IEEE Journal of Translational Engineering in Health and Medicine

Received 5 January 2022; revised 5 March 2022; accepted 8 April 2022.
Date of publication 2 May 2022; date of current version 24 May 2022.

Digital Object Identifier 10.1109/JTEHM.2022.3172025

Multiple Sclerosis Lesions Segmentation Using Attention-Based CNNs in FLAIR Images

MEHDI SADEGHIBAKHI[1], HAMIDREZA POURREZA[1], AND HAMIDREZA MAHYAR[2]

[1]MV Laboratory, Department of Computer Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad 9177948974, Iran
[2]Faculty of Engineering, W Booth School of Engineering Practice and Technology, McMaster University, Hamilton, ON L8S 4L8, Canada

CORRESPONDING AUTHOR: HAMIDREZA POURREZA (hpourreza@um.ac.ir)

ABSTRACT **Objective:**Multiple Sclerosis (MS) is an autoimmune and demyelinating disease that leads to lesions in the central nervous system. This disease can be tracked and diagnosed using Magnetic Resonance Imaging (MRI). A multitude of multimodality automatic biomedical approaches are used to segment lesions that are not beneficial for patients in terms of cost, time, and usability. The authors of the present paper propose a method employing just one modality (FLAIR image) to segment MS lesions accurately. **Methods:**A patch-based Convolutional Neural Network (CNN) is designed, inspired by 3D-ResNet and spatial-channel attention module, to segment MS lesions. The proposed method consists of three stages: (1) the Contrast-Limited Adaptive Histogram Equalization (CLAHE) is applied to the original images and concatenated to the extracted edges to create 4D images; (2) the patches of size $80 \times 80 \times 80 \times 2$ are randomly selected from the 4D images; and (3) the extracted patches are passed into an attention-based CNN which is used to segment the lesions. Finally, the proposed method was compared to previous studies of the same dataset. **Results:**The current study evaluates the model with a test set of ISIB challenge data. Experimental results illustrate that the proposed approach significantly surpasses existing methods of Dice similarity and Absolute Volume Difference while the proposed method uses just one modality (FLAIR) to segment the lesions. **Conclusion:**The authors have introduced an automated approach to segment the lesions, which is based on, at most, two modalities as an input. The proposed architecture comprises convolution, deconvolution, and an SCA-VoxRes module as an attention module. The results show, that the proposed method outperforms well compared to other methods.

INDEX TERMS Medical image processing, multiple sclerosis, convolutional neural network, lesion segmentation, deep learning.

I. INTRODUCTION

Multiple sclerosis is an autoimmune, chronic, and demyelinating disease of axons that causes lesions in the brain's white matter (WM) tissues [1]. Varying from patient to patient, the most common symptoms of MS are weakness, balance issues, depression, fatigue, or visual impairment. As the most prominent visualization method for medical imaging nowadays, Magnetic Resonance Imaging (MRI) can visualize and diagnose this kind of disease [2]. Accurate segmentation of MS lesions in MR images is one of the most critical tasks in figuring out and describing the progression of the disease [3]. To do so, manual and automated segmentation methods are commonly employed to estimate and segment the total number of lesions and total lesion volume. Although manual segmentation is considered a desirable standard method [4], it poses challenges in describing 3-dimensional (3D) MRI information. This method is time-consuming, annoying, and prone to intra- and inter-observer variability. These challenges motivate Deep Learning (DL) and Machine Learning (ML) researchers to propose and develop a fast and accurate approach for the segmentation of MS lesions in MRI [5].

Supervised machine learning algorithms are one group of automated methods that can acquire knowledge from previously labeled training data and present high efficiency in MS lesion segmentation [6]. Generally, traditional supervised machine learning approaches are dependent on hand-crafted or low-level features. So far, plenty of supervised techniques for MS lesion segmentation have been proposed, such as decision random forests [7], [8], ensemble methods [9], non-local

means [10], k-nearest neighbors, [11], [12], and combined inference from patient and healthy populations [13]. Another group of automated methods is unsupervised, which extracts patterns from unlabeled data. Therefore, a significant number of unsupervised methods have also been introduced, which are based on thresholding methods with post-processing to remove False Positive (FP) and False Negative (FN) pixels [14], [15] or probabilistic models [16], [17].

In recent decades researchers have tended to use deep learning algorithms in biomedical image analysis such as brain tumors, brain tissue, diabetic retinopathy, and cardiac image segmentation.

Since 1988, deep learning methods, especially CNNs, have significantly increased performance in biomedical image analysis [18]. They require fewer manual features than standard supervised ML algorithms and can learn by themselves how to extract features directly from data during the training procedure [19]. Deep learning-based approaches provide state-of-the-art results for different problems such as computer vision semantic segmentation [20], as well as Natural Language Processing (NLP) [21]. They have also gained popularity in studying biomedical problems, such as cell classification [22], retinal blood vessel extraction [23], MS lesions [24], brain tumors [25], neuronal structures [26], and brain tissue segmentation [27]. For example, in [28], [29], [30], authors try to segment tumors in 3D MR images using a modified version of Decoder-Encoder networks. For brain tissue segmentation Wu *et al.* [31] proposed a dual encoder residual U-Net architecture to reduce the risk of losing local structure and necessary details. In [32], a Relation Transformer Block (RTB) and Global Transformer Block (GTB) are proposed to segment small diabetic retinopathy lesions accurately. Also, in [33] authors introduce a multi-modal few-shot Unsupervised Domain Adaptation (UDA) to detail cardiac Images.

A significant number of CNN-based algorithms for biomedical image segmentation have been proposed. These approaches can be divided into image-based and patch-based methods. Details of these approaches are discussed in the following paragraph.

Image-based methods, extract features from the whole image as global structure information [34], [35] and are categorized into 2D-based [34], [36] and 3D-based [35], [37] segmentation.

In 3D-based methods, features are first extracted from an original 3D image by employing CNNs through 3D filters. The model then segments each pixel or voxel to the lesion or non-lesion. The high chances of overfitting are one of the disadvantages of the 3D-based method, which usually fits many parameters when the dataset is small. This is a common occurrence in biomedical applications [35], [37].

In 2D-based methods, the 3D image is first divided into 2D slices, and then each slice is eventually segmented. Finally, to reconstruct the 3D prediction, all 2D predictions are concatenated together. Compared to 3D-based methods, 2D-based approaches are not as accurate due to missing part of the contextual information. However, they have fewer parameters for each layer, lowering the risk of overfitting in small datasets [36].

Patch-based methods use two different strategies for lesion segmentation. Utilizing a moving window, the first strategy creates a local representation for each pixel/voxel. Then, a CNN is trained to use all the extracted patches to identify each patch's central pixel/voxel as either a lesion or a non-lesion. Considering the repetitive computations on the overlapping features of the sliding window, it is worth mentioning that there is a long training time. Faster than the previous strategy, the second approach randomly extracts patches from all over the 3D image. The extracted patches are then used to train the CNN-based model. Ultimately, the trained model segments each pixel/voxel of the patch as either a lesion or a non-lesion [38].

The rest of the current paper is organized into five sections and a reference section. Section II reviews works related to the proposed method. Section III explains the dataset used to train and evaluate the proposed method. Section IV clarifies the methodology. Section V explains the evaluation metrics and compares the results of the test dataset against the findings of other methods. Finally, a discussion and summary are provided in Section VI.

A. MOTIVATION

For automatic MS lesion segmentation, the present paper proposes a new method based on deep learning which consists of two 3D convolutional network branches. The current study randomly extracts some large 3D patches to prevent overfitting due to the small number of data, and the lack of global structure information. Then a deep neural network is designed inspired by the 3D version of ResNet [39], which consists of convolution and deconvolution layers, channel-based attention, and special-based attention.

