]. The formula for calculating the $F1$ score is given in Equation (9).

$$F1 - \text{Score} = \frac{(2 \times \text{Pecision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{9}$$

4.3.3 | Sensitivity

Sensitivity is the ratio between accurately predicted positive values to all positive values. The formula for calculating the sensitivity is given in Equation (10).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

4.4 | Experimental setting

We use the Brain MRI dataset for our experiments that consist of 1373 positive samples. The dataset was divided into 961 images for trailing, 206 for validation and 206 for testing purposes. All the experiments are conducted on Google Colab Pro using a Nvidia P100 GPU with 25 GB RAM. A Keras framework with Tensorflow as the backend was used. Details of hyper-parameters are given in Table 3.

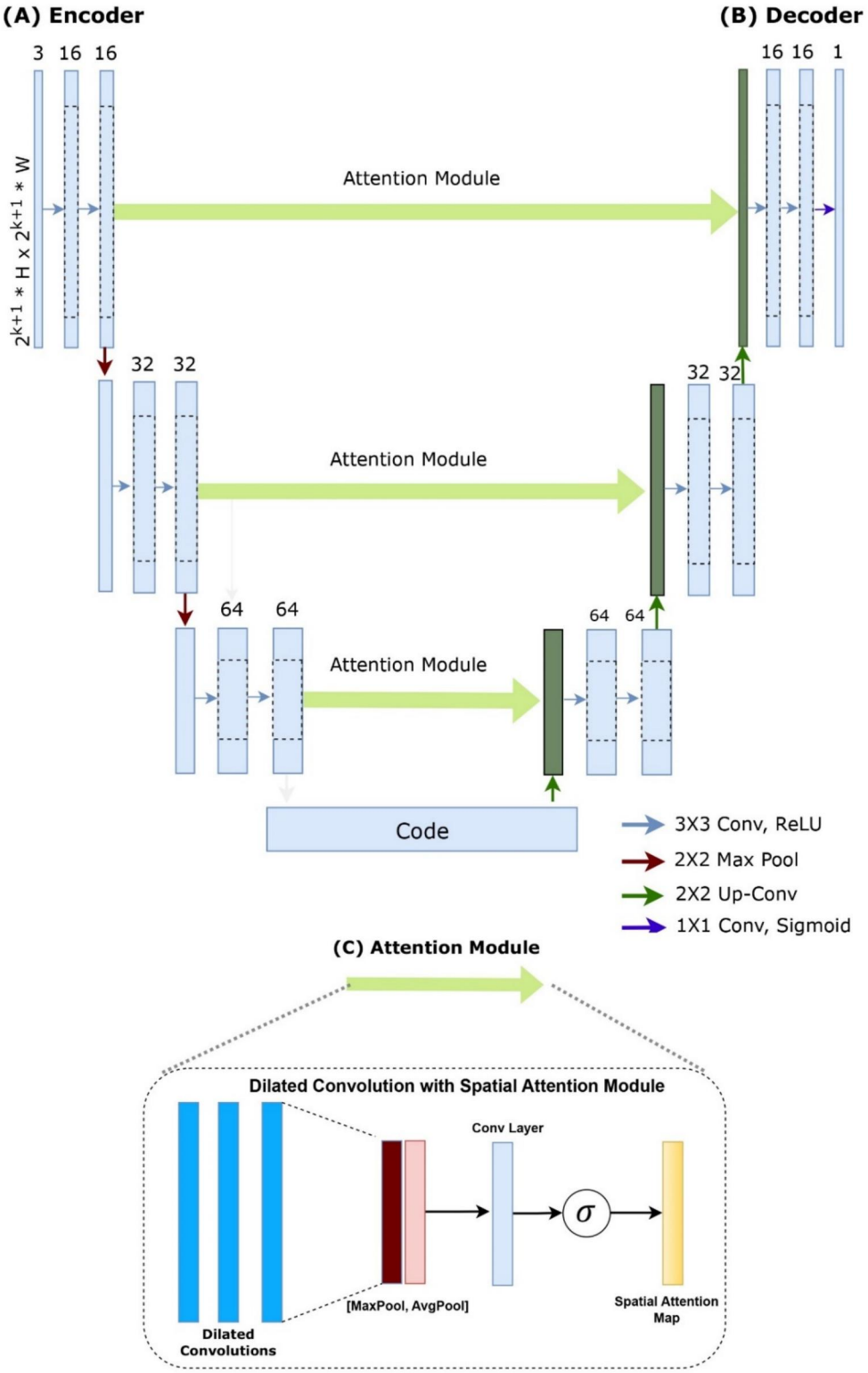

FIGURE 2 Attention-based multi-scale lightweight (AML-Net) for brain MRI segmentation.

4.5 | Results

We proposed dilated convolution with a spatial attention module by using dilated convolution and spatial attention. In this module, we implemented three consecutive dilated convolutions with a dilated rate of three, four and eight respectively. For each dilated convolution, the number of filters is set to 21. For spatial attention, we adopted it from the CBAM network [30]. Spatial attention finds 'where' the important information is in the given input feature map. The

Dilated convolution layer takes a feature map of the encoder layer as input and adds noise to it. The output of the third dilated convolution is provided to the spatial attention as input. Spatial attention applies max and average pooling sequentially to find aggregate feature maps. These two-aggregate feature maps are concatenated and a convolution layer is applied to it that gives a 2D feature map. The output of this attention module is then concatenated with the same layer decoder layer.

We implemented this attention module for up to three stages of our proposed model. Hyper-parameters detail is given in Table 3, where the value of k is set to three. At the first level,

TABLE 2 Institutions and their contribution to the dataset.

Institutions name	Number of patients
Henry Ford Hospital	45 patients
Case Western—St. Joseph's	34 patients
Thomas Jefferson University	16 patients
Case Western	14 patients
UNC	1 patient

TABLE 3 Hyper-parameters used for training the attention-based multi-scale lightweight model.

