Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19252-2

Analysis of multimodal fusion strategies in deep learning for ischemic stroke lesion segmentation on computed tomography perfusion data

Chintha Sri Pothu Raju[1] · **Bala Chakravarthy Neelapu**[2] · **Rabul Hussain Laskar**[1] · **Ghulam Muhammad**[3]

Received: 30 January 2024 / Revised: 13 March 2024 / Accepted: 15 April 2024

Abstract
Stroke poses a significant risk to human life. Segmenting and immediately treating the stroke core stops its further development, therefore, enhancing the likelihood of survival. Convolutional neural networks (CNN) have been very successful in medical image segmentation, namely in the field of deep learning, and have produced the most advanced outcomes. Multimodal images provide superior outcomes in the segmentation of stroke lesions compared to single-modal images. The integration of input from several modalities at various levels is crucial in determining performance and producing diverse outcomes in deep learning models that use multimodalities. Further investigation is required to explore the optimal methods for processing multimodal data in CNNs, the influence of fusion on CNN learning, and the effect of fusion strategies on lesions of varying sizes. To examine the impact of a multi-modal fusion method on lesion segmentation, we assessed four models using distinct fusion techniques, including early, late, bottleneck, and hierarchical fusions. This study discusses the various fusion procedures used in segmenting the lesion using computed tomography perfusion data. In addition, both quantitative and qualitative assessments, including deep feature analysis and feature similarity, were conducted to assess the impact of the fusion technique on the model's performance. Furthermore, we examined the influence of fusion techniques on the size of the lesion. In addition, we analyzed the advantages and disadvantages of several multimodal fusion systems. Our findings demonstrate that the bottleneck fusion technique got the highest dice score, 0.582, on the Ischemic Stroke Lesion Segmentation 2018 validation data as a result of its capacity to construct complex relationships across several modalities.

Keywords CT perfusion · Deep learning · Ischemic stroke segmentation · Multimodal fusion strategies · Comparative analysis

✉ Chintha Sri Pothu Raju
chintha_rs@ece.nits.ac.in

✉ Ghulam Muhammad
ghulam@ksu.edu.sa

Extended author information available on the last page of the article

Published online: 24 April 2024

1 Introduction

The incidence of brain strokes is increasing in India, making it a prominent cause of mortality and long-term disability [17]. Stroke is one of the cerebrovascular diseases that affects the blood flow and blood vessels in the brain. A stroke can be either an ischemic stroke that occurs due to a block in the blood vessels or a hemorrhage caused by the breakage of blood vessels in the brain. Ischemic stroke holds the major share of about 80% of total cases. The stroke has different stages based on the onset time (acute ischemic stroke: $0-24$ hours, sub-acute stroke: 24 hours-2 weeks, and chronic stroke: >2 weeks) [15]. When the stroke occurs, the blood flow gets interrupted to some parts of the brain where the cells in the brain will be dead which is called the core (irreversibly damaged tissue), and brain cells around the core will get the oxygen supply enough for its survival but not enough for cognitive functioning. This part of the brain is referred to as the penumbra (salvageable brain tissue).

The quantification of brain abnormalities (brain tumors, strokes, etc.) requires brain imaging such as computed tomography (CT) and magnetic resonance imaging (MRI). The stroke can be detected in MRI (T1-weighted, T2-weighted, Fluid Attenuated Inversion Recovery (FLAIR), and Diffusion Weighted Imaging (DWI)), MR perfusion, CT (Non-Contrast CT), or CT perfusion. The advantage of perfusion maps is that they can provide sufficient information about the penumbra area, whereas MRI is more sensitive to the core area. With the information about the penumbra area, the radiologists can decide on a more suitable treatment, which benefits the affected persons. For capturing the raw CT perfusion images, a contrast bolus is injected into the blood, and performed a series of scans continuously or at predefined intervals. The raw CT perfusion data is 4-dimensional data (3-dimensional volume data over time). These raw CTP images also referred to as dynamic contrast-enhanced images, are used to generate the blood flow and time parameter maps. The blood flow parameter maps are Cerebral Blood Volume (CBV) and Cerebral Blood Flow (CBF). The time parameters are Mean Transit Time (MTT) and Time to Maximum (Tmax). These parameter maps contain the flow-related variables with each voxel of interest [25]. These four parametric maps contain different information related to the stroke-affected area. This information is vital in segmenting the core region. The stroke lesions can be seen in the brain as darker regions in CBV and CBF, whereas it is brighter regions in the Tmax and MTT [7]. This property can be seen in Fig. 1.

Generally, in medical image processing tasks such as classification and segmentation, the earlier works included only a single modality [2, 6, 41]. But, nowadays, the use of multimodal for a specific task in medical image processing has increased. The studies also suggested that multimodal input performs better than single modal input [27, 45]. The main reason behind this is that the processing of multimodal inputs gives sufficient information present in the multimodalities that helps in performing the tasks much better. Whereas, with a single-modal, the

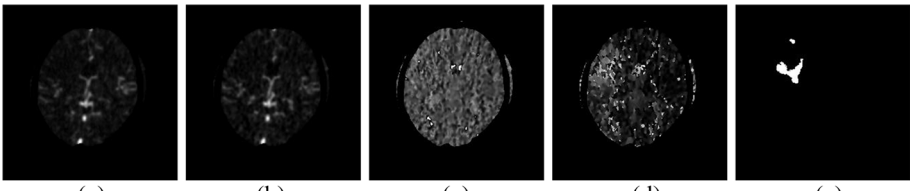

Fig. 1 An example of input images a) CBV b) CBF c) MTT d)Tmax e) Ground Truth

deep learning models may not be able to extract sufficient information for accurate segmentation. The consideration of different modalities reduces the uncertainty [45] and eliminates artifacts [22]. From the literature, extracting information from multimodal images and fusing them was performed in several ways [13, 44]. One of them is extracting the information from one modality and fusing that into the other modality using conventional image processing techniques. Later the fused image is used for further processing and discarded the image from which the features have been extracted. A multilevel fusion strategy has also been proposed by [23] for combining two different modalities that combine the features at the image level, matrix level, and feature level. Few researchers used multimodal images as input to deep learning models as different channels for extracting better features. The observations made from the literature are that feature fusion at different layers or stages yields different results [14]. Effective feature extraction from multimodalities is in demand for better segmentation of the abnormalities in medical image analysis. Even though the information from the different modalities is significant, the stage of fusing the information is also vital in improving performance.

Recently, multimodalities have been frequently used for brain stroke detection and quantification [3]. Researchers mainly followed different fusion strategies i.e. early fusion, late fusion, and other fusion strategies for fusing the features from the multimodalities. More details about these fusion strategies are summarized in Section 2. For the segmentation of stroke lesions using CTP, parameter maps are being used, but most of the developed works used early fusion [4, 8]. Few works have experimented with different feature fusion techniques such as early and late fusions. They concluded that late fusion is performing better [34]. In addition to these fusion strategies, various other fusion methods are also addressed in the literature. Although various fusion strategies are discussed in the literature, there is a lack of detailed discussion on why these strategies exhibit different behaviors. Multimodal imaging, which integrates information from diverse imaging modalities like CTP maps, has demonstrated superior results in stroke lesion segmentation compared to single-modal images. However, determining the optimal methods for integrating multimodal data within CNNs and understanding the impact of fusion strategies on segmentation performance are areas necessitating further investigation. This study aims to address these gaps by exploring and comparing four distinct fusion strategies-early, late, bottleneck, and hierarchical fusions-in the context of stroke lesion segmentation using CTP data. Despite the prevalence of different fusion methods in the literature, there remains a need for a deeper understanding and emphasis on the rationale and efficacy behind these strategies. Our approach involves analyzing the deep features learned by CNN and investigating how the CNN encodes the diversity present in medical images, such as lesion size, while also evaluating the effect of different fusion strategies on CNN learning. Furthermore, in this study, we have demonstrated four feature fusion strategies and investigated how these strategies influence the final segmentation result. The contributions of this paper are as follows:

1. We summarized the available deep learning fusion techniques on CTP data into different fusion strategies. Four multimodal fusion strategies are explored to identify the promising fusion strategy for ischemic stroke lesion segmentation.
2. Quantitative and qualitative analysis is carried out to determine the performance of the different fusion strategies.
3. The analysis of the effect of fusion strategies based on the lesion volume i.e. small and large lesions. In addition to this, the analysis of deep features and feature similarity index is also studied for all multimodal fusion strategies.

The models incorporated with early fusion, late fusion, bottleneck fusion, and hierarchical fusion strategies were designed and trained in our environment for fair comparison purposes. The organization of this paper is as follows. Section 2 gives insight into the literature on multimodal fusion strategies. Section 3 discusses materials and methodology that include the dataset, developed models to carry out the experimentation, and metrics used to compare the performance of the deep learning models. Section 4 describes the quantitative results of all multimodal feature fusion strategies in addition to qualitative analysis such as deep feature analysis, feature similarity index, and the effect of the fusion strategy on small and large lesions. Section 6 is about conclusions and future work.

2 Related works

Many researchers adopted the early fusion strategy, but their main contribution differs from fusing the multimodal images. The segmentation of stroke lesions is achieved by incorporating different conceptual knowledge into the networks. Those concepts are i) processing the multimodal data in different paths for acquiring multi-scale features i.e. Multiscale processing [3, 16, 20], ii) training with the adversarial loss along with the traditional loss, i.e. adversarial learning [42], iii) development of novel network architectures like asymmetrical encoder-decoder architecture for reducing the complexity of the network [8], iv) used dilated convolutions for enhancing the contextual information from the multimodal input images [38], v) transfer learning technique [1] and vi) generating the DWI from the perfusion maps and then segmenting the lesions on the DWI [36, 37, 40].

In the work of [34], the late fusion technique was employed in the segmentation of the stroke lesions on CTP data. All modalities were fed to four different U-Nets and different pixel-level classifiers to segment the lesion. They experimented with the voting classifier, weighted averaging, and logistic regression. The model with the logistic regression performed well. In this case, all modalities are well exploited independently by the networks for better information. The late fusion strategy will become an asset where multimodality information has little complementary information. Raju et al. [30] also utilized the late fusion strategy to segment the stroke lesion on CTP data in which a single model can process the multimodal inputs using group convolutions without the need to develop multiple networks.