The designed network can be employed in each path and each branch is assigned to a particular MRI modality so that the maximum benefits of each modality can be individually utilized. As a result, the present study has introduced a network that can have at most two different modalities (FLAIR and T1) However, for the first time, stunning results were achieved with just one modality (FLAR) when compared to other methods. In summary, the main contributions in this paper are:

- Introduction of a 3D patch-based method to prevent overfitting and a lack of global structure information and to combine the two modalities in order to take advantage of brain tissue information.
- Proposal of a deep learning-based approach with one modality (FLAIR) to segment the MS lesion which, for the first time, due to, the difficulty of providing variant modalities in terms of cost and time in the clinical situation.
- Demonstration of the top dice coefficient on the ISBI dataset using two modalities compared with other two-modality methods.

IEEE Journal of Translational Engineering in Health and Medicine

- Design of a 4D channel-wise and spatial attention module inspired by channel-based [40] attention and special-based attention [41].

B. RELATED WORKS

As already stated, a wide range of methods for MS lesion segmentation has been proposed. Recently, convolution neural network strategies have reported outstanding performance in medical image processing, especially in MS lesion segmentation. Hence, this discussion of related works is related to CNN-based methods.

It should be noted that some approaches are patch-based. A useful illustration is a study by, Ghafoorian *et al.* [42] in 2015 which proposed a 2D CNN-based model that, increases the number of training samples and avoids overfitting during the training process. Similarly, Birenbaum *et al.* introduced multiple 2D patch-based CNNs that use in parallel the benefits of the common information within longitudinal data [43]. Roy *et al.* developed a two-path CNN based on a 2D-patch which employs different MRI modalities in each path as an input and finally concatenates the output of each path to create the final prediction [38]. Afzal *et al.* proposed a system that includes two 2D patch-wise CNNs which accurately segment lesions. Their first CNN network creates a lesion map while the second enhances efficiency by reducing the number of false positives [44]. Shachor *et al.* introduced a multi-view structure based on 2D patches. Each view of the 2D patches is passed to the model as an input and the patches output is fused to create the final prediction [45].

In addition, there are some approaches to segment lesions based on 3D patches. Vaidya *et al.* presented a 3D patch-based CNN for segmenting lesions, after which a WM mask is applied to the output prediction to reduce the FP rate and attain high performance [46]. Valverde *et al.* [24] proposed a cascaded 3D CNN approach whose first model is trained with extracted 3D patches and its second model is then used to reduce the FP of the first model. In addition, Valverde *et al.* developed a model to analyze the effect of intensity domain adaptation on CNN-based models [47]. Andermatt *et al.* introduced a method based on multi-dimensional gated recurrent units and used 3D patches to train the model [48]. Salem *et al.* [49] presented a CNN-based model to create synthesis lesions in MR images as a way to deal with one of the biggest challenges in medical image processing, that is, a small number of data. They reported acceptable results even though only one image is used as a dataset to train Valverde's proposed model [24]. Hashemi *et al.* [50] proposed a 3D patch-based CNN method that employs the idea of a densely connected network. They also introduced a new loss function to deal with imbalanced data.

Furthermore, some approaches have used the whole image as an input. Brosch *et al.* proposed a whole brain-based segmentation method utilizing 3D CNN which takes advantage of some shortcut connections between layers so as to extract the low- and high-level features from the shallowest to the deepest layers. By doing so, the model learns information and features about the locations and structure of MS lesions [35]. Kang *et al.* proposed an attention context U-NET based on 3D images [37]. Aslani *et al.* designed a deep 2D encoder-decoder CNN for the segmentation of MS lesions [51]. Another paper of theirs introduced, a method based on a 2D CNN for slice-by-slice segmentation of lesions in 3D data. Lesions are separately segmented on each slice, and then each segmented slice is concatenated to create a 3D lesion mask [36].

Although all of the proposed patch-based techniques perform well in terms of segmentation, they all lack global structural details. Simply put, the segmentation process does not take into account the brain's overall structure or the exact location of lesions. In contrast, whole brain-based segmentation methods need a large number of data to train the model, which is regrettably a commonplace requirement in biomedical applications that is difficult to meet.

II. METHODS AND PROCEDURES

A. MATERIAL

To evaluate and compare the proposed method's performance to that of other state-of-the-art approaches, the present paper employs the ISBI[1] 2015 Longitudinal MS Lesion Segmentation Challenge dataset, which is publicly available on the challenge website. Further details are provided in the upcoming section.

1) ISBI 2015 LONGITUDINAL MS LESION SEGMENTATION CHALLENGE:

The ISBI dataset includes 19 subjects which are divided into a test set with 14 subjects and training set with five subjects. For each subject, there are varieties of time-points in the range of 4 to 6, for each of which T1-w, T2-w, PDW, and FLAIR image modalities are prepared. The size of each image is $182 \times 256 \times 182$ and the voxel resolution is one millimeter. Two different raters (R1 and R2) manually segment the images, so the data set has two ground truth lesion masks. The ground truth is publicly available for training images, but not for test images. However, the proposed method's performance over the test set is evaluated by submitting binary masks to the challenge website[2] [52].

B. METHOD

The process of constructing the model consists of three stages: first, the preprocessing data, second, the patch extraction from images, and last, lesion prediction using the trained model. The following discusses all processes of the proposed method in detail.

1) DATA PREPROCESSING

The present study utilizes the preprocessed version of the images available on the challenge website. Preprocessing algorithms that are already applied to images of the dataset

[1] International Symposium on Biomedical Imaging.
[2] www.smart-stats-tools.org

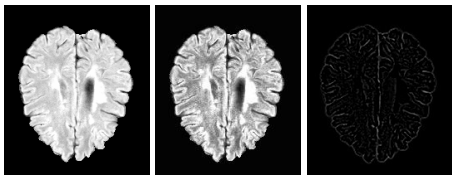

FIGURE 1. Impact of a preprocessing algorithm on the fourth screening of the sixth training sample. From left to right, a slice of the 3D original image, the preprocessed version of the image, and the extracted edge by the Laplacian filters.

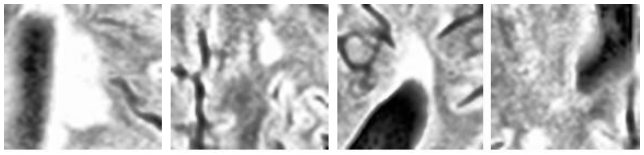

FIGURE 2. Samples of extracted patches from the dataset images. The first and third patches are centered on the lesion voxel from left to right, while the others are centered on the non-lesion voxel.

are skull-stripped by the Brain Extraction Tool (BET) [53] and N3 intensity normalization [54] and rigidly registered to the MNIICBM152 template [55].

The current work applies more preprocessing algorithms to enhance the local contrast of the images and to avoid the distorting differences in the ranges of values. Contrast-Limited Adaptive Histogram Equalization (CLAHE) [56] is applied to achieve this objective. Then, the edges extracted by the Laplacian detectors (To help the model for extracting higher-level features and also find a boundary of the white matter tissue in the T1 image) are concatenated to enhance the image, and the 4D data is created. Finally, before passing the data into the network, the intensities of each image are normalized with a zero mean and unit variance. Figure 1 presents an illustrated example of one slice of an extracted edge and the enhanced image.

2) PATCH EXTRACTION

Extraction of patches from the images begins after data preparation. The input of the model is a bunch of patches of images. Approximately 60% of the selected patches, $80 \times 80 \times 80 \times 2$ in size, are centralized on the lesion voxel. The rest of the patches are centered on the non-lesion voxel. Figure 2 depicts some of the extracted patches.