Parameter	Value
Input size	$2k-1 \times H \times 2k-1 \times W \times 3$, where $k = (1, 2, 3, 4)$, $H = 48$ and $W = 48$
Weights	He_normal
Learning rate	0.0001
B1,B2	0.9, 0.99
Epochs	50
Batch size	2
Optimiser	Adam
Loss function	Dice coefficient
Pyramid levels	3

the train dice coefficient starts at 0.2402 which gradually increases and finally ends at a dice coefficient of 0.8007. At this level, validation dice coefficient starts at 0.3796 and ends at a dice coefficient of 0.7590. The graph for the train and validation dice coefficient is shown in Figure 3. This figure shows that at initial validation the dice coefficient was higher than the train and after nine epochs train dice coefficient goes higher than the validation dice coefficient. Further train loss starts with a loss value of 0.7595 which gradually decreases and ends at a loss value of 0.1771. Like the train, validation loss starts with a loss value of 0.6204 which gradually decreases and ends at a loss value of 0.2332. The graph for train and validation loss is shown in Figure 3. This figure shows that the initial validation loss is lower than the training loss. After epoch nine, the training loss becomes lower than the validation loss. Train loss is stable, but validation loss is a little bit volatile. At the end of the training, there is a small difference between training and validation loss which shows the model is trained smoothly without overfitting. The total number of parameters for the first level is 272,916. This level has some extra parameters because of the input layer in it. Performance metric values of IoU, F-1 score and sensitivity at the first level are 0.6247, 0.7690 and 0.9703 respectively, as shown in Table 4.

At the second level, the train dice coefficient starts at 0.7173 which gradually increases and finally ends at a dice coefficient of 0.8672. At this level validation dice coefficient starts at 0.7423 and ends at a dice coefficient of 0.7723. The graph for the train and validation dice coefficient is shown in Figure 4. This figure shows that at initial validation dice coefficient was higher than the train and after the first epochs train dice coefficient went higher than the validation dice coefficient. Further train loss starts with a loss value of 0.2809 which gradually decreases and ends at a loss value of 0.1328. Like the train, validation loss starts with a loss value of 0.2577 which gradually decreases and ends at a loss value of 0.2263. The graph for train and validation loss is shown in Figure 4. Performance metric values of IoU, F-1 score and sensitivity at the first level are 0.7606, 0.8640 and 0.9517 respectively, as shown in Table 4.

At the third level, the train dice coefficient starts at 0.8081 which gradually increases and finally ends at a dice coefficient of

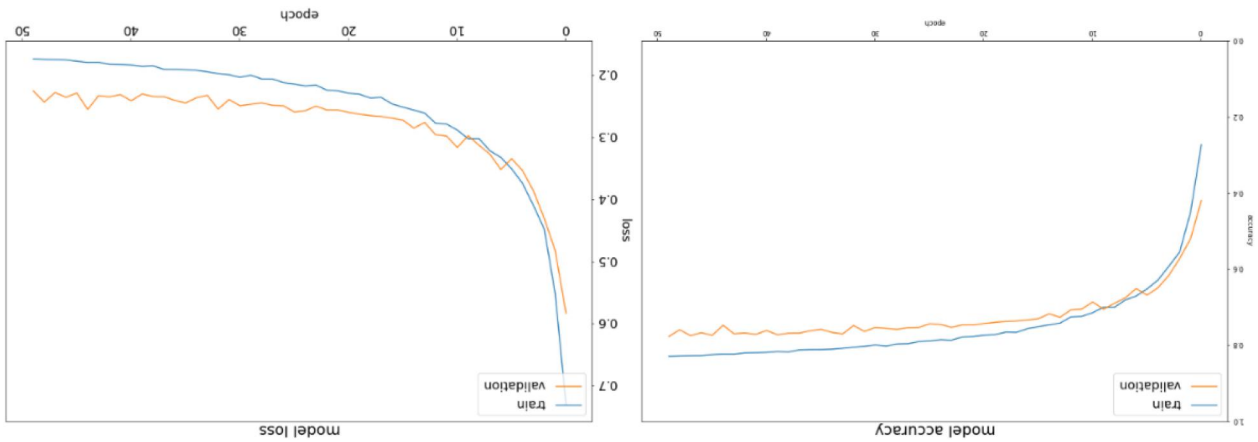

FIGURE 3 Accuracy and loss performance during training at pyramid level 1.

0.9104. At this level validation dice coefficient starts at 0.7972 and ends at a dice coefficient of 0.8019. The graph for the train and validation dice coefficient is shown in Figure 5. This figure shows that at initial validation the dice coefficient was higher than the train and after 15 epochs train dice coefficient goes higher than the validation dice coefficient. Further train loss starts with a loss value of 0.1920 which gradually decreases and ends at a loss value of 0.0896. Like the train, validation loss starts with a loss value of 0.2018 which gradually decreases and ends at a loss value of 0.1981. The graph for train and validation loss is shown in Figure 5. Performance metric values of IoU, *F*-1 score and sensitivity at the third level are 0.8347, 0.9092 and 0.9391 respectively, as shown in Table 4.

TABLE 4 Results of the proposed model at different pyramid levels.

Levels	IoU	*F*-1 score	Sensitivity
Level 1	0.6247	0.7690	0.9703
Level 2	0.7756	0.8730	0.9567
Level 3	**0.8347**	**0.9292**	**0.9391**

Note: The bold values show the efficiency of the proposed work.

5 | COMPARISON OF RESULTS

Figure 6 shows that our proposed model outperformed the other for both training and validation accuracy. We compared our results with state-of-the-art models. The first model is U-Net which is popular for segmentation due to its encoder-decoder structure but it has some limitations [7]. U-Net used a simple skip connection to combine features of the encoder and decoder block. When we combine low-level features of the encoder block with high-level features of the decoder block then there is a semantic gap between them [8, 9]. U-Net accuracy for training is smooth but for validation, it shows a volatile behaviour which shows that it did not optimise well with our experimental settings. Like U-Net, PMED-Net also used simple skip connections and have semantic gap between encoder and decoder [17]. We design an attention module in our proposed model AML-Net that is applied at the place of skip connections. This attention module refines the encoder features and then we combine these refined features with the encoder which provides better results. The PMED-Net provides less accuracy during training and validation. CU-Net used two U-Nets in cascaded form due to which the model became