Chen et al. [7] modelled an encoder that extracts the features from different modalities individually and those latent representations were fused through a hyper-fusion module in the decoder part. Deep supervision is also employed for better convergence. Zhou et al. [45] developed a multiple-encoder network that can segment brain tumour lesions on the MRI sequences. This model processed the T1-weighted, contrast-enhanced T1-weighted, T2-weighted, and FLAIR MRI sequences in individual encoders to derive features from every modality and fuse the information in the decoder using a fusing block. To extract the informative features, both attentions i.e. channel and spatial were used. A dense multipath U-Net was developed by [9] that contains different encoders for extracting the features from the different modalities and also employs the dense connections within and across the encoders. These individual features were fused at every stage in the encoder.

In an article [33], the authors followed a different approach to fuse the features from multimodalities. The architecture comprises two parallel U-Nets and each U-Net with two encoders to process the parametric maps. One U-Net is for processing the blood parameters (CBV and CBF) through different encoders and fusing those individual features at the bottleneck using cross-modality and cross-attention modules. The other U-Net is for

Table 1 Overview of fusion strategies involved in stroke lesion segmentation on CTP data

Reference	Fusion Strategy	Description of fusion involved / Contribution of the paper
Liu et al. [22]	Multiple Layer Fusion	Divided the maps into two groups i.e i) CBV, CBF, and MTT ii) Tmax. These two groups are processed in different paths. These features are combined in the last stages of the model.
Soltanpour et al. [34]	Late	Used separate U-Nets for four different modalities and the final output is generated using the probabilities from all models using the pixel-level classifier.
Abulnaga and Rubin [1]	Early	Used a PSP Net and focal loss to segment ischemic stroke lesions.
Islam et al. [15]	Early	Used the concept of adversarial learning. Bulit a discriminator to correct the higher order inconsistencies between predicted output and ground truth.
Anand et al. [4]	Early	Used the concept of Dense connections. Utilized DenseNet-121 as an encoder in the U-Net.
Song and Huang [37]	Early	Three sub models i.e., extractor for feature extraction, generator for DWI generation using CTP maps, segmentor for segmenting the lesion.
Liu et al. [20]	Early	Generated the DWI from the CTP maps using generative adversarial learning and a segmentor is used for segmenting the lesion. All the inputs are fed to the network together.
Bertels et al. [5]	Early	Used the symmetrical nature of the brain i.e., maps are flipped and registered with the original. All the inputs are fed into the model together.
Dolz et al. [9]	Multiple Layer Fusion	Used HyperDense Connectivity i.e. features are extracted from maps individually through different encoders and the features are transferred from one map to another in the form of dense connections among the encoders.
Clèrigues et al. [8]	Early	Used a more regularized training, symmetric modality and uncertainty filtering.
Pinheiro et al. [29]	Early	Utilized V-Net and U-Net for the segmentation of the lesions with emphasis on voxel normalization and depth of network.
Song [36]	Early	Generated the DWI from the CTP maps then segmented the lesions on the DWI.
Yang [42]	Early	Adversarial learning
Tureckova and Rodríguez-Sánchez [38]	Early	All the input maps are fused from the first stage itself. The work mainly focuses on the use of dilated convolution for the segmentation of lesions.
Wang et al. [40]	Early	It is improved version of work presented by [37]
Chen et al. [7]	Bottleneck	The features are extracted from individual maps separately, and then fused at the bottleneck.

Table 1 continued

Reference	Fusion Strategy	Description of fusion involved / Contribution of the paper
Shi et al. [33]	Multiple Layer Fusion	First, features are extracted from all maps using separate encoders, then blood parameters features are fused through a common decoder and similarly for time parameters. Lastly, features from both decoders are fused.
Soltanpour et al. [35]	Early	Along with four maps, 3 successive slices are also included for 3D context, and a heatmap of Tmax is also included at the input. All these maps are fused at the first stage itself.
Raju et al. [30]	Late	Used group convolution to process the CTP maps individually to extract features by using only one U-Net. This eliminated the use of training multiple U-Nets.
Zhu et al. [46]	Other	At first stage features are extracted from all the maps individually, then fused those at very next stage.
Liu et al. [21]	Early	Used transformer block to extract the global features using U-Net residual blocks. All the maps are fused at the early stages only.
Omarov et al. [28]	Early	Used the modified version of UNet. Added few regularization techniques and augmentation.
Kumar et al. [18]	Early	A modified version of model proposed by [8]
Ghnemat et al. [10]	Early	Utilized an augmentation technique based on generative adversarial networks and mutation model to increase the number of samples.

processing the time parameters (MTT and Tmax) in a similar way that has been followed for the blood parameters. Finally, two feature maps from each U-Net were fused and fined-tuned the weights at the last stage in order to create the final segmentation map. Zhu et al. [46] have developed a network that extracts different features from all the parameter maps separately and concatenates them after the first stage. Then those features are processed through the encoder and decoder structure. In the paper published by [19], the features from three distinct modalities were extracted by separate encoders and concatenated the features from all the modalities at every stage. The concatenated features from all stages in the encoders were upsampled to produce the same dimensions, concatenated, and convolved to generate the segmentation map. Table 1 gives the overview of the fusion strategies involved in the literature on stroke lesion segmentation on CTP data.

In recent years, several studies [26, 32, 39] have been proposed to segment stroke lesions using CTP data. These studies primarily utilize perfusion-weighted imaging (3D CTP images over time) to identify and segment lesions, without utilization of the CBV, CBF, MTT, and Tmax modalities. These works were not emphasized in our study as they did not incorporate any fusion strategies.

Despite learning from the literature that multimodal image segmentation outperforms single modal image segmentation. In deep learning models incorporating multimodal inputs, the level of fusion significantly impacts performance. The aforementioned papers made use of several fusion strategies. However, due to the wide variations in the cross-modalities, a

simple fusion strategy may not effectively exploit the features of the modalities [7]. Fusing these complementary pieces of information from different modalities without proper care may lead to cross-modal interference. So, implementing any fusion arbitrarily does not extract the better features, and also there are some standard questions about this multimodal fusion in deep learning that are left unanswered. Those are i) how the CNN will encode the features of different multimodal inputs ii) the analysis behind the success of the fusion strategies, iii) how fusion strategies affect the learning of CNN models, and iv) the effect of multimodal fusion strategies on small and large volume lesions. It is very difficult and unfair to compare all developed strategies as they have been implemented and analyzed in different conditions and on different data. Also, understanding the feature fusion across multiple modalities, particularly in the ischemic stroke segmentation application, is still untouched.

3 Materials and methods

3.1 Dataset

Ischemic Stroke Lesion Segmentation (ISLES) is a medical image segmentation challenge in which the delineation of stroke lesions needs to be performed on the CT Perfusion data. In ISLES 2018 [24], the data provided is from 103 patients. 63 out of 103 are for training the network and the remaining 40 are used for testing the network. Each data contains five different modalities, i.e. CBV, CBF, MTT, Tmax, and CT. The automatic segmentation model can take either all or a few of these modalities as input segmentation of stroke lesions. The ground truth is also included for the train data and it is delineated manually based on the DWI. For a fair comparison of deep learning models' performance, the test set's ground truth is not made available to the general public. To know the model's performance, segmentation maps of the test cases need to be uploaded to the SMIR website. The dimension of each modality is 256 X 256. The depth of the volume, i.e. slices in each case, ranges from 2 to 22. For this reason, the implementation of 3D segmentation models is not feasible.

3.2 Pre-processing and augmentation

The ISLES 2018 data is skull-stripped and co-registered across the modalities. The only pre-processing technique applied to data is intensity normalization. In this technique, the variance and mean of all images are set to one and zero, respectively. One of the significant limitations of the medical image datasets is the amount of the dataset, i.e. the number of training images. Moreover, deep learning models are data-driven models, which thrive on a huge number of images for training the models. In these small dataset scenarios, the models tend to overfit the data, resulting in the model's poor performance. To overcome this issue, data augmentation is necessary. Hence, a few primitive augmentation techniques have been applied to the data, such as random rotation of -45^o to $+45^o$, random horizontal flip, and random vertical flip with probabilities of 0.5. These techniques have been applied "On the fly" to save memory and dynamically change the input data patterns in every epoch.

3.3 DL models incorporated with multimodal fusion strategies

The models utilized in this paper are largely derived from U-Net [31], which was created specifically for segmenting medical images. In this paper, four multimodal fusion strategies

are investigated. They are i) Early fusion, ii) Late fusion, iii) Bottleneck fusion, and iv) Hierarchical fusion. For an extensive study of these fusion strategies, these are incorporated separately into a simple basic U-Net architecture resulting in the development of four different models. All the convolution layers involved in these models are 3 x 3 except for the last layer.

The U-Net architecture comprises an encoder section, where input images are processed to generate latent representations, and a decoder section, which produces segmentation outputs. The encoder begins with 64 feature maps, doubling at each step until reaching a maximum of 1024 feature maps. Each step includes two convolutional layers followed by batch normalization [12] and ReLU activation. Max-pooling layers reduce feature map resolution. Direct connections from an encoder to a decoder alleviate this reduction. The decoder increases feature map resolution, ultimately matching the input image size. Starting with 1024 feature maps, the decoder halves feature maps at each step, ending with 64 feature maps matching the input image resolution. A final 1 X 1 convolution generates a final feature map, which is thresholded to produce the segmentation map. Multimodal fusion strategies are also integrated into the models.

Early fusion of multimodalities is implemented on the basic U-Net structure without any changes in the network. All the modalities are fed to the model as different channels to acquire the fused feature representations of all the modalities together. Majorly the early fusion technique is adopted in most of the networks, which fuse the information in the initial layer itself. Thus, the complementary information from all the modalities is fused at the initial layer itself. The deep learning model structure that implements the early fusion is shown in Fig. 2.