3) NETWORK ARCHITECTURE

This section presents the proposed architecture outline. Inspired by the spatial attention [41] and channel attention [40] strategies, the present study integrates an adapted version of these two attentions into 3D ResNet architecture to capture a better contextual image representation. Subsequently, decision-level fusion [57] is employed to learn the complementary information independently from the different modalities.

a: Base model

Inspired by the 3D ResNet, the architecture of a base model for MS lesion segmentation is illustrated in

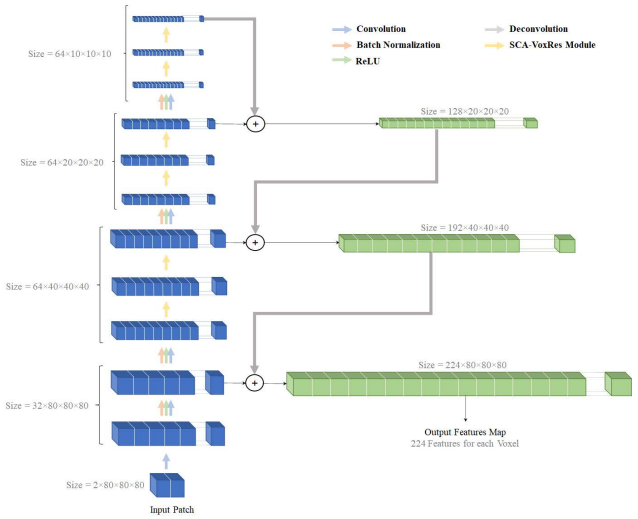

FIGURE 3. Overview of the proposed Base Model (CNN-based architecture to extract features for lesion segmentation) as described in Section II. Data is presented in input by 4D patches (The first dimension is the number of input channels, and then are the depth, height, and width, respectively), and the model generates the feature maps with 224 features for each voxel input patch. Finally, the output of the Base Model is used as input to the segmentation layer, which is responsible for the segmentation of each voxel to lesion or non-lesion.

Figure 3. The model consists of convolutional, deconvolutional, batch normalization (BN), and rectified linear unit (ReLU) layers, as well as six stacked residual modules (i.e., SCA-VoxRes modules) with a total of 25 volumetric convolutional/deconvolutional layers. As shown in Figure 4, each SCA-VoxRes module includes two convolutions, two BN/ReLU layers, and an SCA module. In this module, the transformed feature and input feature are added together by the skip connection. This connection can propagate information directly to the forward and backward passes. In addition, the SCA module includes spatial and channel-wise attention, which will be explained in upcoming sections.

It should be pointed out that filters and operations are implemented in a 3D shape to learn and extract a more robust volumetric feature representation [67]. Due to their computation efficiency and representation capability, the small kernels (i.e., $3 \times 3 \times 3$) are employed in the convolutional layers. To reduce the resolution of the input image and features, three convolutional layers are employed with a stride of 2. As a result, a sizeable receptive field network is obtained to extract more contextual information to improve discrimination capability. Four BN layers are inserted into the network to overcome the internal covariance shift in the training process and improve network performance. This network uses rectified linear units as the activation function for nonlinear transformation. Lastly, the extracted features in the 3rd, 5th, 9th, and 15th layers are deconvolved and concatenated to use as an input for the segmentation step.

b: Spatial and channel-wise attention

Diagnosis of the brain's white matter lesions is very challenging due to similar pixels/voxels in brain tissue and the

IEEE Journal of Translational Engineering in Health and Medicine

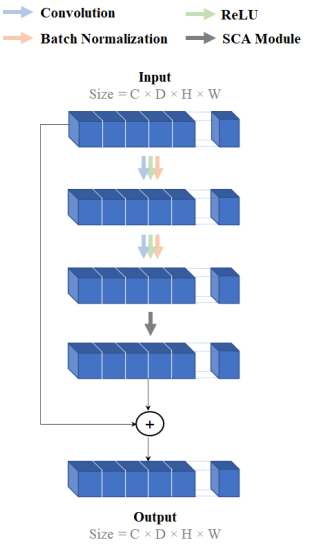

FIGURE 4. SCA-VoxRes Module. The input should be in the size of (C, D, H, W), where C is the number of channels, D is the depth, and H and W are height and width, respectively.

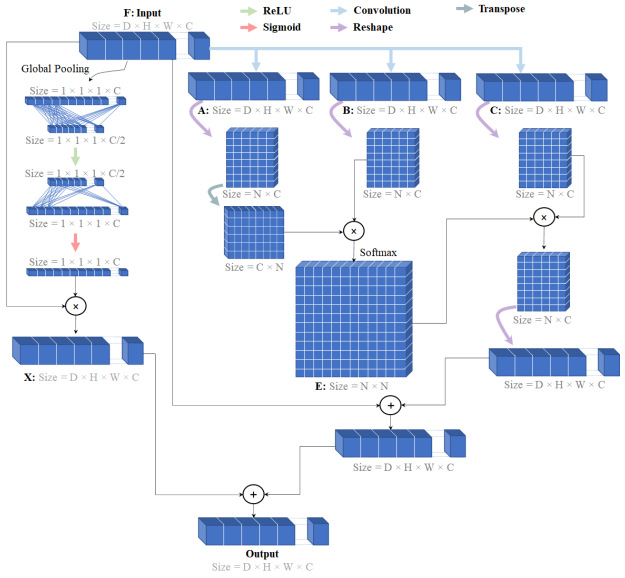

FIGURE 5. Spatial Channel Attention (SCA) module based on 4D input. D, H, W, C are depth, height, width, and channel. As well, as N is equal to $D \times H \times W \times$.

resulting intensity. Furthermore, in CNN-based approaches, convolution operators represent a local receptive field. Consequently, the features extracted from pixels with the same label may somewhat differ, and these differences cause intra-class inconsistency and affect the accuracy of the model. For this reason, an attention module is developed, as illustrated in Figure 5, which is based on a combination of spatial and channel-wise attention for exploring global contextual information.

i) Channel-wise attention

The objective of channel-wise attention is to enhance the network feature representation capability and emphasize interdependent feature map-specific semantics by exploiting the interdependencies between the channels of the feature maps. The structure of channel-wise attention is illustrated in Figure 5. First, the channel-wise attention map, $X \in \mathbb{R}^{1\times1\times1\times C}$, is directly calculated by an average pooling and two fully connected layers from the original features, $F \in \mathbb{R}^{D\times H\times W\times C}$. Then, the channel-wise attention map is multiplied by the input feature to scale each channel of the feature map.

ii) Spatial attention

The extraction of discriminatory representation is essential for the segmentation of WM lesions, and this can be achieved by capturing long-range contextual information. Thus, a spatial attention mechanism can encode a global representation with a broader field of view into local features. As illustrated in Figure.5, the input feature map, $F \in \mathbb{R}^{D\times H\times W\times C}$, is fed into the three one-by-one convolution layers to generate three new feature maps, called A, B, C, while they are the same size, $\mathbb{R}^{D\times H\times W\times C}$ Next, A and B are reshaped to $\mathbb{R}^{N\times C}$, where $N = D \times H \times W$ is the number of voxels. Later, the spatial attention map, $E \in \mathbb{R}^{N\times N}$, is calculated by the matrix multiplication of B and the transpose of A, followed by applying a SoftMax function to the result. In the next step, the feature map, $C \in \mathbb{R}^{D\times H\times W\times C}$, is reshaped to the $\mathbb{R}^{N\times C}$ matrix. Then, matrix multiplication is performed between $C \in \mathbb{R}^{N\times C}$, and $E \in \mathbb{R}^{N\times N}$, and this reshapes the result to $\mathbb{R}^{D\times H\times W\times C}$. Finally, the output of spatial attention is calculated by Equation 1.

$$O_j = \omega \sum_{i=1}^{n} \left(E_{ij} C_i\right) + F_j \tag{1}$$

where ω refers to a learnable parameter. The output of spatial attention can be obtained from Equation 1, as each voxel is a weighted sum of the features.

4) TWO-PATH ARCHITECTURE

In medical image processing applications, datasets provide different imaging modalities for analyzing various tissue structures robustly. For example, in the ISBI dataset, four modalities are available such as T1w, T2w, PDW, and FLAIR. The most significant reason for providing multi-modality images is because the information of the modalities complements each other. Therefore, utilizing the basic model, the current study designed a decision-level fusion architecture with two individual paths. Figure 6 presents this architecture. First, two modalities, T1w and FLAIR, are individually processed, and two different feature maps are extracted. Second, the extracted feature maps from both paths are concatenated together and then used as an input for classification. This strategy allows for individual feature learning for each modality before aggregating the feature maps.