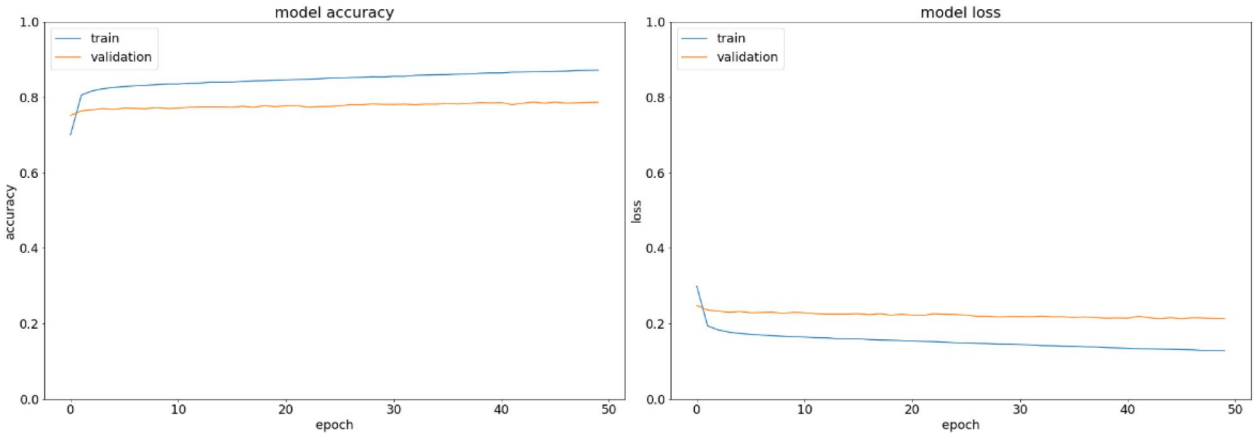

FIGURE 4 Accuracy and loss performance during training at pyramid level 2.

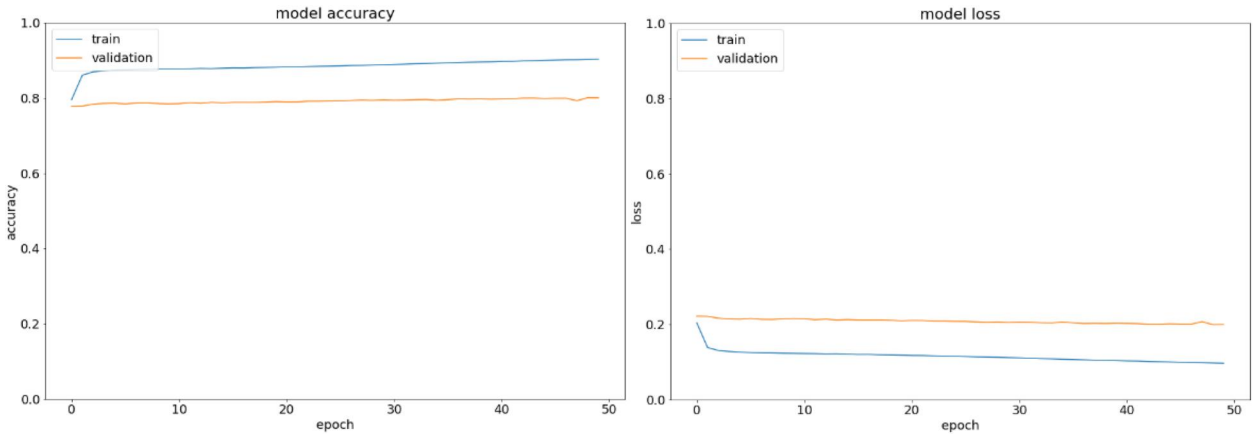

FIGURE 5 Accuracy and loss performance during training at pyramid level 3.

complex and slow and required more computational power [13]. Besides this, our AML-Net is lightweight which requires less time and computational resources. Our model also provides better results than CU-Net. RCA-Iunet used a cross spatial attention module in long skip connection for breast ultrasound segmentation [43]. But this model underfits our dataset and provides 0.4136 accuracy on the training dataset and 0.3796 on the validation dataset. Thus our proposed model AML-Net trained efficiently due to its attention module and pyramid structure.

Figure 7 shows that our proposed model has less loss than others for both training and validation datasets. U-Net loss on the training dataset is smooth, but on the validation dataset, it shows volatile behaviour which shows that it does not generalise well. PMED-Net loss is higher than our proposed model on both the train and validation datasets. RCA-Iunet on the validation dataset is lower than our proposed model, but it underfits on our dataset and did not provide better performance metrics and prediction than our proposed model.

Table 5 compares the performance metrics of our proposed model to the basic model in terms of IoU, *F*1-score and sensitivity. Our model gives results with IoU = 0.834, *F*1-score = 0.909 and sensitivity = 0.939 which is greater than U-Net, CU-Net and PMED-Net. Moreover, our proposed model has 1/43.4, 1/30.3, 1/28.5, 1/20.2 and 1/16.7 fewer parameters as compared to Z-Net, U-Net, Double U-Net, BCDU-Net and CU-Net respectively.

Figure 8 shows a prediction comparison of our proposed model with other models and the original label. RCA-Iunet did not give us accurate predictions because it underfits during training in our experimental setup. U-Net, CU-Net and RCA-Iunet did not provide accurate predictions for samples two and three while our model provides predictions close to the original mask label.

TABLE 5 Evaluation of the proposed model with other models.

Network	IoU	*F*-1 score	Sensitivity
U-Net [7]	0.815	0.898	0.874
CU-Net [13]	0.813	0.896	0.924
PMED-Net [17]	0.795	0.886	0.906
RCA-IUnet [44]	0.775	0.873	0.864
Proposed model (AML-Net)	**0.834**	**0.909**	**0.939**

Note: The bold values show the efficiency of the proposed work.

Abbreviations: AML-Net, attention-based multi-scale lightweight; CU-Net, cascaded U-Net.

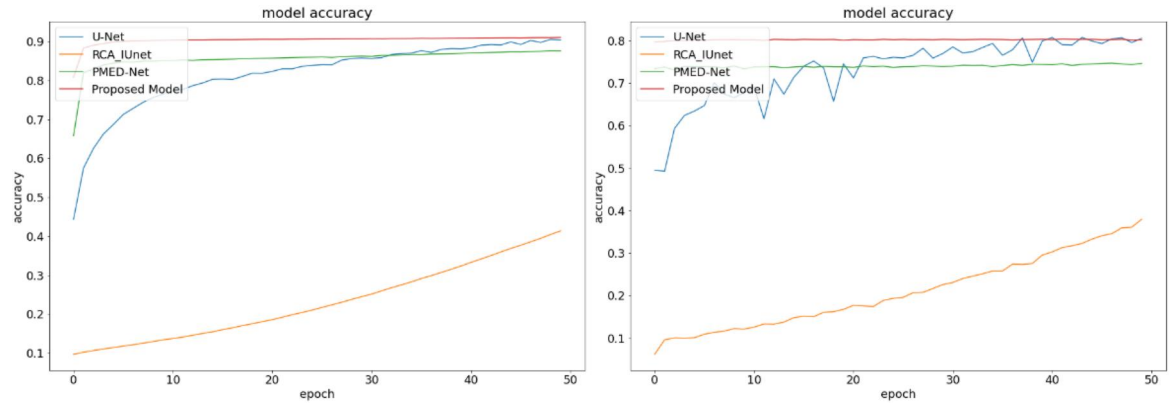

FIGURE 6 Train and validation accuracy comparison of the proposed model with others during training.