In a network utilizing the late fusion strategy, each modality is independently processed to extract unique features before combining them to generate the final segmentation map for the multimodal input images, as illustrated in Fig. 3. Upon the basic U-Net structure, our model incorporates several modifications. Previous studies on late fusion typically employed

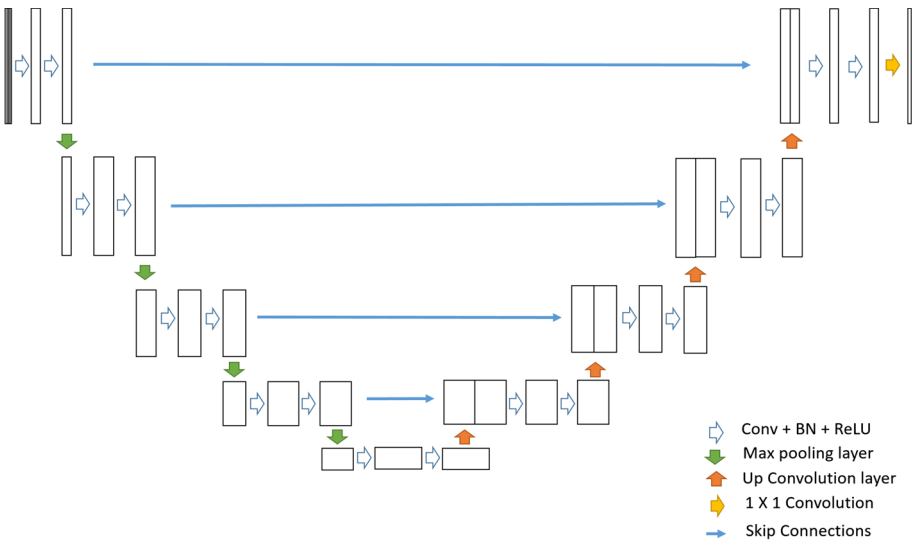

Fig. 2 Architecture of U-Net model with early fusion strategy. Conv, BN, and ReLU represent 3 X 3 convolution layer, batch normalization and rectified linear unit respectively. The number of feature maps are mentioned in the Table 3

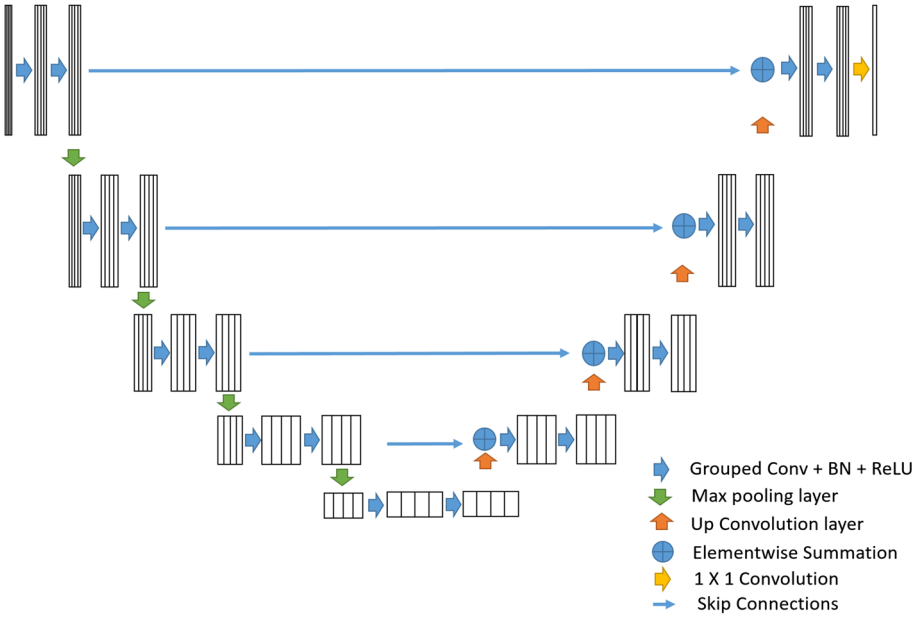

Fig. 3 Architecture U-Net model with late fusion strategy. grouped Conv, BN and ReLU represent 3 X 3 convolution layer with four groups, batch normalization and rectified linear unit respectively. The number of feature maps are mentioned in the Table 3

a method where each modality is processed separately by individual networks before feature fusion. However, in this study, we developed an alternative network architecture where all modalities are processed together within the same network, ensuring isolation among the features of each modality. Key alterations in the network to accommodate the late fusion strategy include replacing conventional convolutional layers with group convolutional layers and incorporating element-wise summation instead of concatenating features from the encoder to the decoder. This approach ensures that low-level features from the contracting path are added to high-level features of the same modality in the expanding path without interference from features of other modalities, thereby providing feature isolation.

The idea behind the bottleneck fusion strategy is that the different networks extract the features from the different modalities. The latent representations from all the modalities are fused at the middle layers of the network, i.e. the feature maps from individual modalities are fused at the end of the encoder, which is referred to as a bottleneck. Thus this fusion is called bottleneck fusion. In this fusion strategy also, we used group convolutions to avoid the multiple networks to process the input modalities independently. The architecture of the network is shown in Fig. 4.

The hierarchical fusion technique is more specific to the ISLES data, whereas the other fusion strategies can also be applied to other databases. As explained in Section 3.1, input data has four parameters, divided into two groups, i.e. blood parameter group and time parameter group. Based on this concept, in the early layers, the features are extracted from the individual maps. First, the feature maps are fused within the group and then fused across the groups in later layers. Thus, this fusion is referred to as a hierarchical fusion strategy shown in Fig. 5. In detail, the first two stages of the network extract the features from the modalities independently. Then the features from the same groups are fused after the second stage. Later

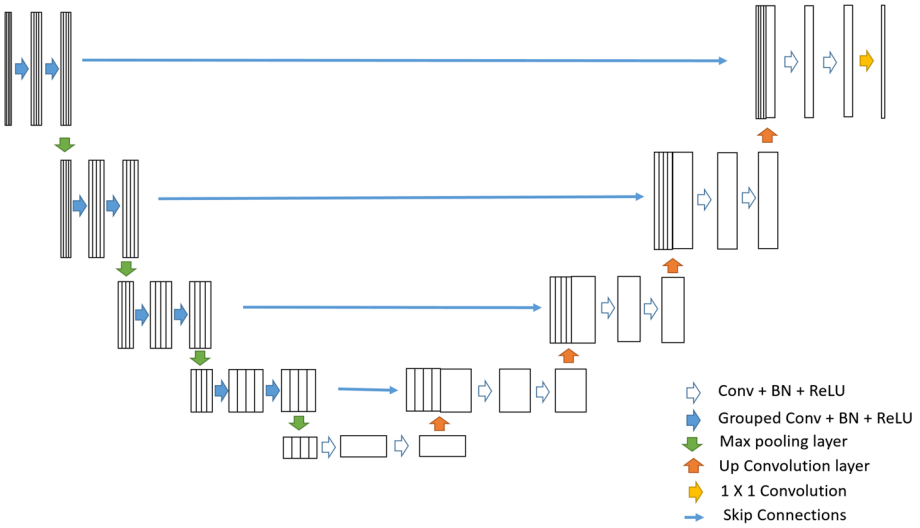

Fig. 4 Architecture of U-Net model with bottleneck fusion strategy. Conv, Group Conv, BN and ReLU represent 3 X 3 convolution layer, 3 X 3 convolution layer with four groups, batch normalization and rectified linear unit respectively. The number of feature maps are mentioned in the Table 3

two stages extract the combined features from the same group resulting in features from the two different groups. Finally, these two group features are fused at the bottleneck of the network. The latent representations of these modalities are fed to the decoder that uses the conventional convolution layers to generate the final segmentation results.

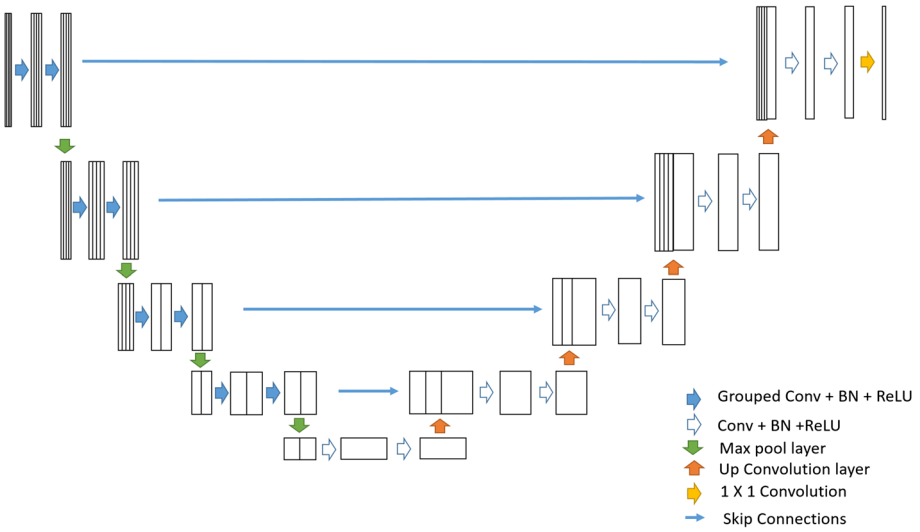

Fig. 5 Architecture of U-Net model with hierarchical fusion strategy. Conv, Group Conv, BN and ReLU represent 3 X 3 convolution layer, 3 X 3 convolution layer with groups, batch normalization and rectified linear unit respectively. The number of feature maps are mentioned in the Table 3

3.4 Group convolution

The primary distinction between conventional and group convolutional layers is the number of feature maps from the preceding layer is taken into account when creating the current feature maps. The output channel (feature map) in a traditional convolutional layer is dependent upon every channel in the preceding layer. In group convolution, the output channel depends on all the channels in a particular group instead of all channels in the preceding layer. The graphical representation is shown in Fig. 6. The input channels are shown at the top, while the output channels are shown at the bottom. Each group of feature maps is represented by a black box with a thin line, and each channel (feature maps) is represented by a rectangular box with a different color.