C. LOSS FUNCTION

One of the significant problems in medical image processing algorithms is dealing with unbalanced data. Since 3D images

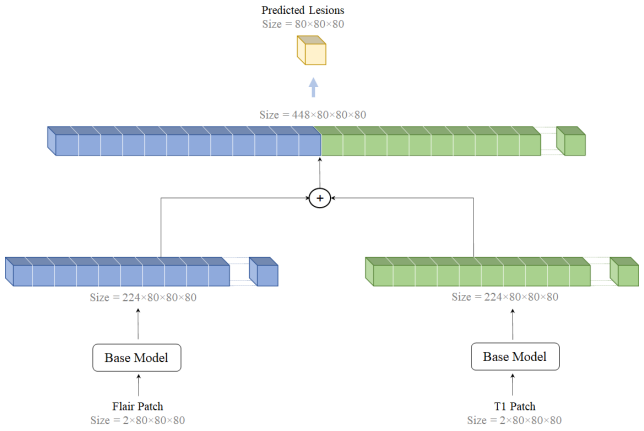

FIGURE 6. The architecture of the Two-path Network integrates the feature maps, which the Base Model prepares, and segments them using a one-by-one convolution layer.

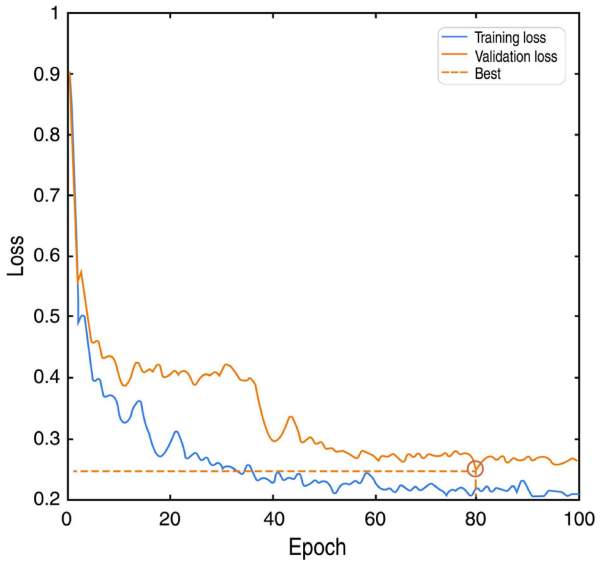

FIGURE 7. The decrease of loss function during the training step on the ISBI dataset. Due to overfitting, the best model is selected for the test step based on the validation set loss. According to the plot, the best model is at epoch 80.

have fewer lesion voxels than non-lesion voxels, an imbalance in data is faced. Therefore, if the used loss function cannot deal with this problem, the model converges to the minority class. In other words, samples in a class with a small number (minority class) cannot be classified accurately by using the model. The proposed method uses a combination of Tversky [50] and Focal [58] loss functions to handle the imbalanced data. As illustrated in Equation 2, the Tversky loss function allows the assignment of different weights to False Negative (FN) and False Positive (FP) to improve the recall rate.

$$TverskyLoss = \sum_{c} 1 - \frac{\sum_{i=1}^{N} p_{ic} g_{ic} + \varepsilon}{\sum_{i=1}^{N} p_{ic} g_{ic} + \alpha \sum_{i=1}^{N} p_{ic} - g_{ic} + \beta \sum_{i=1}^{N} p_{i\bar{c}} g_{ic} + \varepsilon} \quad (2)$$

where p_{ic} is the probability that pixel i is of lesion class c and $p_{i\bar{c}}$ is the probability that pixel i is of non-lesion class g_{ic}, and the same can be said for g_{ic} and $g_{i\bar{c}}$ which are related to the grand truth. In addition, hyperparameters α and β can be tuned by assigning a number in the [0, 1] range.

The Tversky loss function limitation is the low convergence speed due to the segmentation of small ROIs. This does not contribute significantly to losses. However, to overcome this problem, the current work utilizes the Focal Tversky loss function, which has a parameter, γ, to control the segmentation of small ROIs. Equation 3 defines the Focal Tversky Loss.

D. TRAIN AND TEST DETAILS

At the beginning of the training procedure, the data should be split to make the train, test, and validation dataset. The ISBI dataset includes two sets of images. Consisting of 21 images from five subjects, the first set of images with available ground truth is randomly divided into validation and training sets. Sixty images are considered as a train set and five images as a validation set. In addition, the second set of images is used as a test set.

One of the main challenges for gradient-based optimization methods is choosing an optimal starting point for the learning rate. Classical methods employ a fixed number for the learning rate in all stages of training. However, adjusting the learning rate during the training procedure is better by reducing it according to a predefined schedule function. In this paper, exponential decay is used as the schedule function. In our experiments, the Adam [59] optimizer outperforms performed better than other optimizers, such as AdaGrad [60], AdaDelta [61], and RMSprop [62]. So, Adam is selected as an optimizer.

The input image is divided into 80 × 80 × 80 image patches to test the model. The model predicts the label for each part, and, in the end, all predictions are integrated as the label of the given image.

E. IMPLEMENTATION DETAILS

The proposed method is implemented in Python with the Pytorch framework. The experiments are performed on Google Colaboratory[3] with 12 GB RAM. The two-path network is trained end-to-end and, to do so, 4D patches are randomly extracted from the 4D data as described in Section II. Then, the Focal Tversky loss function deals with the imbalanced data problem, as explained in Section II, with $\alpha = 0.7$, $\beta = 0.3$, $\gamma = 4/3$.

In addition, to optimize the network parameters and find the best model, the current work utilizes validation data and

[3]colab.research.google.com

IEEE Journal of Translational Engineering in Health and Medicine

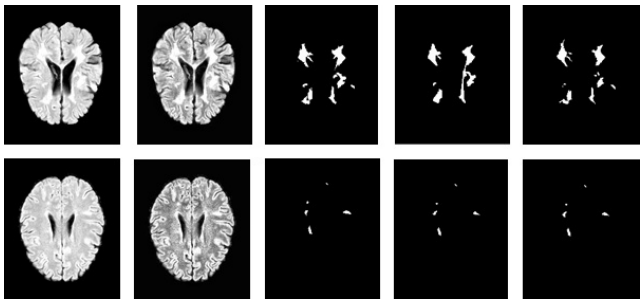

FIGURE 8. Output segmentation by the present study on two ISBI training dataset subjects (the first row is Subject 2 and the second row is Subject 3) compared to the ground truth provided by Rater 1 and Rater 2. From left to right, the first column is the original FLAIR image, the second column is after the preprocessing method, the third column is the segmentation result of the proposed method, and the following two columns are the ground truth provided by Rater 1 and Rater 2.

saves the model with the best validation data results. The model is trained in 100 epochs with the Adam optimizer at an initial learning rate of 0.0001. The exponential decay function reduces the learning rate in each epoch. According to Figure 7, the best performance obtained is at epoch 80. The training computation time of 100 epochs is approximately 12 hours.

III. EXPERIMENTAL SETUPS AND RESULT

The ISBI 2015 dataset evaluates the proposed network, and a direct comparison is made with plenty of state-of-the-art approaches. The main reason for choosing this dataset is the challenge posed by lesions regarding size, shape, and location. Therefore, future sections shall discuss evaluation criteria and outcomes.