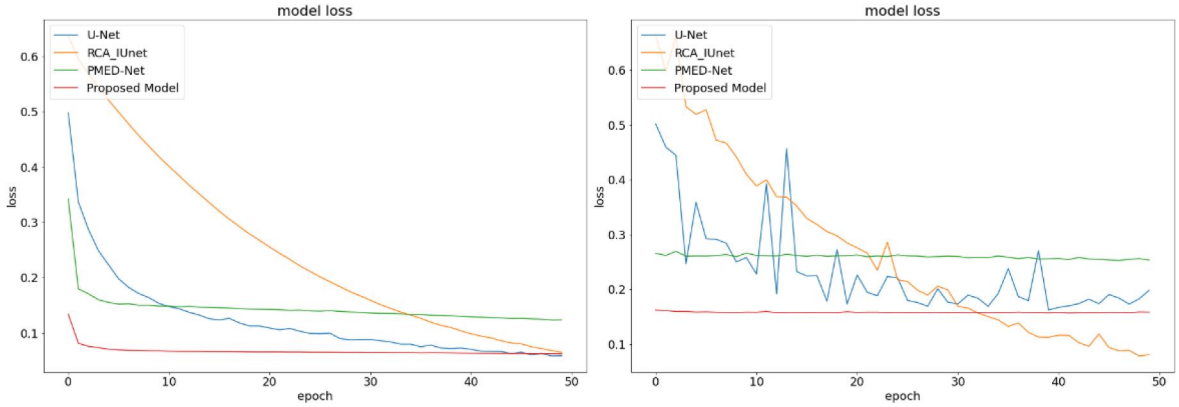

FIGURE 7 Train and validation loss comparison of the proposed model others during training.

6 | ABLATION STUDY

We conducted an ablation study of our proposed attention-based encoder-decoder models by trying different attention components. The first four experiments are based on different attention components. We do experiment with using dilated convolution with channel, channel and spatial, spatial and channel and spatial attention. We also conducted experiments by using different numbers of kernels in dilated convolution to check which performed well. Furthermore, we conducted three experiments to choose the number of kernels in dilated convolution. For this, we used 21, 32 and 2, 4, 8 different numbers of kernels in dilated convolution with spatial attention.

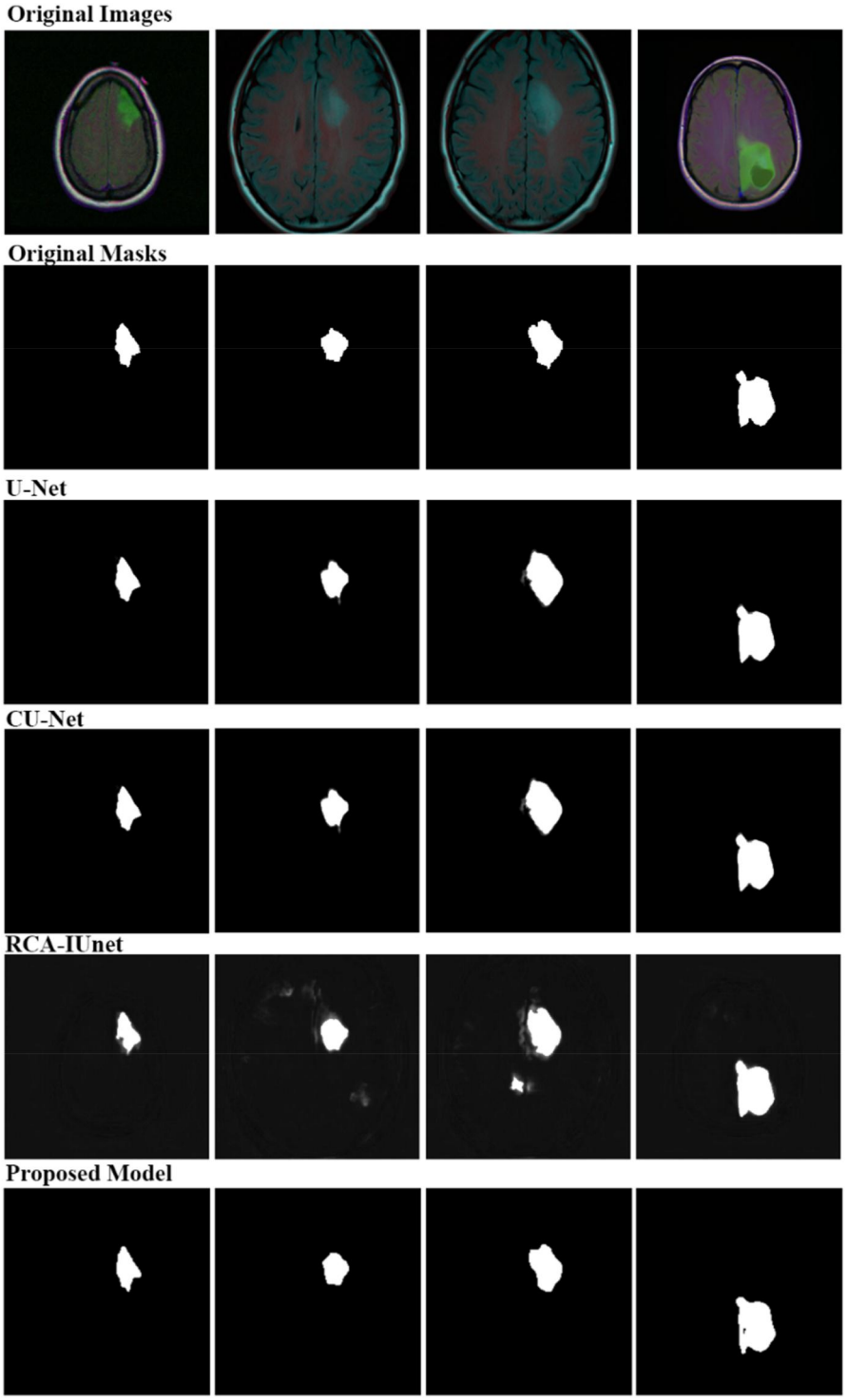

FIGURE 8 Predictions comparison with other models.