Consider an encoder-decoder network which consists of layers with 'n' number of convolution filters with a input depth D. Here, $I = \{I_1, I_2, I_3 \cdots\cdots I_D\}$ represents either the input image stack or the feature maps at a particular layer is the input to generate the output feature maps $O = \{O_1, O_2, \cdots\cdots O_g\}$ where g denotes the number of groups in the convolution layer, { } represents concatenation operation and $O_l = \{O_l^1, O_l^2, \cdots, O_l^m\}$ where O_l^m represents the m^{th} feature map in l^{th} group and calculation of O_l^m is shown in (1).

$$O_l^m(i, j) = \left(\sum_{d=1}^{D/g} \sum_{x=2k-1}^{2k+1} \sum_{y=2k-1}^{2k+1} F_{md}(x, y) X_d(i - x, j - y)\right) + b_m, m = 1, 2, 3, \cdots, (n/g) \tag{1}$$

Where O_{lm} denotes the m^{th} feature map in the l^{th} group, $2k + 1$ and $2k - 1$ represents the parameters of the filter side length and b_m represents bias for the m^{th} feature map. The filters are represented with F.

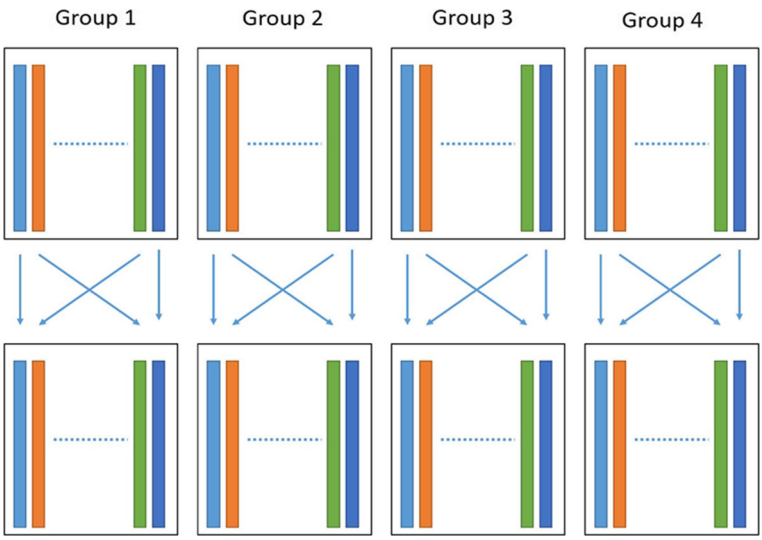

Fig. 6 Formation of feature maps in group convolutions

4 Experimental results

4.1 Experimental setup

All the experiments were carried out on an Intel-based Xeon processor configured with 128GB of RAM and an NVIDIA Tesla T4 graphic card with 16 GB of memory. The models were developed using PyTorch deep learning libraries. The deep learning models were trained for 200 epochs with a learning rate equal to 0.0001 and used Adam as an optimizer. The batch size is kept at 4. The combination of cross entropy and soft dice loss has been used as a loss function to combat class imbalance, which is very common in medical image datasets. Cross entropy loss is responsible for calculating the pixel-wise loss, whereas the soft dice loss is responsible for volume difference. The ISLES 2018 training data is divided into an 80:20 ratio for training and validation data. The metrics that are used for evaluating the models are Dice score, Hausdorff Distance (HD), Average Symmetric Surface Distance (ASSD), precision, recall, and Average Volume Difference (AVD).

4.2 Quantitative results

The architectures incorporated with different fusion strategies, discussed in Section 3.3, were developed especially to find the effectiveness of the fusion of multimodalities at the different layers. Hence, other modules which responsible for enhancing performance are not included in the architectures of deep learning models. Experiments were carried out in the same environment mentioned in Section 4.1. The same hyperparameters are followed across all the experiments. All the metrics are calculated for all the different fusion strategies and are mentioned in Table 2.

Among all the fusion strategies, bottleneck fusion achieved the highest dice score. This is a considerable increase compared to the early and late fusion strategies. In the next place, the hierarchical fusion achieved a slightly lower value than the bottleneck fusion. The mean and median scores are denoted with the 'x' symbol and solid line respectively in Fig. 7 and mean Dice scores are mentioned in Table 2. The median dice scores of early, late, bottleneck, and hierarchical fusions are 0.659, 0.693, 0.696, and 0.6871 respectively.

In addition to the dice score, other metrics such as HD, ASSD, precision, and recall are also calculated and represented in the Table 2. Lower HD and ASSD values indicate better performance of the model. Lower HD and ASSD values were achieved by the bottleneck fusion i.e., 18.72 and 3.49 respectively, indicating the closest proximity between the original and predicted surface. The precision and recall provide information about the false positives and false negatives. The higher the values the better the performance. Bottleneck fusion achieved better precision with 46.21 indicating fewer false positives. Whereas the highest

Table 2 The quantitative results of all the fusion strategies

Fusion Type	Dice Score	HD	ASSD	Precision	Recall
Early Fusion	0.564 ± 0.28	23.19	3.84	45.11	42.64
Late Fusion	0.569 ± 0.28	20.30	4.43	44.38	43.60
Bottleneck Fusion	$\mathbf{0.582 \pm 0.26}$	**18.72**	**3.49**	**46.21**	45.44
Hierarchical Fusion	0.574 ± 0.27	19.41	3.79	42.62	**46.37**

The values highlighted in bold are the best values. The values presented are averages across all fusion strategies

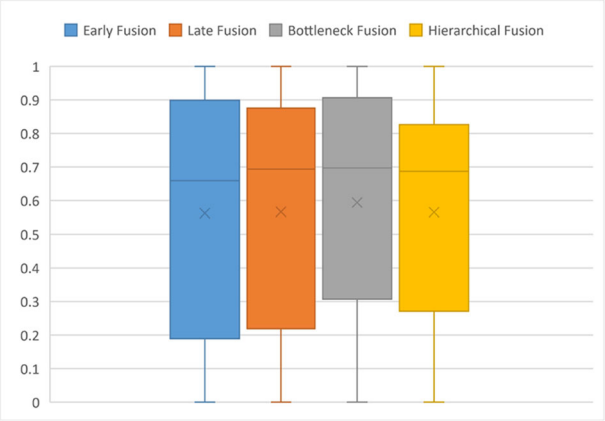

Fig. 7 Boxplot representation of Dice scores achieved in different fusion strategies

recall is achieved by the hierarchical fusion represents fewer false negatives. Bottleneck fusion shows a good balance between precision and recall indicating better performance by balancing the false positives and negatives.

The architectures exhibit close similarities with slight variations. Table 3 gives more insights into the architectural differences of all fusion strategies. The first column is in the form of [–,–,–], and that the first number indicates the number of features, and the later two numbers denote the resolution of feature maps. The following two columns represent the type of layer and the number of parameters associated with that corresponding layer. All models contain the same resolutions for feature maps across all stages, but their formation diverges. As outlined in the methodology, differences among the models primarily lie in their convolution layers. Instead of conventional convolutional layers, group convolutions are utilized, as specified in the layer column of Table 3 in the format (G=n). For instance, Conv2d (G=4) denotes a convolution layer with four groups. The adoption of group convolutions does not alter map resolution. However, parameters within the layer differ based on the number of groups utilized.

The last row of Table 3 indicates the total number of parameters in deep architectures. The computational cost is also an important factor in deep models and is decided by the number of trainable parameters in the network. The computational time is proportional to the number of parameters in the architecture. The feature fusion strategies are arranged in ascending order based on the computational complexity. They are i) late fusion strategy, ii) bottleneck fusion, iii) hierarchical fusion, and iv) early fusion. The late fusion strategy requires a minimum number of parameters around 7 million. The bottleneck fusion network employed a moderate amount of parameters equal to 16.91 million. The hierarchical feature fusion strategy utilized 28.63 million parameters. Finally, the early fusion strategy requires the highest number of parameters at about 31 million.

4.3 Analysis based on the volume of the lesion

This analysis is to explore the effect of multimodal fusion on small and large lesions. The images within the validation set are divided into two groups based on the volume of lesions present in each slice i.e., small and large lesions. Specifically, the average volume is computed

Table 3 Architecture details of all models and layers with the number of parameters

Output Shape	Early Fusion		Late Fusion		Bottleneck Fusion		Hierarchical Fusion	
	Layer	Param[a]	Layer	Param	Layer	Param	Layer	Param
[64, 256, 256]	Conv2d[b]	2,368	Conv2d (G=4)	640	Conv2d (G=4)	640	Conv2d (G=4)	640
[64, 256, 256]	B[c] + R[d]	128	B + R	128	B + R	128	B + R	128
[64, 256, 256]	Conv2d	36,928	Conv2d (G=4)	9,280	Conv2d (G=4)	9,280	Conv2d (G=4)	9,280
[64, 256, 256]	B + R	128	B + R	128	B + R	128	B + R	128
[64, 128, 128]	Maxpool		Maxpool		Maxpool		Maxpool	
[128, 128, 128]	Conv2d	73,856	Conv2d (G=4)	18,560	Conv2d (G=4)	18,560	Conv2d (G=4)	18,560
[128, 128, 128]	B + R	256	B + R	256	B + R	256	B + R	256
[128, 128, 128]	Conv2d	1,47,584	Conv2d (G=4)	36,992	Conv2d (G=4)	36,992	Conv2d (G=4)	36,992
[128, 128, 128]	B + R	256	B + R	256	B + R	256	B + R	256
[128, 64, 64]	Maxpool		Maxpool		Maxpool		Maxpool	
[256, 64, 64]	Conv2d	2,95,168	Conv2d (G=4)	73,984	Conv2d (G=4)	73,984	Conv2d (G=2)	1,47,712
[256, 64, 64]	B + R	512	B + R	512	B + R	512	B + R	512
[256, 64, 64]	Conv2d	5,90,080	Conv2d (G=4)	1,47,712	Conv2d (G=4)	1,47,712	Conv2d (G=2)	2,95,168
[256, 64, 64]	B + R	512	B + R	512	B + R	512	B + R	512
[256, 32, 32]	Maxpool		Maxpool		Maxpool		Maxpool	
[512, 32, 32]	Conv2d	11,80,160	Conv2d (G=4)	2,95,424	Conv2d (G=4)	2,95,424	Conv2d (G=2)	5,90,336
[512, 32, 32]	B + R	1024	B + R	1,024	B + R	1,024	B + R	1,024
[512, 32, 32]	Conv2d	23,59,808	Conv2d (G=4)	5,90,336	Conv2d (G=4)	5,90,336	Conv2d (G=2)	11,80,160
[512, 32, 32]	B + R	1024	B + R	1,024	B + R	1,024	B + R	1,024
[512, 16, 16]	Maxpool		Maxpool		Maxpool		Maxpool	
[1024, 16, 16]	Conv2d	47,19,616	Conv2d (G=4)	11,80,672	Conv2d	47,19,616	Conv2d	47,19,616
[1024, 16, 16]	B + R	2048	B + R	2,048	B + R	2,048	B + R	2,048
[1024, 16, 16]	Conv2d	94,38,208	Conv2d (G=4)	23,60,320	Conv2d	94,38,208	Conv2d	94,38,208