A. EVALUATION METRICS

Generally, with the usage of the model, the metrics between the ground truth and the predicted image measure the evaluation metrics of segmentation methods. Hence, the following describes the evaluation metrics employed by the current research. With the assumption that $\mathscr{M}_A$ represents the ground truth created by one of the experts and $\mathscr{M}_R$ is the output generated by the model, each of the evaluation metrics is defined as [52]:

Dice Similarity Coefficient (DSC): Dice is a commonly used volume metric for measuring the similarity between the ground truth and the model's output.

$$DSC\left(\mathscr{M}_R, \mathscr{M}_A\right) = \frac{2\left|\mathscr{M}_A \cap \mathscr{M}_R\right|}{\left|\mathscr{M}_A\right| + \left|\mathscr{M}_R\right|} \tag{4}$$

Lesion True Positive Rate (LTPR): With the assumption of the list of lesions, $\mathscr{L}_R$, as the 18-connected components of $\mathscr{M}_R$ and a similar definition for $\mathscr{L}_A$, the lesion-wise ratio of true positives, LTPR, is defined as:

$$LTPR\left(\mathscr{M}_R, \mathscr{M}_A\right) = \frac{\left|\mathscr{L}_R \cap \mathscr{L}_A\right|}{\left|\mathscr{L}_R \cap \mathscr{L}_A\right| + \left|\mathscr{L}_R \cap \mathscr{L}_A^C\right|} \tag{5}$$

where $\mathscr{L}_A^C$ is the complement of $\mathscr{L}_A$.

Lesion False Positive Rate (LFPR): LFPR is the lesion-wise ratio of false positives, which is given as:

$$LFPR\left(\mathscr{M}_R, \mathscr{M}_A\right) = \frac{\left|\mathscr{L}_R^C \cap \mathscr{L}_A\right|}{\left|\mathscr{L}_R^C \cap \mathscr{L}_A\right| + \left|\mathscr{L}_R^C \cap \mathscr{L}_A\right|} \tag{6}$$

Absolute Volume Difference (AVD): The total volume of the absolute difference is divided by the total volume of the ground truth.

$$AVD\left(\mathscr{M}_R, \mathscr{M}_A\right) = \frac{\max\left(\left|\mathscr{M}_R\right|, \left|\mathscr{M}_A\right|\right) - \min\left(\left|\mathscr{M}_R\right|, \left|\mathscr{M}_A\right|\right)}{\mathscr{M}_R} \tag{7}$$

Although providing more information (anatomical and tissue-based features) for learning-based methods can lead to getting effective and accurate learning, the goal of this paper is not only to achieve high accuracy but also to provide a system with a minimum of modality and at the same time high accuracy. Therefore, the number of input modalities acts as one of the evaluation metrics, because, as the number of input modalities lowers, the method becomes more beneficial for patients in terms of cost, time, and usability.

B. RESULTS

The proposed method's efficiency on the ISBI dataset is evaluated by a process that is carried out in two stages.

In the first stage, the evaluation is with training data, in which the ground truth of the images is available. Table 1 provides the results of the comparison to those of other methods. As seen in the table, the current study's approach outperforms other methods in terms of DSC and LTPR. For images of high and low lesion loads, Figure 8 compares the present paper's segmentation results to ground truths.

In the second stage, ground truths are not available in the ISBI test set (with 14 subjects) used to evaluate the proposed method and the evaluation metrics are calculated utilizing the challenge web service. The current work trains its model with the training data's four subjects, one of which serves as the validation data. Then, the segmentation of the test set is predicted and, finally, the 3D segmentation results are submitted to the challenge web service for evaluation. Table 2 presents the results of the ISBI test set and compares these to those of other published papers. Clearly, in some of the evaluation

$$FocalTverskyLoss = \sum_c 1 - \left(\frac{\sum_{i=1}^N p_{ic} g_{ic} + \varepsilon}{\sum_{i=1}^N p_{ic} g_{ic} + \alpha \sum_{i=1}^N p_{i\bar{c}} g_{ic} + \beta \sum_{i=1}^N p_{ic} g_{i\bar{c}} + \varepsilon}\right)^{\gamma} \tag{3}$$

TABLE 1. The first ISBI data results of the proposed model in comparison to those of other models. In this experiment, the ISBI dataset includes images with available ground truth. The mean values of DSC, LTPR, and LFPR for different methods are shown.

Method	R1			R2		
	Dice	LTPR	LFPR	Dice	LTPR	LFPR
R1	-	-	-	0.7320	0.6550	0.1740
R2	0.7320	0.8260	0.3550	-	-	-
DCEN. [35] (Trained by R1)	0.6844	0.7455	0.5455	0.6444	0.6333	0.5288
DCEN [35] (Trained by R2)	0.6833	0.7833	0.6455	0.6588	0.6933	0.6199
DED CNN [51] (Trained by R1)	0.6980	0.7460	0.4820	0.6510	0.6410	0.4506
DED CNN [51] (Trained by R2)	0.6940	07840	0.4970	0.6640	0.6950	0.4420
Multi-Branch CNN [36] (Trained by R1)	0.7649	0.6697	**0.1202**	0.6989	0.5356	**0.1227**
Multi-Branch CNN [36] (Trained by R2)	0.7646	0.7002	0.2022	0.7128	0.5723	0.1896
Proposed-Method (Two Modality)	**0.7982**	0.8013	0.3676	**0.7978**	0.7295	0.2628
Proposed-Method (One Modalities)	0.7865	**0.8017**	0.3923	0.7856	**0.7298**	0.2945

TABLE 2. The results of the proposed method on the official ISBI test set when compared to the results of other methods. The metrics with the best and second-best performances are indicated by bold and underlined values, respectively.

Method	Modalities	Dice	LTPR	LFPR	AVD
Asmsl [48]	4 (T1, T2, FLAIR, PD)	0.6298	**0.4871**	0.2013	0.4045
ACU-NET [37]	4 (T1, T2, FLAIR, PD)	<u>0.6345</u>	<u>0.4787</u>	0.1299	0.3949
Multi-View CNN [43]	4 (T1, T2, FLAIR, PD)	0.6271	0.5678	0.4975	0.3585
IMAGINE [50]	4 (T1, T2, FLAIR, PD)	0.5841	0.4558	**0.0866**	0.4972
Cascaded CNN [24]	3 (T1, T2, FLAIR)	0.6304	0.3669	0.1529	**0.3384**
Multi-Branch CNN [36]	3 (T1, T2, FLAIR)	0.6114	0.4103	0.1393	0.4537
DED CNN [51]	3 (T1, T2, FLAIR)	0.4864	0.3034	0.1708	0.4768
SDA U-NET [49]	3(T1, T2, FLAIR)	0.6305	0.3670	0.1529	0.3585
FLEXCONN [38]	2 (T1, FLAIR)	0.5243	----	<u>0.1103</u>	0.5207
One-shot [47]	2 (T1, FLAIR)	0.5774	0.2967	0.1885	0.3848
Proposed Method	<u>2 (T1, FLAIR)</u>	**0.6430**	0.4543	0.3524	<u>0.3524</u>
Proposed Method	**1 (FLAIR)**	0.6321	0.4547	0.3868	0.3880

metrics, the present study's results for the two modalities are superior to those of other studies. Even in the single modality (the FLAIR image), the current paper's results are satisfactory when compared to the findings of other approaches.

As mentioned in Section II.A, there are two different ground truths for each training sample and this difference indicates the challenge of accurately labeling lesion areas. In the proposed method, the knowledge of both experts is employed to train the model. As seen in Table 2, in comparison to other studies, the proposed method has a high LFPR. However, the visualization of the results shows that most FP pixels are in the connected neighborhood of TP pixels. In other words, the algorithm is unlikely to predict non-lesion pixels as lesions unless these pixels are connected to a lesion area, thus producing a slight increase in the lesion area. As shown in Figure 9, the created false positive pixels are all connected in the vicinity of actual positive pixels, which slightly expands the area of the lesion.