6.1 | Different attention components

In the first experiment, we designed an attention module by using dilated convolution following channel attention. In the second experiment, we developed an attention module by using dilated convolution following channel and spatial attention. In the third experiment, we developed an attention module by using dilated convolution along with spatial and channel attention. In the fourth experiment, we designed an attention module by using spatial attention. For the above four experiments, we adopted channel and spatial attention given in the CBAM network [30]. For these experiments, there are three consecutive dilated convolutions with a dilated rate of two, four and eight respectively. For each dilated convolution, the number of filters is set to 21. Detailed results of these four experiments are given in Table 6, in which dilated convolution with spatial attention performs well.

6.2 | Different number of kernels with spatial attention

We conducted different experiments by utilising different combinations of spatial and channel attention and found that dilated convolution with spatial attention outperforms all others. Next, we conducted experiments by changing the number of kernels in dilated convolution to find an optimal number of kernels that perform well to solve our problems. We design an attention module by using three consecutive dilated convolution layers followed by spatial attention. In the fifth experiment, we used two kernels in the first dilated convolution, four in the second dilated convolution and eight in the third dilated convolution layers. In the sixth experiment, we used 32 numbers of kernels in all three dilated convolution layers. From the above experiments, 21 numbers of kernels perform well. Details of these experiments are given in Table 7.

TABLE 6 Result comparison of the ablation study for four different attention modules.

		L1	L2	L3
DC + C	IoU	0.6041	0.7009	0.7558
	*F*1-score	0.7532	0.8241	0.8609
	Sensitivity	0.9447	0.9144	0.8870
DC + C + S	IoU	0.6491	0.7303	0.7716
	*F*1-score	0.7872	0.7444	0.8710
	Sensitivity	0.9374	0.9102	0.9034
DC + S + C	IoU	0.6107	0.6884	0.7527
	*F*1-score	0.7583	0.8154	0.8589
	Sensitivity	0.9398	0.9124	0.8904
DC + S	IoU	0.6247	0.7606	**0.8347**
	*F*1-score	0.7690	0.8640	**0.9092**
	Sensitivity	0.9703	0.9517	**0.9391**

Note: The bold values show the efficiency of the proposed work.
Abbreviations: C, channel; DC, dilated convolution; L, level; S, spatial.

7 | CONCLUSION

In this work, we proposed a pyramid of three small attention-based encoder-decoder architectures that are trained independently to diminish the loss. Each small encoder-decoder architecture is trained with different scale input images along with previously learned features. Features of different scales are extracted at different levels to make coarse-to-fine predictions. Moreover, we designed an attention module which replaced the traditional skip connection. The attention module is designed by conducting different experiments utilising spatial and channel attention. Among them, dilated convolution with spatial attention performs well which we proposed as our research solution. We also conducted experiments by using a different number of kernels in dilated convolution from which 21 kernels in all three dilated convolutions perform well. This attention module has three dilated convolutions which make a relatively large receptive field followed by spatial attention to extract global context from encoder low-level features. Then these fine features are combined with the decoder's same-layer high-level features. We perform our experiment on a LGG dataset provided by The Cancer Genome Atlas which has FLAIR modality and compare our results with state-of-the-art models. Our proposed model has 1/43.4, 1/30.3, 1/28.5, 1/20.2 and 1/16.7 fewer parameters as compared to Z-Net, U-Net, Double U-Net, BCDU-Net and CU-Net respectively. Moreover, our model gives results with IoU = 0.834, *F*1-score = 0.909 and Sensitivity = 0.939 which is greater than U-Net, CU-Net, RCA-IUnet and PMED-Net.

8 | LIMITATION OF THE RESEARCH

There are a lot of limitations to this research. Firstly, this research is only focused on low-grade-glioma and mainly FLAIR modality. There are a lot of attention mechanisms available in the literature. To develop an attention module, we utilise only two attention mechanisms: spatial and channels. We

TABLE 7 Results of the ablation study for three different numbers of kernels in dilated convolutions.

# of K		L1	L2	L3
2, 4, 8	IoU	0.5497	0.6930	0.7560
	*F*1-score	0.7094	0.8187	0.8610
	Sensitivity	0.9354	0.8984	0.8983
32, 32, 32	IoU	0.6410	0.7168	0.7668
	*F*1-score	0.7813	0.8350	0.8680
	Sensitivity	0.9335	0.8680	0.8953
21, 21, 21	IoU	0.6247	0.7606	**0.8347**
	*F*1-score	0.7690	0.8640	**0.9002**
	Sensitivity	0.9703	0.9517	**0.9391**

Note: The bold values show the efficiency of the proposed work.
Abbreviations: K, kernels; L, level.

implemented our proposed model AML-Net on 3 pyramids and for each pyramid, we used a different scale input image. There is a computational overhead to prepare input data for the three times instead of one time.

ACKNOWLEDGEMENTS

The authors are highly grateful to their affiliated universities and institutes for providing research facilities.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interests.

DATA AVAILABILITY STATEMENT

The data will be available upon request to the corresponding author.

CODE AVAILABILITY

The code will be available upon request to the corresponding author.