Table 3 continued

Output Shape	Early Fusion		Late Fusion		Bottleneck Fusion		Hierarchical Fusion	
	Layer	Param[a]	Layer	Param	Layer	Param	Layer	Param
[1024, 16, 16]	B + R	2048	B + R	2,048	B + R	2,048	B + R	2,048
[512, 32, 32]	ConvT[e]	20,97,664	ConvT (G=4)	5,24,800	ConvT	20,97,664	ConvT	20,97,664
[512, 32, 32]	Conv2d	47,19,104	Conv2d (G=4)	5,90,336	Conv2d	47,19,104	Conv2d	47,19,104
[512, 32, 32]	B + R	1024	B + R	1,024	B + R	1,024	B + R	1,024
[512, 32, 32]	Conv2d	23,59,808	Conv2d (G=4)	5,90,336	Conv2d	23,59,808	Conv2d	23,59,808
[512, 32, 32]	B + R	1024	B + R	1,024	B + R	1,024	B + R	1,024
[256, 64, 64]	ConvT	5,24,544	ConvT (G=4)	1,31,328	ConvT	5,24,544	ConvT	5,24,544
[256, 64, 64]	Conv2d	11,79,904	Conv2d (G=4)	1,47,712	Conv2d	11,79,904	Conv2d	11,79,904
[256, 64, 64]	B + R	512	B + R	512	B + R	512	B + R	512
[256, 64, 64]	Conv2d	5,90,080	Conv2d (G=4)	1,47,712	Conv2d	5,90,080	Conv2d	5,90,080
[256, 64, 64]	B + R	512	B + R	512	B + R	512	B + R	512
[128, 128, 128]	ConvT	1,31,200	ConvT (G=4)	32,896	ConvT	1,31,200	ConvT	1,31,200
[128, 128, 128]	Conv2d	2,95,040	Conv2d (G=4)	36,992	Conv2d	2,95,040	Conv2d	2,95,040
[128, 128, 128]	B + R	256	B + R	256	B + R	256	B + R	256
[128, 128, 128]	Conv2d	1,47,584	Conv2d (G=4)	36,992	Conv2d	1,47,584	Conv2d	1,47,584
[128, 128, 128]	B + R	256	B + R	256	B + R	256	B + R	256
[64, 256, 256]	ConvT	32,832	ConvT (G=4)	8,256	ConvT	32,832	ConvT	32,832
[64, 256, 256]	Conv2d	73,792	Conv2d (G=4)	9,280	Conv2d	73,792	Conv2d	73,792
[64, 256, 256]	B + R	128	B + R	128	B + R	128	B + R	128
[64, 256, 256]	Conv2d	36,928	Conv2d (G=4)	9,280	Conv2d	36,928	Conv2d	36,928
[64, 256, 256]	B + R	128	B + R	128	B + R	128	B + R	128
[1, 256, 256]	Conv2d	65	Conv2d	65	Conv2d	65	Conv2d	65
		31.04M		6.99M		16.91M		28.63M

G = n in the layer column represents the number of groups. Where n indicates the number of groups in that particular layer
[a]Parameters, [b]Convolution, [c]Batch Normalization, [d]ReLU, [e]Transposed Convolution. ReLU does not require any parameters

Table 4 Comparison of performance of different fusion strategies for small and large volume lesions

Fusion Type	Dice Score		HD		ASSD		AVD	
	Small	Large	Small	Large	Small	Large	Small	Large
Early Fusion	0.2024	0.6958	21.80	25.38	6.447	2.773	1.638	3.496
Late Fusion	0.1992	0.7146	17.98	24.02	5.959	1.961	**1.147**	3.748
Bottleneck Fusion	**0.2392**	**0.7351**	**15.39**	23.77	**4.433**	2.057	1.589	**3.488**
Hierarchical Fusion	0.2221	0.7341	16.75	**23.57**	5.064	**1.792**	1.530	3.520

The bold entries indicate the best values obtained

across all cases. Images with stroke lesion volumes lower than the calculated average fall under the small volume group, while those with volumes equal to or greater than the average are categorized as large-volume lesions. Table 4 denotes values for the different evaluation metrics for both small and large lesions. All the fusion strategies achieved good performance in detecting large lesions. Among all these, bottleneck fusion produced the best results for the large lesions, and hierarchical fusion also achieved similar results.It can be noted that, in the case of small lesions, bottleneck fusion achieved the best dice over the other fusion strategies. As a whole, the performance of the bottleneck fusion strategy worked better than all others in almost all the metrics for small lesions. The late fusion achieved the best than the other fusion strategies in AVD metric indicating that the prediction of lesion volume is nearer to the reference volume. Even though the AVD is less in late fusion, the higher dice score achieved by the bottleneck fusion represents more overlap between the predicted and reference volumes.

5 Discussion

The utilization of CNN models, especially U-Net, in medical image processing has been phenomenal in achieving state-of-the-art results in medical image segmentation tasks by overcoming the drawbacks of conventional image processing techniques. The deep models developed in this paper were almost similar in terms of architecture except for the level of fusion of multimodal information. Even though the networks are identical, the results produced by those networks are different. This is mainly due to the fusion of multimodal information at the various levels in the deep learning model. Although the models are producing the results, how well CNN learns to encode the multimodal data, especially in this context, and how the feature fusion at various levels affects the feature maps and the learning procedure of the deep learning model. The aforementioned research questions can be addressed to an extent with the help of concepts such as complementary information, joint representation learning, information integration, and feature representation in deep learning [43].

Complementary Information: Each modality comprises different information. The modalities have core and penumbra information which implies these modalities will have different characteristics. These contribute differently to the final core segmentation result. So, fusing at different levels the characteristics or features from multiple modalities will definitely influence the result. The fusion strategy enables the model to leverage the complementary information from multimodalities leading to enhanced performance and improved understanding of the data.

Joint representations: Joint representation learning is utilized by multimodal fusion techniques, which also enable the model to reflect complex relations and interactions between modalities. The model can learn more robust and discriminative representations that better capture the underlying patterns of multimodal data by integrating data from several modalities.

Information integration: The fusion strategy regulates how information from various modalities is combined and incorporated within the CNN. Early fusion integrates the information from the multiple modalities at the input level, allowing joint representative learning from the beginning. Whereas the late fusion technique combines the information at the decision in which the joint feature learning is minimal. The bottleneck fusion, in the encoder, extracts the features from the multiple modalities separately and all the information is fused at the bottleneck between the encoder and decoder. The bottleneck fusion allows joint representative learning from the bottleneck after the fusion of the information from multiple modalities.

Feature representation: Different fusion strategies can result in variation in feature representation learned by the CNN. As discussed above, the early fusion allows the joint representative learning which can capture the shared modality feature. Late fusion extracts the modal-specific features from multiple models which may be helpful in the modalities having very different characteristics. The fusion strategy can impact the level of integration of the learned feature representations. The models trained on the different fusion strategies have different feature response characteristics. Analyzing the deep features across the various fusion strategies may help in understanding the effects of feature fusion at different levels. We adopted a CNN visualization strategy to visualize the inherent deep features of CNN. Understanding the features of the deep model helps not only in ensuring how the model learns practical information from the images but also in connecting this information with patterns recognizable by humans [11]. The similarity between the features gives more insights into the learning of the models.

5.1 Deep feature analysis

The deep learning network has the ability to extract the different features on its own. This ability made CNN different from machine learning algorithms. The filters are the main ones responsible for the feature generation at each stage. The features in the deep learning model are represented in (2) as follows:

$$Z = f(I; W_c) \tag{2}$$

Where Z represents the deep features, I is the input image stack or the feature maps from the previous layer, and W_c is the weight parameters of the deep learning model.