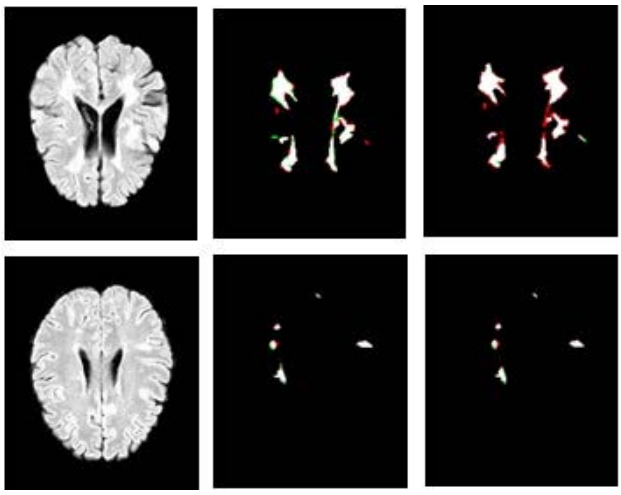

FIGURE 9. Output segmentation of the introduced model compared to two ground truths. Left to right are the FLAIR image, the results compared to the first expert, and the results compared to the second expert. In all images, true positives are denoted in white pixels, false positives in red pixels, and false negatives in green pixels.

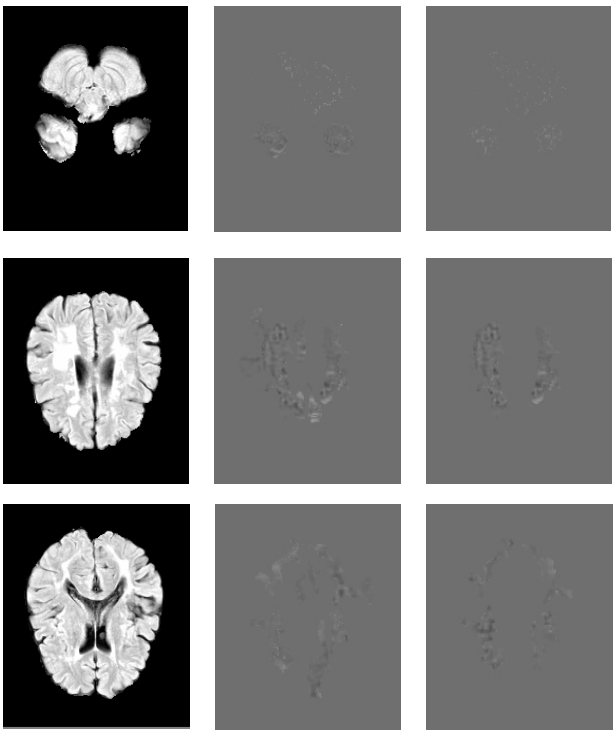

FIGURE 10. Attention maps generated by M3d-CAM. From left to right, the first column original Flair image, the second column is the attention map for the model without an SCA module, and the third column is the attention map for the model with an SCA module.

Also, to show the efficiency of the SCA module, we use Explainable Artificial Intelligence (XAI) to visualize the features which are understandable for humans and experts [63]. Several XAI models for 2D or 3D segmentation and classification have been proposed until now [64] [65]. This

IEEE Journal of Translational Engineering in Health and Medicine

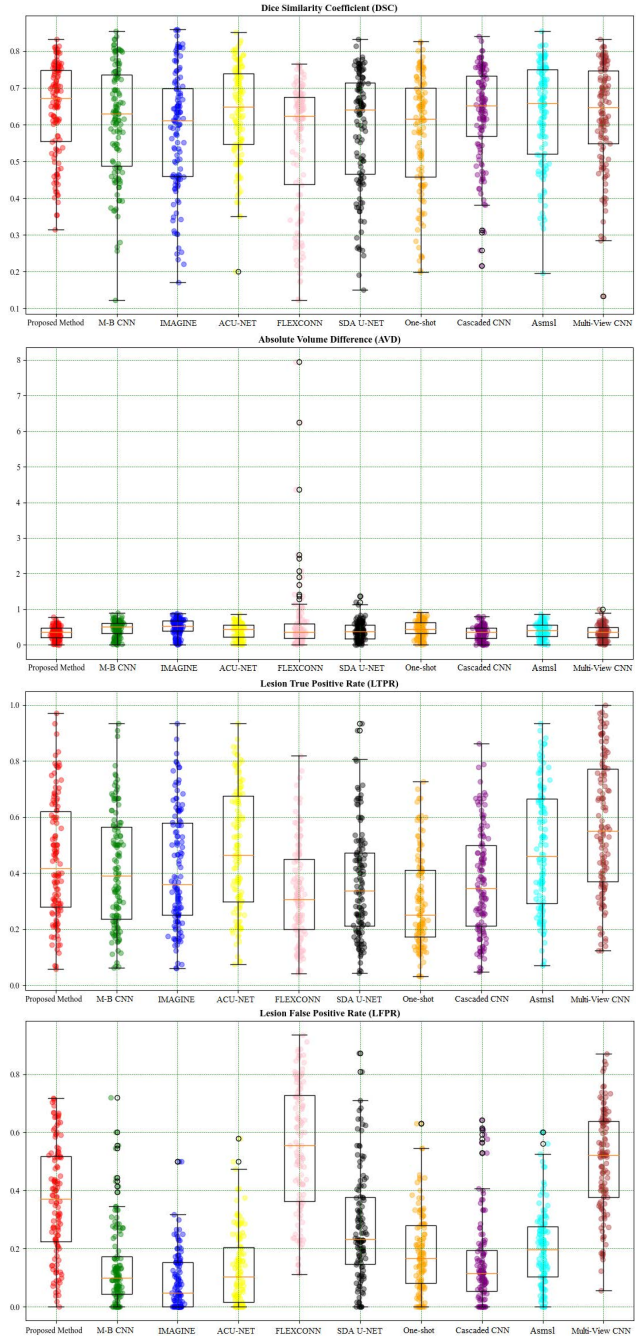

FIGURE 11. Demonstrate the tested models' boxplot with all measures on the ISBI dataset.

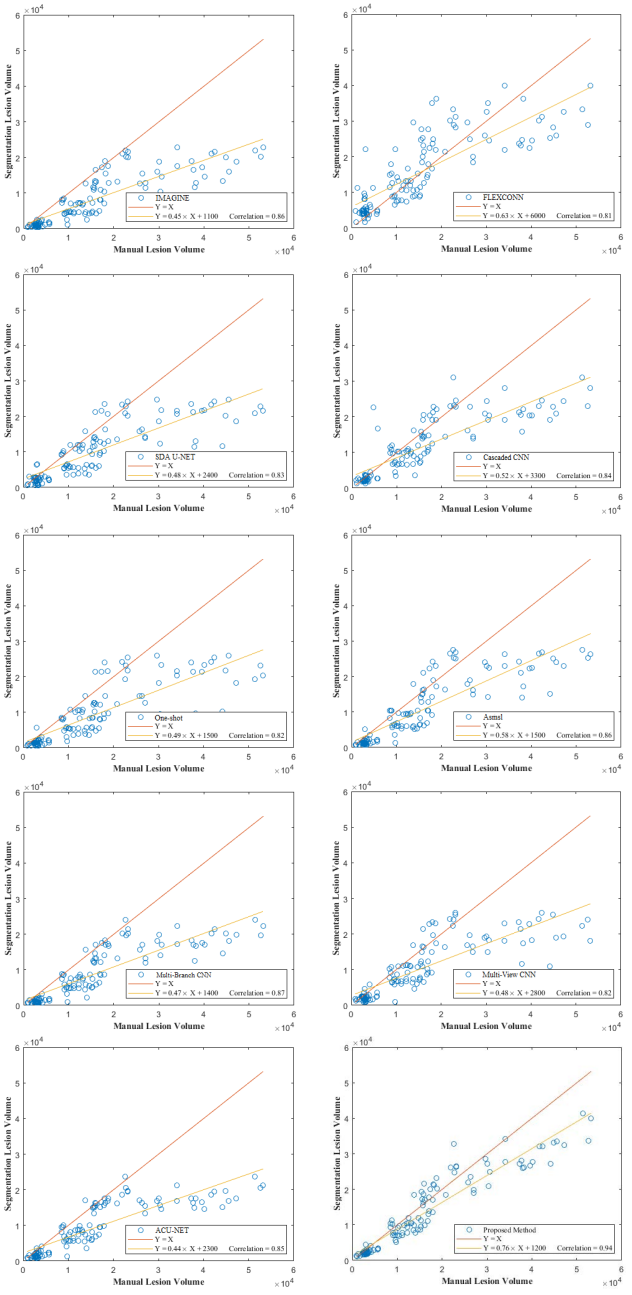

FIGURE 12. Comparing the lesion volumes produced by automatic and manual segmentation on the ISBI dataset. Each point is related to a single lesion.

paper uses a PyTorch library M3d-CAM to generate 3D attention maps [66]. Figure 10 shows some samples of the attention maps. It is obvious when the SCA module is used the unimportant parts are filtered especially in the first slices where the pixels have high gray levels and are more similar to the lesion pixels.