CONSENT TO PARTICIPATE

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

ORCID

Muhammad Faheem https://orcid.org/0000-0003-4628-4486

REFERENCES

1. Kesav, N., Jibukumar, M.G.: Efficient and low complex architecture for detection and classification of brain tumor using RCNN with two channel CNN. J. King Saud Univ. - Comput. Inf. Sci. 34(8), 6229–6242 (2022)
2. Han, S.: Understanding brain tumors. https://www.healthline.com/health/brain-tumor. Accessed 12 March 2023
3. Ullah, Z., et al.: Cascade multiscale residual attention CNNs with adaptive ROI for automatic brain tumor segmentation. Inf. Sci. 608, 1541–1556 (2022). https://doi.org/10.1016/j.ins.2022.07.044
4. Latif, U., et al.: An end-to-end brain tumor segmentation system using multi-inception-UNET. Int. J. Imag. Syst. Technol. 31(4), 1803–1816 (2021). https://doi.org/10.1002/ima.22585
5. Wang, J., et al.: Boosted EfficientNet: detection of lymph node metastases in breast cancer using convolutional neural networks. Cancers 13(4), 1–14 (2021). https://doi.org/10.3390/cancers13040661
6. KararKhan, M.Z., Alshahrani, H., Reyad, O.: Smart IoMT-based segmentation of coronavirus infections using lung CT scans. Ann. Oncol. 8, 19–20 (2020)
7. Ronneberger, O., et al.: Convolutional networks for biomedical image segmentation. In: International Conference Medical Image Computing and Computer-Assisted Intervention, pp. 234–241 (2015)
8. Ibtehaz, N., Rahman, M.S.: MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Network. 121, 74–87 (2020). https://doi.org/10.1016/j.neunet.2019.08.025
9. Alarood, A.A., et al.: Secure medical image transmission using deep neural network in e-health applications. Healthc. Technol. Lett. 10(4), 87–98 (2023). https://doi.org/10.1049/htl2.12049
10. Jyothi, P., Singh, A.R.: Deep learning models and traditional automated techniques for brain tumor segmentation in MRI: a review. Artif. Intell. Rev. 8(4), 1–47 (2022). https://doi.org/10.1007/s10462-022-10245-x
11. Zhou, Z., Siddiquee, M.R.: UNet++: a nested U-Net architecture for medical image segmentation. Deep Learn. Med. image Anal. Multimodal Learn. Clin. Decis. Support 9, 3–11 (2018)
12. Huang, H., et al.: UNet 3 +: a full-scale connected UNet for medical image segmentation. IEEE Int. Conf. Acoust. Speech Signal Process. 3 (3), 1055–1059 (2020)
13. Liu, H., Shen, X., Shang, F.: CU-Net: cascaded U-Net with loss weighted. Multimodal Brain Image Anal. Math. Found. Comput. Anat., 102–111 (2019)
14. Zhuang, J., et al.: Fully convolutional networks for semantic segmentation. In: Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019, pp. 847–856 (2019)
15. Li, R., et al.: ABCNet: attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. ISPRS J. Photogrammetry Remote Sens. 181(9), 84–98 (2021). https://doi.org/10.1016/j.isprsjprs.2021.09.005
16. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39(12), 2481–2495 (2017)
17. Khan, A., Kim, H., Chua, L.: PMED-Net: pyramid based multi-scale encoder-decoder network for medical image segmentation. IEEE Access 9(4), 55988–55998 (2021)
18. Khan, A.A., et al.: D2PAM: epileptic seizures prediction using adversarial deep dual patch attention mechanism. CAAI Trans. Intell. Technol. 8(1), 755–769 (2023)
19. Ali, G., et al.: A hybrid convolutional neural network model for automatic diabetic retinopathy classification from fundus images. IEEE J. Transl. Eng. Health Med. 11(3), 341–350 (2023)
20. Naqvi, R.A., Hussain, D., Loh, W.: Artificial intelligence-based semantic segmentation of ocular regions for biometrics and healthcare applications. Comput. Mater. Continua 66(1), 715–732 (2021)
21. Liu, H., et al.: CU-Net: cascaded U-Net with loss weighted, pp. 1–9
22. Azad, R., Escalera, S.: Bi-directional ConvLSTM U-Net with Densley connected convolutions (2019)
23. Cao, H., et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: ECCV 2022 Med. Comput. Vis. Work., pp. 1–14 (2022)
24. Kurdi, S.Z., et al.: Brain tumor classification using meta-heuristic optimized convolutional neural networks. J. Personalized Med. 13(2), 181 (2023). https://doi.org/10.3390/jpm13020181
25. Rajinikanth, V., Kadry, S., Nam, Y.: Convolutional-neural-network assisted segmentation and SVM classification of brain tumor in clinical MRI slices. Inf. Technol. Control 50(2), 342–356 (2021). https://doi.org/10.5755/j01.itc.50.2.28087
26. Khan, M.A., et al.: Multimodal brain tumor detection and classification using deep saliency map and improved dragonfly optimization algorithm. Int. J. Imag. Syst. Technol. 33(2), 572–587 (2023). https://doi.org/10.1002/ima.22831
27. Badjie, B., Deniz Ülker, E.: A deep transfer learning based architecture for brain tumor classification using MR images. Inf. Technol. Control 51(2), 332–344 (2022). https://doi.org/10.5755/j01.itc.51.2.30835
28. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
29. Yuan, M., Liu, Z., Wang, F.: Using the wide-range attention u-net for road segmentation. Remote Sens. Lett. 10(5), 506–515 (2019). https://doi.org/10.1080/2150704X.2019.1574990
30. Woo, S., et al.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
31. Park, J., et al.: BAM: bottleneck attention module. In: British Machine Vision Conference, BMVC 2018, no. 7 (2018)
32. Li, R., et al.: Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. IEEE Geosci. Remote Sens. Lett. 19(–5), 1–5 (2022). https://doi.org/10.1109/TGRS.2021.3093977
33. Yu, C., et al.: BiSeNet: bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 334–349 (2018)
34. Jafari, M., et al.: DRU-NET: an efficient deep convolutional neural network for medical image segmentation. In: IEEE 17th Int. Symp. Biomed. Imaging, no. 4, pp. 1144–1148 (2020)