Several observations have been made after visualizing the feature maps from several stages and different fusion strategies. Figure 8 shows the features at various stages of deep learning models. It contains five columns, each with four distinct feature maps generated from the same level. Each column represents various layers of feature maps starting from early to a deeper level. It can be observed that the early-level feature maps do contain low-level features such as texture and edges. As the feature maps go deeper into the layers, feature maps contain the higher-level features extracted from the weighted combination of the previous layer features. The higher-level features are more abstract in nature, clearly highlighting a few regions in the image. The highlighted regions in the feature maps represent the most significant

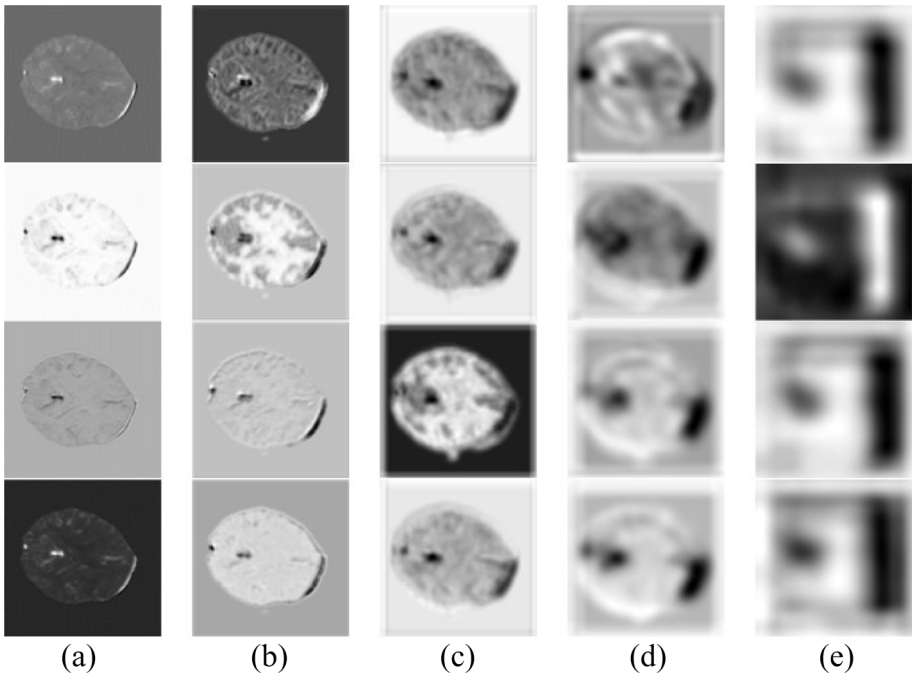

Fig. 8 Feature maps generated at different stages of a fusion strategy. From (a) to (e), columns represent feature maps at the first stage, and the last column represents feature maps at the bottleneck. All feature maps shown here are selected randomly, and these are the features generated after the second convolution layer of each stage. All images are resized to the same size for a good visual appearance

differentiable patterns learned by the filters. Most of the other regions are predominantly less significant as per the model learned weights. The above observation suggests that the regions in the early layers contain the most helpful information, which can be interpreted easily as they directly take the images as input. Even though the later layers contain the information, it is difficult to interpret as they receive the information from the previous layers, which can be seen in Fig. 8. Moreover, the regions that are found to be significant in one feature map, the same region may not be significant in the other feature map. One more important observation is that most of the feature maps have similar significant regions as the network goes deeper.

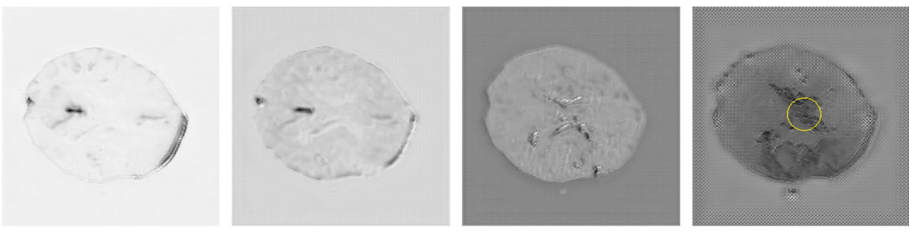

Fig. 9 Features maps derived in the last stage of the late fusion strategy. (a) Feature map from CBV (b) Feature map from CBF (c) Feature map from MTT (d) Feature map from Tmax

The feature maps shown in Fig. 9 are generated in the last stage of the late fusion strategy. The late fusion features belong to four different groups derived from different modalities, and features across the groups are also different in nature. The similarity among the features is discussed in detail in Section 5.2. One of the observations made from the features of the late fusion strategy is that the final segmentation result is biased towards one particular group of features obtained from a specific modality. The highlighted region in the last image (particular to one group's features) dominates the features of the other groups. That highlighted part is falsely recognized as a part of the lesion. This phenomenon may be due to the lack of common contextual information about the different modalities.

5.2 Feature similarity analysis

The ability to extract the features by actively modifying the filters based on the data fed to the network puts deep learning forth among all other techniques. How well these feature maps are generated and how these are distinct from each other are also factors that affect the performance of the deep learning model. To know how these features are different from each other, we made use of the cosine similarity concept. It measures how close or similar the two features are. This similarity measure was calculated for the features generated after the first stage. Figure 10(a) shows the cosine similarity index between each and every feature map at the first stage of the encoder path. In it, each pixel (value) indicates the relation between the two feature maps, and the color of the pixel indicates how strongly they are correlated. The green color denotes the less correlated features, and the yellow represents a high correlation between the features.

From the architectures, we know that 64 feature maps are generated in the first stage using 64 different filters. Each feature map is compared to all feature maps in the same stage and generates a similarity value. Hence, a 64 X 64 matrix contains features' similarity index values of the features. It is also observed that the diagonal elements are indicated with value 1 (high similarity) due to features comparing with the same features. Figure 10 illustrates the similarity index between the features at the first stage of all fusion strategies implemented in this paper. An important observation from it is that the features in the early fusion strategy are more similar compared to the other strategies. The reason behind this is that all features are fused at the first stage itself. Such generated features will have the information of all four CT perfusion parameter maps, which are applied as the input to the network. In the other strategies (late fusion, bottleneck fusion, hierarchical fusion), all CTP maps are processed independently without fusing the information across the modalities, so the feature maps generated from each modality are different in nature. Hence, the similarity index among the features generated in the fusion strategies, except early fusion, is very low. In other words,

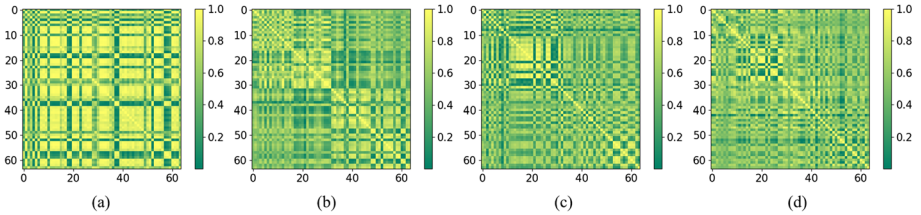

Fig. 10 Graphical representation of cosine similarity index between feature maps generated after the first stage in (a) Early Fusion (b) Late Fusion (c) Bottleneck Fusion (d) Hierarchical Fusion

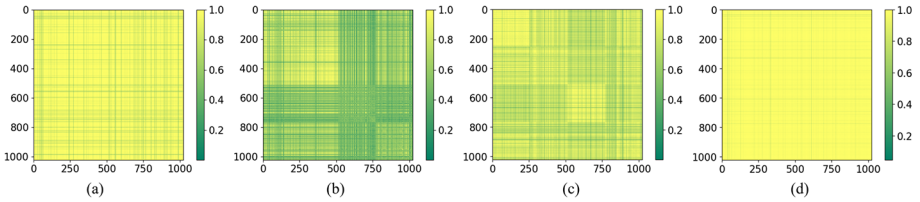

Fig. 11 Graphical representation of cosine similarity index between feature maps generated at bottleneck stage in (a) Early Fusion (b) Late Fusion (c) Bottleneck Fusion (d) Hierarchical Fusion

models that embedded fusion strategies other than early fusion extracted more dissimilar features which may help boost the performance of the model.

Figure 11 is conceptually similar to Fig. 10, but the sole difference is that feature maps are generated at the bottleneck. For all the fusion strategies, the number of feature maps at the bottleneck is the same, i.e. 1024. From Fig. 11(b), it can be observed that the feature maps generated in late fusion are more diverse in nature and show less similarity among them since feature maps are derived from each modality independently. These feature maps belong to four different groups derived from the four CT perfusion maps without any information exchange between these modalities. Upon observing the feature similarity index diagram of bottleneck fusion, Fig. 11(c), we can interpret that the feature maps are also different from each other but not as diverse as the features in the case of late fusion. One of the main reasons behind this is that the feature maps are fused just preceding the bottleneck layer. The feature maps in the hierarchical fusion strategy exhibit more similarity as compared to the other strategies.

By examining Figs. 10 and 11, we can interpret that feature maps in the initial layer are more diverse as compared with the feature maps at the deeper levels. The same may be inferred by observing the feature maps in Fig. 8(e), and these are more or less looking like similar features.

5.3 Effect of fusion strategy on small and large volume lesions

The interpretation of the scatter plot between the original and predicted volumes represented in Fig. 12 can be understood in the following way. The points on the diagonal (45^0 line) are predicted correctly. The points above the line represent the over-prediction of the lesions and below the line represent the underestimation of the lesions. For a better understanding and view, one more diagram for each fusion strategy has also been shown here. Figure 12(b),(d),(f), and (h) are zoomed and cropped versions of Fig. 12(a),(c),(e), and (g) respectively, to enhance the appearance of small volume lesion details.

For a better understanding of Fig. 12 the R-square values are also calculated. The achieved R-square value for early fusion is 0.903, Late fusion is 0.933, Bottleneck fusion is 0.919, and hierarchical fusion is 0.916. The late fusion has achieved the best R-square value followed by bottleneck fusion. A point to mention here is that the late fusion strategy has been achieved, which means that it is able to predict the lesion with almost similar volume as the reference but the overlap between the predicted and reference is not quite good in late fusion. This can be inferred by the dice and AVD metrics in Table 4. But, in the case of bottleneck fusion, it has more overlap between the predicted and reference volume, indicated by the dice value. Looking at Fig. 12(d), we can ascertain that late fusion underestimates the small lesions.

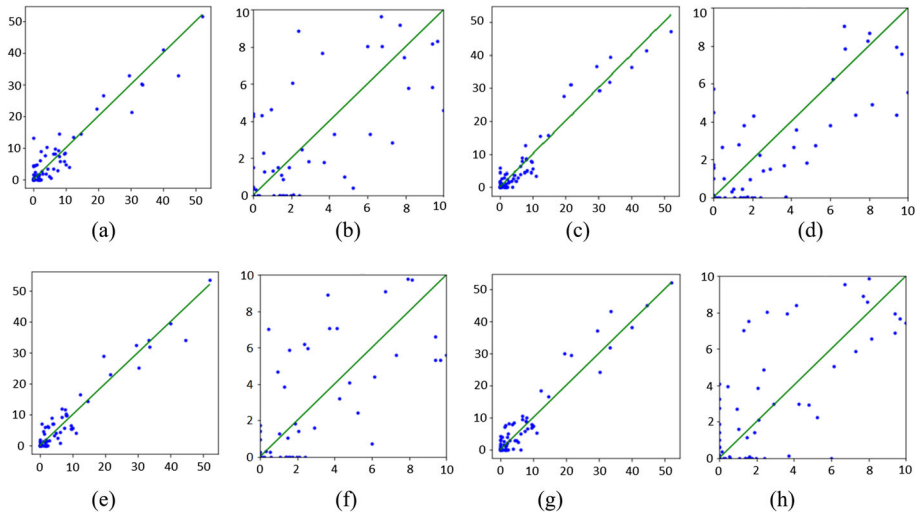

Fig. 12 The comparison of the original volume and the predicted volume on validation data. (a),(c),(e), and (g) represent the details of early fusion, late fusion, bottleneck fusion, and hierarchical fusion. (b),(d),(f), and (h) are zoomed and cropped versions of (a),(c),(e), and (g), respectively. The X-axis indicates the original volume (ml) and the y-axis indicates the predicted volume (ml)

Similarly, the hierarchical fusion overestimates the small lesions. For the large lesions, all fusion strategies behave more or less the same.