The boxplots of the DSC, LFPR, LTPR and AVD evaluation metrics for different approaches are illustrated in Figure 11. The Figure shows our proposed method performs well in terms of DSC and AVD compare to other state-of-the-art methods. As it is evident, the DSC is in the range of [0.3, 0.85] and most of the samples are over 0.6. In addition, although the mean of the LFPR is high for the proposed method, there are no samples in which the LFPR is over 0.7, and most of the samples are located under 0.45.

Figure 12 shows the relationship between the ground truth and predicted volumes of lesions (each point per lesion). In terms of qualitative evaluation, it can be seen that most of the methods tend to underestimate lesions as many of the points are under the red line significantly when the size of lesions is increasing. On the contrary, the FLEXCONN

method which is proposed by Roy *et al.* [38] tends to overestimate the lesion sizes. However, with quantitative analysis, the proposed method estimated lesion size the slope closest to unity (0.76) and the highest Pearson correlation coefficient (0.94). So, it means the proposed method provides a more robust global agreement between predicted lesion volumes and ground-truth lesion volumes. It is worth mentioning, that a better deal does not mean the model has better accuracy.

IV. DISCUSSION AND CONCLUSION

The present paper introduces an automated lesion segmentation approach that is based on one modality as an input which is rare in the state-of-the-art. The proposed architecture is an end-to-end 3D patch-wise composed of convolution, deconvolution, and an SCA-VoxRes module as an attention module.

In medical image processing, it is common to overcome the limitation of the single modality approach by combining different MRI modalities. Holding that patients cannot afford the cost of combining modalities in medical image segmentation, the current research presents a single modality-based architecture that is more accurate and robust than other multi-modality methods.

Furthermore, patch size is the limitation of patch-based CNNs which suffer from a lack of spatial information for the lesion. Therefore, the proposed method uses an attention module to capture long-range contextual information as a way of attaining spatial information. Consequently, the patch-based CNNs do not lack data, because many random patches can be extracted from 3D data.

Although there are several advantages to the proposed method the limitation of the proposed method should be taken into account. The most important limitation of the proposed method is the high range of LFPR which is due to the existence of two different grand truths in the dataset. In this case, when the model is training, it should try to optimize the loss function according to the logical-or of two grand truths. On the other hand, this research tried to propose a method based on features like using one modal as input and introducing an attention module to improve the robustness of the output segmentation which is more suitable for clinical diagnosis. However, lack of access to clinical data caused we could not show the advantages and disadvantages of the proposed method well, but the proposed method with the public data outperforms well even when the FLAIR is an input.

REFERENCES

[1] L. Steinman, "Multiple sclerosis: A coordinated immunological attack against myelin in the central nervous system," *Cell*, vol. 85, no. 3, pp. 299–302, 1996, doi: 10.1016/s0092-8674(00)81107-1.

[2] A. Compston, H. Winedl, and B. C. Kieseier, "Multiple sclerosis," *Lancet*, vol. 359, no. 9313, pp. 1221–1231, 2002, doi: 10.1016/S0140-6736(02)08220-X.

[3] L. A. Rolak, "Multiple sclerosis: It's not the disease you thought it was," *Clin. Med. Res.*, vol. 1, no. 1, pp. 57–60, 2003, doi: 10.3121/.cmr.1.1.57.

[4] J. Simon *et al.*, "Standardized MR imagingprotocol for multiple sclerosis: Consortium of MS centers consensus guidelines," *Amer. J. Neuroradiol.*, vol. 27, no. 2, pp. 455–461, 2006.

[5] E. M. Sweeney *et al.*, "OASIS is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in MRI," *NeuroImage: Clin.*, vol. 2, pp. 402–413, 2013, doi: 10.1016/j.nicl.2013.03.002.

[6] M. Cabezas, "Boost: A supervised approach for multiple sclerosis lesion segmentation," *J. Neurosci. Methods* vol. 237, pp. 108–117, Nov. 2014, doi: 10.1016/j.jneumeth.2014.08.024.

[7] E. Geremia, O. Clatz, B. H. Menze, E. Konukoglu, A. Criminisi, and N. Ayache, "Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images," *NeuroImage*, vol. 57, no. 2, pp. 378–390, Jul. 2011, doi: 10.1016/j.neuroimage.2011.03.080.

[8] A. Jesson and T. Arbel, "Hierarchical MRF and random forest segmentation of MS lesions and healthy tissues in brain MRI," in *Proc. Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, 2015, pp. 1–2.

[9] X. Lladó *et al.*, "Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches," *Inf. Sci.*, vol. 186, no. 1, pp. 164–185, 2012, doi: 10.1016/j.ins.2011.10.011.

[10] N. Guizard, P. Coupé, V. S. Fonov, J. V. Manjón, D. L. Arnold, and D. L. Collins, "Rotation-invariant multi-contrast non-local means for MS lesion segmentation," *NeuroImage, Clin.*, vol. 8, pp. 376–389, Jan. 2015, doi: 10.1016/j.nicl.2015.05.001.

[11] M. D. Steenwijk *et al.*, "Accurate white matter lesion segmentation by K nearest neighbor classification with tissue type priors (kNN-TTPs)," *NeuroImage, Clin.*, vol. 3, pp. 462–469, Jan. 2013, doi: 10.1016/j.nicl.2013.10.003.

[12] M. J. Fartaria *et al.*, "Automated detection of white matter and cortical lesions in early stages of multiple sclerosis," *J. Magn. Reson. Imag.*, vol. 43, no. 6, pp. 1445–1454, Jun. 2016, doi: 10.1002/jmri.25095.

[13] X. Tomas-Fernandez and S. K. Warfield, "A model of population and subject (MOPS) intensities with application to multiple sclerosis lesion segmentation," *IEEE Trans. Med. Imag.*, vol. 34, no. 6, pp. 1349–1361, Jun. 2015, doi: 10.1109/TMI.2015.2393853.

[14] P. Schmidt *et al.*, "An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis," *Neuroimage*, vol. 59, no. 4, pp. 3774–3783, 2012, doi: 10.1016/j.neuroimage.2011.11.032.

[15] E. Roura *et al.*, "A toolbox for multiple sclerosis lesion segmentation," *Neuroradiology*, vol. 57, no. 10, pp. 1031–1043, Oct. 2015, doi: 10.1007/s00234-015-1552-2.

[16] R. Harmouche, N. K. Subbanna, D. L. Collins, D. L. Arnold, and T. Arbel, "Probabilistic multiple sclerosis lesion classification based on modeling regional intensity variability and local neighborhood information," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 5, pp. 1281–1292, May 2015, doi: 10.1109/TBME.2014.2385635.

[17] M. Strumia, F. R. Schmidt, C. Anastasopoulos, C. Granziera, G. Krueger, and T. Brox, "White matter MS-lesion SegmentationUsing a geometric brain model," *IEEE Trans. Med. Imag.*, vol. 35, no. 7, pp. 1636–1646, Jul. 2016, doi: 10.1109/TMI.2016.2522178.

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, p. 436, Nov. 2016, doi: 10.1038/nature14539.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105, doi: 10.1145/3065386.

[22] X.-H. Han, J. Lei, and Y.-W. Chen, "HEp-2 cell classification using K-support spatial pooling in deep CNNs," in *Deep Learning and Data Labeling for Medical Applications*. Balkans, Greece: Springer, 2016, pp. 3–11, doi: 10.1007/978-3-319-46976-8_1.