35. Arora, R., et al.: Automated skin lesion segmentation using attention-based deep convolutional neural network. Biomed. Signal Process Control 65(12), 1–10 (2021). https://doi.org/10.1016/j.bspc.2020.102358
36. Huang, Y.I.H., et al.: UNET 3 +: a fullscale connected Unet for medical image segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1055–1059 (2020)
37. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. Proc. IEEE Int. Conf. Comput. Vis., 9992–10002 (2021). https://doi.org/10.1109/ICCV48922.2021.00986
38. Zhou, Z., Siddiquee, M.R., Tajbakhsh, N.: UNet ++: A Nested U-Net Architecture for Medical Image Segmentation UNet ++: A Nested U-Net Architecture, no. July. Springer International Publishing, Granada (2018)
39. Coefficient, M.C., Rahman, H.A., Dinov, I.D.: Znet: deep learning approach for 2D MRI brain tumor segmentation. IEEE J. Transl. Eng. Health Med. 10(3), 1–8 (2022). https://doi.org/10.1109/jtehm.2022.3176737
40. Buda, M., Saha, A., Mazurowski, M.A.: Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Comput. Biol. Med. 109(3), 218–225 (2019). https://doi.org/10.1016/j.compbiomed.2019.05.002
41. Ahuja, S., Panigrahi, B.K., Gandhi, T.K.: Fully automatic brain tumor segmentation using DeepLabv3+ with variable loss functions. In: 2021 8th Int. Conf. Signal Process. Integr. Networks, no. 8, pp. 522–526 (2021)
42. Rosebrock, A.: Intersection over union (IoU) for object detection. *pyimagesearch. com*. https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-objectdetection/. Accessed 13 March 2023
43. Rashid, J., et al.: Mouth and oral disease classification using InceptionResNetV2 method. Multimed. Tool. Appl., 1–19 (2023). https://doi.org/10.1007/s11042-023-16776-x
44. Punn, N.S., Agarwal, S.: RCA-IUnet: a residual cross-spatial attention-guided inception U-Net model for tumor segmentation in breast ultrasound imaging. Mach. Vis. Appl. 33(2), 1–10 (2022). [Online]. https://doi.org/10.1007/s00138-022-01280-3
45. Akram, A., et al.: Segmentation and classification of skin lesions using hybrid deep learning method in the Internet of Medical Things. Skin Res. Technol. 29(11), e13524 (2023). https://doi.org/10.1111/srt.13524
46. Ullah, K.A., et al.: Machine learning-based prediction of osteoporosis in postmenopausal women with clinical examined features: a quantitative clinical study. Health Sci. Rep. 6(10), e1656 (2023). https://doi.org/10.1002/hsr2.1656
47. Ali, G., et al.: Lyme rashes disease classification using deep feature fusion technique. Skin Res. Technol. 29(11), e13519 (2023). https://doi.org/10.1111/srt.13519
48. Kumar, K.K., et al.: Brain tumor identification using data augmentation and transfer learning approach. Comput. Syst. Sci. Eng. 46(2), 1845–1861 (2023). https://doi.org/10.32604/csse.2023.033927

How to cite this article: Zeeshan Aslam, M., et al.: AML-Net: attention-based multi-scale lightweight model for brain tumour segmentation in internet of medical things. CAAI Trans. Intell. Technol. 1–17 (2024). https://doi.org/10.1049/cit2.12278

APPENDIX A1

MACHINE LEARNING

In history, humans invented different types of machines to make daily life easier. These machines assist humans in transportation, industry and computing. One of the inventions to assist humans is Machine learning. Arthur Samuel defined Machine learning as 'the field of study that gives computers the ability to learn without being explicitly programed'. Machine learning enables us to handle large data efficiently. Sometimes data is very complex, and we cannot interpret it manually. So, we apply machine learning algorithms that identify the semantics among them. Many programmers and mathematicians develop machine learning algorithms to find relevant information from a huge set of datasets [45]. Different algorithms are used to solve different data problems. Classification is a machine learning technique in which we predict label based on the given data. Classification algorithms predict discrete value output, for example, 0 or 1. Regression is a machine learning technique in which we investigate the relationship between dependent variables and independent variables. Regression algorithms predict continuous values output.

Supervised Learning is a machine learning approach in which we provide input along with label and model learn mapping based on this labelled dataset. Input can be an image, text, audio or a video. These learning methods required external assistance to accomplish their task. In this, we divide an input dataset into the train and test set. Train dataset contains input X and original label y of the input. Based on this dataset algorithms learn patterns and use them to predict the output.

(i) *Decision Tree* is the supervised learning algorithm which we make predictions based on the previously answered questions. It is basically a graph that represents their data in the form of trees. A tree contains nodes and branches in which node represent an event and edges represent any condition or rule. Decision trees can be used in regression as well as in classification problems. (ii) *Naive Bayes* assumes attributes are independent given the label y and provide an estimated conditional probability of the class. It is used in classification and regression tasks depending upon the conditions an event is happening. Equation (A1) shows the formula to calculate Navie Bayes.

$$P(X|Y=y) = \prod_{i=1}^{d} P(X_i|Y=y) \tag{A1}$$

where each attribute $X = \{X1, X2, \ldots Xn\}$ consists of n attributes.

Unsupervised Learning is an approach in which we do not provide a class label during training. Datasets are unlabelled, and this approach does not require any external assistance in training. Unsupervised algorithms try to learn the structure of the data given to them. Based on the learned features, this approach assigns labels to the new data given to it. This approach is mostly used in feature reduction and clustering algorithms. (i) *K-Mean Clustering* is an unsupervised based technique in which we partition our set of observations into k clusters based on the nearest mean. In this task, we define k centres in order to find k clusters. These centres are chosen wisely because different centres give us different results. The best way is to choose a centre far away from each other. (ii) *Principal Component Analysis* is used to reduce the dimension of the data. It used interpretability and minimised information loss. We can plot 2D or 3D data easily so we reduce the dimension of the data so we can have a look at the structure of the data.

APPENDIX A2

NEURAL NETWORKS

Neural Networks also known as artificial neural networks (ANNs) are inspired by biological neurons of human brains which consist of artificial neurons that behave like a human brain. ANNs captures patterns of data and solve problems in artificial intelligence. An artificial neural network has three layers. The input layer which takes input, the hidden layer which has neurons and the output layer which provides us with the output of the model. The neuron is the basic computational component of the ANNs in which calculation of the networks is placed [46, 47]. These neurons used some activation function, and if their value exceeds some threshold value then neuron is activated, and the value is passed to the next layer. Otherwise, no neuron is activated, and data is not passed to the next neuron.

In *Supervised Neural Networks* we have original input X without label y. We give inputs to the network that performs forward propagation and provides us predicted output y. We compare this predicted output with the original label and calculate loss or error of the network. Based on this loss we do backward propagation and optimise the network weight parameters. Then the forward propagation is performed by using these updated weights. This forward and backward propagation continues to the network is fully optimised and optimal weights are achieved.

In *Unsupervised Neural Networks*, we have original input X but no label y. This network observes data and makes categories of them according to data similarity. It checks similarity of the data, for example, the correlation between input data and group them together.