Figure 13 shows three examples (each row) of different-sized lesions and the predicted lesions of all fusion strategies are compared with the reference regions. The bottleneck

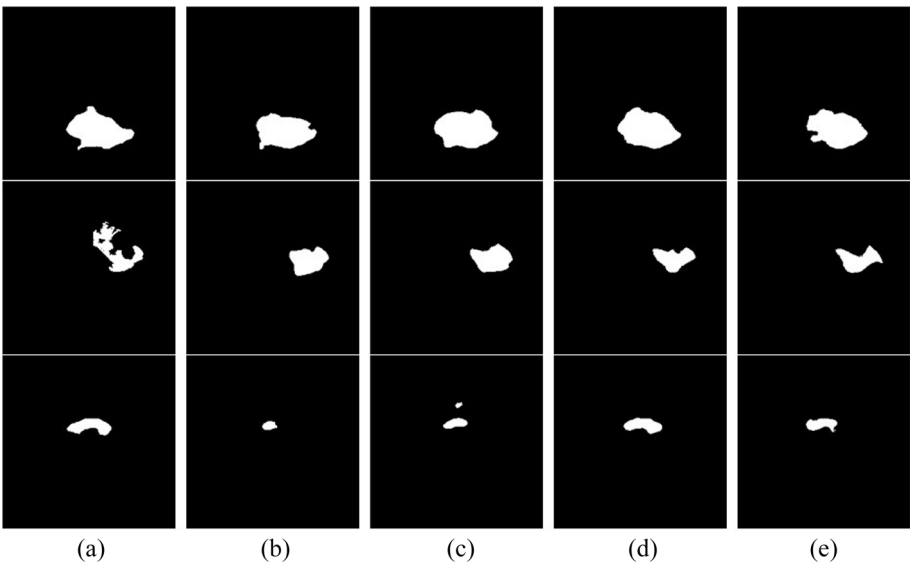

Fig. 13 The segmentation results of fusion strategies for three examples. From left to right (a) Ground truth (b) Early fusion (c) Late fusion (d) Bottleneck fusion (e) Hierarchical fusion

fusion strategy is able to predict better approximate segmentation results compared to the other fusion strategies.

Multiple modalities contain different information, particularly in this case, few modalities are responsive to the core where whereas other modalities to penumbra information. In most of the papers, the fusion at the input space (early fusion) has been employed. In the early fusion, the complementary information from the different modalities is integrated into the early layers while facilitating joint representative learning. This fusion model struggles to exploit the complementary information present in the modalities since they are fused in early stages and also cause cross-modal interference. In the late fusion, the information from all the modalities is well exploited using separate paths facilitating the well extraction of complementary information with minimum utilization of joint representative learning. The main problem with the late fusion is that learned features from the different modalities are independent and have no relation among them i.e., no cross-reference among the multimodal features. The bottleneck fusion overcomes the weaknesses of early and late fusion by properly integrating the complementary information present across the multimodalities. The complex relationship between the different modalities has been exploited by extracting the features from the modalities differently and combining them at a particular layer. This capacity to extract the relation among the different modalities makes the models more accurate than the fusion at the input space or the decision level. From the analysis, it is observed that how the multimodal data is CNN encoded the diversity of multimodal information presents the multiple modalities and effect of fusion on CNN learning. The above discussion presents the strengths and weaknesses of particular fusion strategies.

One of the primary reasons for the better results from bottleneck feature fusion is that this fusion strategy avoids the problem of cross-modal interference by introducing separate paths for feature extraction from the different modalities in the encoder. By implanting the bottleneck fusion strategy into the deep learning models, the models tend to learn the most complex relationship between the modalities. Moreover, each modality contributes different features and sometimes complementary information that is highly significant for better results. Fusing the features from all different modalities at bottleneck stages facilitates the efficient exploitation of the features. In bottleneck strategy, the deep learning model will have the flexibility to learn the appropriate scale for fusing the modalities together. It is beneficial to separate the information that would otherwise be combined by processing each modality separately in different networks. From the points mentioned above, it is evident that the fusion of multimodal features plays a major role in producing better segmentation results and bottleneck fusion achieved better results. Here we performed simple fusion operations such as concatenation and element-wise summation, but this can be extendable to the multimodal fusion through the attention-based module. This gives them more flexibility to capture the proper information from the multimodalities. The limitations and future work are as follows: Firstly, the primary limitations of our study may be the availability and size of the dataset used for training and evaluation. Limited access to large, diverse datasets can restrict the generalizability of our findings. These experiments are conducted only on this dataset. For more exploration of these fusion strategies, we need to experiment on the different datasets. Secondly, these models were developed to process the data in a 2D manner since the depth of the data varies from 2 to 22. Our future work includes the experimentation of these fusion strategies on the other datasets and provides more interpretability and explainability of the models on the diverse datasets.

6 Conclusion

In this paper, we reviewed and analyzed different fusion strategies i.e., early fusion, late fusion, bottleneck fusion, and hierarchical fusion, involved in the stroke lesion segmentation on CTP data. We analyzed the process of encoding the multimodal data, and the effect of the fusion strategy on CNN learning with the help of complementary information, joint representation learning, information integration, and feature representation. In addition, we also analyzed the deep features and feature similarity between the features to explore the encoding process of different deep learning models which involves various fusion strategies. After analyzing the results of the models, the bottleneck fusion strategy performed better. Implanting the early fusion in the deep learning model fails to exploit the complementary features among the modalities because the fusion is involved in the early stages of the network. The late fusion strategy fails to extract better features due to no cross-reference among the features from different modalities. The bottleneck fusion strategy balanced both the problems mentioned in early and late fusion. It is because of its capacity to establish the relation between the multimodal inputs. Bottleneck fusion also worked well for the small lesions as well. The bottleneck fusion strategy performed well in its raw form without any specialized modules which are responsible for the increase in performance, and the other fusion strategies also may work better if they are incorporated with the specialized module for improving performance. All the experiments were conducted on the ISLES 2018 dataset. The performance of the bottleneck fusion can be improved further by adding modules that are responsible for increasing the performance. In the future, we will apply these strategies to other databases and compare which strategy yields better results.

Funding This work was supported by the Researchers Supporting Project number (RSP2024R34), King Saud University, Riyadh, Saudi Arabia.

Data Availability The ISLES 2018 dataset, which was used in the experiments, is available at https://www.isles-challenge.org/ISLES2018/

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Abulnaga SM, Rubin J (2019) Ischemic stroke lesion segmentation in ct perfusion scans using pyramid pooling and focal loss. In: Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T (eds) Crimi A. Glioma multiple sclerosis stroke and traumatic brain injuries springer international publishing. Cham, Brainlesion, pp 352–363
2. Al Jowair H, Alsulaiman M, Muhammad G (2023) Multi parallel u-net encoder network for effective polyp image segmentation. Image Vis Comput 137:104767
3. Alshehri F, Muhammad G (2023) A few-shot learning-based ischemic stroke segmentation system using weighted mri fusion. Image Vis Comput 140:104865
4. Anand V.K, Khened M, Alex V, Krishnamurthi G, (2019) Fully automatic segmentation for ischemic stroke using ct perfusion maps In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T (Eds.) Brainlesion: Glioma Multiple Sclerosis stroke and traumatic brain injuries springer international publishing. Cham, pp 328–334