[23] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels WithDeep neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 11, pp. 2369–2380, Nov. 2016, doi: 10.1109/TMI.2016.2546227.

[24] S. Valverde *et al.*, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach," *NeuroImage*, vol. 155, pp. 159–168, Jul. 2017, doi: 10.1016/j.neuroimage.2017.04.034.

[25] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, Jan. 2017, doi: 10.1016/j.media.2016.05.004.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[27] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. N. L. Benders, and I. Isgum, "Automatic segmentation of MR brain images with a convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1252–1261, May 2016, doi: 10.1109/TMI.2016.2548501.

[28] X. Guan *et al.*, "3D AGSE-VNet: An automatic brain tumor MRI data segmentation framework," *BMC Med. Imag.*, vol. 22, no. 1, pp. 1–18, Dec. 2022.

[29] H. Huang *et al.*, "A deep multi-task learning framework for brain tumor segmentation," *Frontiers Oncol.*, vol. 11, Jun. 2021.

[30] W. Zhang *et al.*, "ME-Net : Multi-encoder net framework for brain tumor segmentation," *Int. J. Imag. Syst. Technol.*, vol. 31, no. 4, pp. 1834–1848, Dec. 2021.

[31] L. Wu, S. Hu, and C. Liu, "MR brain segmentation based on DE-ResUnet combining texture features and background knowledge," *Biomed. Signal Process. Control*, vol. 75, May 2022, Art. no. 103541.

[32] S. Huang, J. Li, Y. Xiao, N. Shen, and T. Xu, "RTNet: Relation transformer network for diabetic retinopathy multi-lesion segmentation," *IEEE Trans. Med. Imag.*, early access, Jan. 20, 2022, doi: 10.1109/TMI.2022.3143833.

[33] M. Gu, S. Vesal, R. Kosti, and A. Maier, "Few-shot unsupervised domain adaptation for multi-modal cardiac image segmentation," 2022, *arXiv:2201.12386*.

[34] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang, "Joint sequence learning and cross-modality convolution for 3D biomedical segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6393–6400, doi: 10.1109/CVPR.2017.398.

[35] T. Brosch, L. Y. W. Tang, Y. Yoo, D. K. B. Li, A. Traboulsee, and R. Tam, "Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1229–1239, May 2016, doi: 10.1109/TMI.2016.2528821.

[36] S. Aslani *et al.*, "Multi-branch convolutional neural network for multiple sclerosis lesion segmentation," *NeuroImage*, vol. 196, pp. 1–15, Aug. 2019, doi: 10.1016/j.neuroimage.2019.03.068.

[37] C. Hu, G. Kang, B. Hou, Y. Ma, F. Labeau, and Z. Su, "ACU-Net: A 3D attention context u-net for multiple sclerosis lesion segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1384–1388, doi: 10.1109/ICASSP40776.2020.9054616.

[38] S. Roy, J. A. Butman, D. S. Reich, P. A. Calabresi, and D. L. Pham, "Multiple sclerosis lesion segmentation from brain MRI via fully convolutional neural networks," 2018, *arXiv:1803.09172*.

[39] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage*, vol. 170, pp. 446–455, Apr. 2018, doi: 10.1016/j.neuroimage.2017.04.041.

[40] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3146–3154, doi: 10.1109/CVPR.2019.00326.

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[42] M. Ghafoorian and B. Platel, "Convolutional neural networks for ms lesion segmentation, method description of diag team," in *Proc. Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, 2015, pp. 1–2.

[43] A. Birenbaum and H. Greenspan, "Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks," in *Deep Learning and Data Labeling for Medical Applications*. Balkans, Greece: Springer, 2016, pp. 58–67, doi: 10.1007/978-3-319-46976-8_7.

[44] H. M. R. Afzal *et al.*, "Automatic and robust segmentation of multiple sclerosis lesions with convolutional neural networks," *Comput., Mater. Continua*, vol. 66, no. 1, pp. 977–991, 2020, doi: 10.32604/cmc.2020.012448.

[45] Y. Shachor, H. Greenspan, and J. Goldberger, "A mixture of views network with applications to multi-view medical imaging," *Neurocomputing*, vol. 374, pp. 1–9, Jan. 2020, doi: 10.1016/j.neucom.2019.09.027.

[46] S. Vaidya, A. Chunduru, R. Muthuganapathy, and G. Krishnamurthi, "Longitudinal multiple sclerosis lesion segmentation using 3D convolutional neural networks," in *Proc. Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, 2015, pp. 1–2.

[47] S. Valverde *et al.*, "One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks," *NeuroImage, Clin.*, vol. 21, Jan. 2019, Art. no. 101638, doi: 10.1016/j.nicl.2018.101638.

[48] S. Andermatt, S. Pezold, and P. C. Cattin, "Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units," in *Proc. Int. MICCAI Brainlesion Workshop*. Toronto, ON, Canada: Springer, 2017, pp. 31–42, doi: 10.1007/978-3-319-75238-9_3.

[49] M. Salem *et al.*, "Multiple sclerosis lesion synthesis in MRI using an encoder-decoder U-NET," *IEEE Access*, vol. 7, pp. 25171–25184, 2019, doi: 10.1109/ACCESS.2019.2900198.

[50] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, "Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection," *IEEE Access*, vol. 7, pp. 1721–1735, 2019, doi: 10.1109/ACCESS.2018.2886371.

[51] S. Aslani, M. Dayan, V. Murino, and D. Sona, "Deep 2D encoder-decoder convolutional neural network for multiple sclerosis lesion segmentation in brain MRI," in *Proc. Int. MICCAI Brainlesion Workshop*. Madrid, Spain: Springer, 2018, pp. 132–141, doi: 10.1007/978-3-030-11723-8_13.

[52] A. Carass *et al.*, "Longitudinal multiple sclerosis lesion segmentation: Resource and challenge," *NeuroImage*, vol. 148, pp. 77–102, Mar. 2017, doi: 10.1016/j.neuroimage.2016.12.064.

[53] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *Neuroimage*, vol. 17, no. 2, pp. 825–841, 2002, doi: 10.1016/S1053-8119(02)91132-8.

[54] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Trans. Med. Imag.*, vol. 17, no. 1, pp. 87–97, Feb. 1998.

[55] K. Oishi *et al.*, "Human brain white matter atlas: Identification and assignment of common anatomical structures in superficial white matter," *NeuroImage*, vol. 43, no. 3, pp. 447–457, Nov. 2008, doi: 10.1016/j.neuroimage.2008.07.009.

[56] S. M. Pizer *et al.*, "Adaptive histogram equalization and its variations," *Comput. Vis., Graph., Image Process.*, vol. 39, no. 3, pp. 355–368, 1987, doi: 10.1016/S0734-189X(87)80186-X.

[57] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vols. 3–4, Sep. 2019, Art. no. 100004, doi: 10.1016/j.array.2019.100004.

[58] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988, doi: 10.1109/iccv.2017.324.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[60] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. 7, pp. 1–39, Jul. 2011.

[61] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012, *arXiv:1212.5701*.

[62] Y. N. Dauphin, H. de Vries, and Y. Bengio, "Equilibrated adaptive learning rates for non-convex optimization," 2015, *arXiv:1502.04390*.

[63] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Inf. Fusion*, vol. 77, pp. 29–52, Jan. 2022.

[64] Q. Ye, J. Xia, and G. Yang, "Explainable AI for COVID-19 CT classifiers: An initial comparison study," in *Proc. IEEE 34th Int. Symp. Computer-Based Med. Syst. (CBMS)*, Jun. 2021, pp. 521–526.

[65] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[66] K. Gotkowski, C. Gonzalez, A. Bucher, and A. Mukhopadhyay, "M3d-CAM: A PyTorch library to generate 3D data attention maps for medical deep learning," 2020, *arXiv:2007.00453*.

[67] M. Rezanejad *et al.*, "Medial spectral coordinates for 3D shape analysis," 2021, *arXiv:2111.13295*.