APPENDIX A3

IMPORTANCE OF BRAIN TUMOUR SEGMENTATION

Cancer is the most prevalent cause of death nowadays, and among the numerous forms of cancer, cancer of the brain or brain tumour is the most dangerous. The brain is the most complicated organ, and it plays an important part in our daily lives. A brain tumour is a grouping of superfluous or defective brain cells. The human skull, which protects the brain, is quite rigid. Any expansion in such a tiny area may cause problems. The tumour of the brain is benign (not cancerous) or malignant (cancerous). When a benign or malignant tumour form, the pressure inside the human skull may increase. This can cause brain damage, which is potentially deadly. Gliomas are the frequently occurring primary brain tumours inside adults that can damage the central nervous system. Glioma is a primary brain tumour that develops from glial cells [48].

Image segmentation is important in the detection and treatment of gliomas. A precise glioma segmentation mask, for example, may aid in operational planning, postoperative monitoring, and overall survival. A precise segmentation is required to detect the accurate size of the brain tumour and its location that will help in medical diagnosis, surgery and treatment planning which increases the survival chances. There are some regions where hospitals are available and have limited equipment. Also, expert radiologists are not available at every hospital. Therefore, to assist healthcare there is needed to utilise advanced technology and develop an automated system.

APPENDIX A4

DIFFERENT APPROACHES FOR SEGMENTATION

Image segmentation is important in the detection and treatment of gliomas. A precise glioma segmentation mask, for example, may aid in operational planning, postoperative monitoring, and overall survival. Segmentation is 'Differentiating the tumour part from the normal tissues such as the White Matter, Grey matter and Cerebrospinal Fluid'. Previous research shows that there are three approaches for segmentation.

To make models computationally inexpensive, many researchers tried different methods which have their own strengths and weaknesses. Some approaches to restrict the input size are: (i) Input Restricting. ErfNet reduces the input size to make models computationally inexpensive and increases speed of the models. Although, by doing this model becomes simple and cost effective but it loses some spatial details which decreases accuracy. (ii) Channel Pruning. Some researchers prune network channels to increase speed. It also weakens the spatial capacity of the models which decreases the accuracy of the model. (iii) Drop Stage. PMED-Net and Attention-based DCNN drop the last stages of the models to make it computationally inexpensive. By dropping the last stage, downsample operation of the last stage is not performed which make small receptive field. A small receptive field does not cover the large objective that results in poor performance.

To overcome the above-mentioned problems, mostly researchers used U-Shape like architectures. U-Shape architectures consist of an encoder part and a decoder part. The encoder part gradually down-sample the input and converted it to the smaller size context representation. The decoder part takes this small representation and tried to reconstruct the context representation to the original image. However, these U-Shape architectures have two weaknesses. (i) As U-shape architecture consists of two parts, it has high resolution images and more computation which reduces the speed of the model. (ii) Most of the spatial information is lost by cropping and channel pruning which cannot be recovered.

The most popular network for medical image segmentation is U-Net. This encoder-decoder based network has skipped connections from encoder towards decoder. In encoder feature size is down sample with the help of max-pooling and up sample in decoder with D2 strides to obtain same resolution. U-Net is made up of two paths: contraction and expansion. Contraction paths can aid in the extraction of more advanced features, but they also decrease the size of feature maps. To recover the size of the segmentation map, the path is expanded. However, the preceding process decreases the

'where' while increasing the 'what'. That is, we gain advanced features but lose the localisation information.

Recently, some researchers have identified that for semantic segmentation there is a need to have spatial information as well as contextual information. Bilateral models were proposed to take advantage of both spatial and contextual information. Bilateral architectures consist of a spatial path which computes spatial information of the input and a context path which takes contextual information of the inputs that are further combined to make final prediction. At first, BiSeNet was proposed based on bilateral path. They also proposed attention refinement module (ARM) which refined contextual features and feature fusion module (FFM) which combines the spatial and context path features. ABCNet takes advantage of bilateral architecture and proposed attention enhancement module (AEM) and feature aggregation module (FAM). They deploy two AEM module to enhance the contextual features. FAM efficiently combining spatial and contextual features then simple concatenation or summation.

Bilateral segmentation network (BiSeNet) consisting of a spatial path and context path. Spatial path has 3 layers, and each layer contains convolution, batch normalisation and RelU as activation functions that get rich spatial information. Context path used Xception and obtain large receptive field by down sampling fast and at the end of this lightweight model attach global averaging pooling to obtain global information and relatively large receptive field. They also proposed attention refinement module (ARM) to refine features and features fusion module (FFM) to fuse features of the spatial and context path efficiently. They make it computationally inexpensive as the spatial path have only three layers and context path uses lightweight model. This model also provides high accuracy due to rich spatial information and relatively large receptive field provided by spatial path and context path respectively.

APPENDIX A5

IMPORTANCE OF LIGHT-WEIGHT MODELS

Light-weight models are simple, have a smaller number of parameters due to which they are trained faster and use less computing and storage resources. These models tried to reduce the model calculation, decrease the parameters and reduced running time. Lightweight models are achieved by parameter engineering, reducing the number of channels and spatial resolution and using new activation. Lightweight models overcome the problem of limited computational resources, can easily be integrated in real life scenario and improve economic values of neural networks.

Semantic segmentation methods like SegNet used complex architecture and required a lot of cost in order to obtain higher accuracy. Espnet used small encoder-decoder structure for segmentation which increases the model speed, but it significantly reduced the accuracy. To reduce the number of parameters some work crops the input size which reduces the spatial details or reduce the number of channels which make the small receptive field. A light-weight model LAENet was proposed to segment road scene segmentations which take care of both number of parameters and accuracy.

BiSeNet was proposed which is a light-weight bilateral model for segmentation. It consists of a spatial path and a context path that are further combined with FFM. They make it computationally inexpensive as the spatial path have only three layers and context path uses lightweight model. This model also provides high accuracy due to rich spatial information and relatively large receptive field provided by spatial path and context path respectively. Later, ABCNet was proposed which used BiSeNet architecture and proposed AEM and FAM module. They use LAM in context path to make it computationally inexpensive. They used ResNet-18 in context path with attention enhancement module (AEM) to extract contextual information. They also introduced the feature aggregation module (FAM) to combine features of spatial and context paths. The PMED-Net was proposed, a lightweight architecture that takes advantage of multi-scale inputs. This architecture has only three layers in each of the encoder and decoder block. In order to model computationally inexpensive, they prune numbers of channels which make a very small receptive field. Although the model is light weight, but reduces the accuracy of the model due to the small receptive field.