5. Bertels J, Robben D, Vandermeulen D, Suetens P (2019) Contra-lateral information CNN for core lesion segmentation based on native CTP in acute stroke, vol 11383. Springer International Publishing, LNCS. https://doi.org/10.1007/978-3-030-11723-8_26
6. Chen L, Bentley P, Rueckert D (2017) Fully automatic acute ischemic lesion segmentation in dwi using convolutional neural networks. Neuroimage Clin 15:633–643. https://doi.org/10.1016/j.nicl.2017.06.016
7. Chen Y, Chen J, Wei D, Li Y, Zheng Y (2020) Octopusnet: A deep learning segmentation network for multi-modal medical images In: Li Q, Leahy R, Dong B, Li X (Eds.) Multiscale multimodal medical imaging springer international publishing. Cham. pp 17–25
8. Clèrigues A, Valverde S, Bernal J, Freixenet J, Oliver A, Lladó X (2019) Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks. Comput Biol Med 115. https://doi.org/10.1016/j.compbiomed.2019.103487
9. Dolz J, Ben Ayed I, Desrosiers C (2019) Dense multi-path u-net for ischemic stroke lesion segmentation in multiple image modalities. In: Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T (eds) Crimi A. Glioma Multiple Sclerosis Stroke and Traumatic Brain Injuries Springer International Publishing. Cham, Brainlesion, pp 271–282
10. Ghnemat R, Khalil A, Abu Al-Haija Q (2023) Ischemic Stroke Lesion Segmentation Using Mutation Model and Generative Adversarial Network. Electronics (Switzerland) 12. https://doi.org/10.3390/electronics12030590
11. Huff DT, Weisman AJ, Jeraj R (2021) Interpretation and visualization techniques for deep learning models in medical imaging. Phys Med Biol 66. https://doi.org/10.1088/1361-6560/abcd17
12. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift In: Proc. Int. Conf. Mach. Learn, JMLR.org, pp 448-456
13. Islam M, Nooruddin S, Karray F, Muhammad G (2023) Internet of things: Device capabilities architectures protocols and smart applications in healthcare domain. IEEE Internet Things J 10(4):3611–3641
14. Islam M, Nooruddin S, Karray F, Muhammad G (2023) Multi-level feature fusion for multimodal human activity recognition in internet of healthcare things. Inf Fusion 94:17–31
15. Islam M, Vaidyanathan NR, Jose VJM, Ren H (2019) Ischemic stroke lesion segmentation using adversarial learning In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T (Eds.) Brainlesion: glioma multiple sclerosis stroke and traumatic brain injuries springer international publishing. Cham, pp 292–300
16. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B (2017) Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Med Image Anal 36:61–78. https://doi.org/10.1016/j.media.2016.10.004
17. Khurana S, Gourie-Devi M, Sharma S, Kushwaha S (2021) Burden of stroke in India during 1960 to 2018: a systematic review and meta-analysis of community based surveys. Neurol India 69:547
18. Kumar A, Ghosal P, Kundu SS, Mukherjee A, Nandi D (2022) A lightweight asymmetric u-net framework for acute ischemic stroke lesion segmentation in ct and ctp images. Computer Methods and Programs in Biomedicine 226. https://doi.org/10.1016/j.cmpb.2022.107157
19. Li J, Yu ZL, Gu Z, Liu H, Li Y (2019) Mman: Multi-modality aggregation network for brain segmentation from mr images. Neurocomputing 358:10–19. https://doi.org/10.1016/j.neucom.2019.05.025
20. Liu L, Yang S, Meng L, Li M, Wang J (2019) Multi-scale deep convolutional neural network for stroke lesions segmentation on ct images. In: Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T (eds) Crimi A. Glioma Multiple Sclerosis Stroke and Traumatic Brain Injuries Springer International Publishing. Cham, Brainlesion, pp 283–291
21. Liu R, Pu W, Zou Y, Jiang L, Ye Z (2022) Pool-unet: Ischemic stroke segmentation from ct perfusion scans using poolformer unet In: 2022 6th Asian conference on artificial intelligence technology (ACAIT) IEEE, pp 1–6. https://ieeexplore.ieee.org/document/10137834/. https://doi.org/10.1109/ACAIT56212.2022.10137834
22. Liu Z, Cao C, Ding S, Liu Z, Han T, Liu S (2018) Towards clinical diagnosis: Automated stroke lesion segmentation on multi-spectral mr image using convolutional neural network. IEEE Access 6:57006–57016
23. Lv W, Ashrafinia S, Ma J, Lu L, Rahmim A (2019) Multi-level multi-modality fusion radiomics: application to pet and ct imaging for prognostication of head and neck cancer. IEEE J Biomed Health Inform 24:2268–2277
24. Maier O, Menze BH, von der Gablentz J, Häni L, Heinrich MP, Liebrand M, Winzeck S, Basit A, Bentley P, Chen L, Christiaens D, Dutil F, Egger K, Feng C, Glocker B, Götz M, Haeck T, Halme HL, Havaei M, Iftekharuddin KM, Jodoin PM, Kamnitsas K, Kellner E, Korvenoja A, Larochelle H, Ledig C, Lee JH, Maes F, Mahmood Q, Maier-Hein KH, McKinley R, Muschelli J, Pal C, Pei L, Rangarajan JR, Reza SM, Robben D, Rueckert D, Salli E, Suetens P, Wang CW, Wilms M, Kirschke JS, Krämer UM, Münte TF, Schramm P, Wiest R, Handels H, Reyes M (2017) Isles 2015 - a public evaluation benchmark for

ischemic stroke lesion segmentation from multispectral mri. Med Image Anal 35:250–269. https://doi.org/10.1016/j.media.2016.07.009
25. Mair G, Wardlaw J (2014) Imaging of acute stroke prior to treatment: current practice and evolving techniques. Br J Radiol 87:20140216
26. Mittermeier A, Reidler P, Fabritius MP, Schachtner B, Wesp P, Ertl-Wagner B, Dietrich O, Ricke J, Kellert L, Tiedt S, Kunz WG, Ingrisch M (2022) End-to-end deep learning approach for perfusion data: A proof-of-concept study to classify core volume in stroke ct. Diagnostics 12. https://doi.org/10.3390/diagnostics12051142
27. Muhammad G, Alshehri F, Karray F et al (2021) A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. Inf Fusion 76:355–375
28. Omarov B, Tursynova A, Postolache O, Gamry K, Batyrbekov A, Aldeshov S, Azhibekova Z, Nurtas M, Aliyeva A, Shiyapov K (2022) Modified unet model for brain stroke lesion segmentation on computed tomography images. Comput Mater Contin 71:4701–4717 https://doi.org/10.32604/cmc.2022.020998
29. Pinheiro GR, Voltoline R, Bento M, Rittner L (2019) V-net and u-net for ischemic stroke lesion segmentation in a small dataset of perfusion data, vol 11383. Springer International Publishing, LNCS. https://doi.org/10.1007/978-3-030-11723-8_30
30. Raju CSP, Kirupakaran AM, Neelapu BC, Laskar RH (2022) Ischemic stroke lesion segmentation in ct perfusion images using u-net with group convolutions In: International Conference on Computer Vision and Image Processing. Springer, pp 276–288
31. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation In: Navab N, Hornegger J, Wells WM, Frangi AF (Eds.) Medical image computing and computer-assisted intervention – MICCAI 2015 Springer International Publishing. Cham, pp 234–241
32. de la Rosa E, Sima DM, Menze B, Kirschke JS, Robben D (2021) AIFNet: Automatic vascular function estimation for perfusion analysis using deep learning. Med Image Anal 74:102211. https://doi.org/10.1016/j.media.2021.102211
33. Shi T, Jiang H, Zheng B (2021) C2MA-Net: Cross-modal cross-attention network for acute ischemic stroke lesion segmentation based on CT perfusion scans. IEEE Trans Biomed Eng 69:108–118. https://doi.org/10.1109/TBME.2021.3087612
34. Soltanpour M, Greiner R, Boulanger P, Buck B (2019) Ischemic stroke lesion prediction in ct perfusion scans using multiple parallel u-nets following by a pixel-level classifier In: In Proc. Int. Conf. BIBE IEEE Computer Society. Los Alamitos CA USA. pp 957–963. https://doi.org/10.1109/BIBE.2019.00179
35. Soltanpour M, Greiner R, Boulanger P, Buck B (2021) Improvement of automatic ischemic stroke lesion segmentation in ct perfusion maps using a learned deep neural network. Comput Biol Med 137:104849. https://doi.org/10.1016/j.compbiomed.2021.104849
36. Song T (2019) Generative model-based ischemic stroke lesion segmentation. https://doi.org/10.48550/ARXIV.1906.02392
37. Song T, Huang N (2019) Integrated extractor generator and segmentor for ischemic stroke lesion segmentation. In: Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T (eds) Crimi A. Glioma multiple sclerosis stroke and traumatic brain injuries springer international publishing. Cham, Brainlesion, pp 310–318
38. Tureckova A, Rodríguez-Sánchez AJ (2019) Isles challenge: U-shaped convolution neural network with dilated convolution for 3d stroke lesion segmentation. In: Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T (eds) Crimi A. Glioma multiple sclerosis stroke and traumatic brain injuries springer international publishing. Cham, Brainlesion, pp 319–327
39. Vries LD, Emmer BJ, Majoie CB, Marquering HA, Gavves E (2023) Perfu-net: Baseline infarct estimation from ct perfusion source data for acute ischemic stroke. Med Image Anal 85. https://doi.org/10.1016/j.media.2023.102749
40. Wang G, Song T, Dong Q, Cui M, Huang N, Zhang S (2020) Automatic ischemic stroke lesion segmentation from computed tomography perfusion images by image synthesis and attention-based deep neural networks. Med Image Anal 65:101787. https://doi.org/10.1016/j.media.2020.101787
41. Wang Y, Katsaggelos AK, Wang X, Parrish TB (2016) A deep symmetry convnet for stroke lesion segmentation In: In Proc. Int. Conf. Image. Proc, pp 111–115. https://doi.org/10.1109/ICIP.2016.7532329
42. Yang HY (2019) Volumetric adversarial training for ischemic stroke lesion segmentation. In: Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T (eds) Crimi A. Glioma multiple sclerosis stroke and traumatic brain injuries springer international publishing. Cham, Brainlesion, pp 343–351
43. Zhang Y, Sidibé D, Morel O, Mériaudeau F (2021) Deep multimodal fusion for semantic image segmentation: A survey. Image Vision Comput 105. https://doi.org/10.1016/j.imavis.2020.104042
44. Zhou T, Ruan S, Canu S (2019) A review: Deep learning for medical image segmentation using multi-modality fusion. Array 3–4:100004. https://doi.org/10.1016/j.array.2019.100004

45. Zhou T, Ruan S, Guo Y, Canu S (2020) A multi-modality fusion network based on attention mechanism for brain tumor segmentation In: In Proc. IEEE Int. Symp. Biomed. Imaging, pp 377–380. https://doi.org/10.1109/ISBI45749.2020.9098392
46. Zhu H, Chen Y, Tang T, Ma G, Zhou J, Zhang J, Lu S, Wu F, Luo L, Liu S, Ju S, Shi H (2022) Isp-net: Fusing features to predict ischemic stroke infarct core on ct perfusion maps. Comput. Methods. Programs. Biomed 215:106630. https://doi.org/10.1016/j.cmpb.2022.106630

Authors and Affiliations

Chintha Sri Pothu Raju[1] · Bala Chakravarthy Neelapu[2] · Rabul Hussain Laskar [1] · Ghulam Muhammad[3]

Bala Chakravarthy Neelapu
neelapubc@nitrkl.ac.in

Rabul Hussain Laskar
rhlaskar@ece.nits.ac.in

1 Speech and Image Processing Lab, Department of ECE, National Institute of Technology Silchar, Silchar 788010, Assam, India

2 Department of Biotechnology and Medical Engineering, National Institute of Technology Rourkela, Rourkela 769008, Odisha, India

3 Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia