TransUNet for psoriasis lesion segmentation

Samiksha Soni
Electrical Engineering Department
National Institue of Technology Raipur
Raipur, India
ssoni.jrf2022.ee@nitrr.ac.in

Dr. Narendra D. Londhe
Electrical Engineering Department
National Institue of Technology Raipur
Raipur, India
nlondhe.ele@nitrr.ac.in

Ritesh Raj
Electrical Engineering Department
National Institue of Technology Raipur
Raipur, India
rraj.phd2018.ee@nitrr.ac.in

Dr. Rajendra S. Sonawane
Psoriasis Clinic and Research Centre
Psoriatreat,
Pune, India
drrajss@gmail.com

***Abstract*— Outstanding performance of the transformer-based model in the field of natural language processing has piqued the interest of researchers in investigating these techniques for computer vision. And the most popular UNet model is considered a major player in the field of image segmentation. Thus, in this paper, we have proposed the transformer-based UNet model for the complex task of psoriasis lesion segmentation from raw color images. One of the major challenges for our segmentation task is the scarcity of datasets and to overcome this challenge we have exploited the EfficientNetB1 transfer learned model as a backbone for our segmentation model. The proposed model is evaluated for the 70:30 hold-out data division technique and the segmentation performance is evaluated using the Dice Score (DS) and Jaccard Index (JI). The value of DS and JI obtained for the intended task are 0.9571 and 0.9201 respectively with the proposed model. Comparative analysis with different derivatives of the UNet model and state-of-the-art literary work shows the better performance of our proposed model.**

Keywords*— *Psoriasis, Skin lesion, Image segmentation, Vision transformer, UNet, Deep learning.

I. INTRODUCTION

In many current clinical procedures, medical image segmentation is a crucial step. It has a variety of uses, such as in diagnostic procedures, planning, and better management of those treatments. Recent research shows that skin lesion identification based on deep leaning framework has outperformed dermatologists [1]. Image processing advancements and the incorporation of deep learning techniques, particularly CNN, have given medical science a distinct advantage. The application of CNN and its automated analysis methods to process images has proven to be a vital tool. This can give an automated intelligent system for the diagnosis of skin lesions outside the clinical environment [2-3]. As a result, computer-aided diagnostics now includes the automatic examination of skin lesions and their severity for objective diagnosis of different skin diseases.

Using two-dimensional digital images of Psoriasis patients, several research works have been implemented in the past few years. Most of those machine-learning approaches made use of color, texture, and gray-level features for psoriasis lesion identification. Shrivastav et. al. [4] used manually crafted feature-based techniques for psoriasis lesion analysis. However, their method is computationally bulkier for the analysis of a large dataset. Dash et. al. [5] shifted the trend of machine learning psoriasis lesion assessment toward deep learning approaches. But, most of the previous research works in the field of psoriasis lesion analysis were on the dataset having cropped single patches of psoriasis lesions. Such approaches have only been found to be useful for the subsequent classification or detection challenge. However, the major advancement is done by Raj et. al. [6] to analyze the efficacy of CNN models for multiple lesion segmentation on psoriasis images which is very helpful in calculating the PASI score for assessing the overall severity of Psoriasis. The proposed method achieved remarkable accuracy in psoriasis lesion segmentation using raw images. In [7], the authors extended the previous work and highlighted the impact of transfer learning on the compact UNet model. The work reflected the great improvements in the segmentation performance with the aid of transfer learning.

The automatic psoriasis lesion segmentation is difficult due to ambiguity like some lesions have a light tint or high similarity with healthy skin. It gets more difficult when the color or visual patterns are similar and the lightning conditions vary. As can be seen from Fig. 1, the images have varying backgrounds, shadowing effects, skin hair, and different illumination conditions.

Fig. 1: Sample images from the dataset

978-1-6654-9291-1/22/$31.00 ©2022 IEEE.

Despite the superior representational power of CNN-based techniques, they are often limited in modeling explicit long-range relationships due to the inherent locality of convolution processes. As a result, these architectures frequently generate unsatisfactory outcomes, especially for target components that vary significantly between patients in context with texture, shape, and size. To solve this issue, existing works proposed developing a self-attention mechanism on CNN features. Transformers (i.e. built for sequence-to-sequence prediction) have developed as alternative architectures that do not use convolution operators at all and instead rely exclusively on attention processes [8]. Transformers, in contrast to earlier CNN-based techniques, are not only stronger at modeling global settings, but also exhibit better transferability for downstream tasks when trained on a broad scale. In [9], authors proposed a feature adaptive transformer for skin lesion segmentation. They replace the single branch of the encoder with a dual branch one branch for the transformer and another for CNN to have a better feature extraction in terms of local as well as global features. A blend of CNN and transformer is proposed by Wang et al. [10] to have a better segmentation performance on ISIC 2018 dataset. Numerous works are going on for transformer-based skin lesion analysis for melanoma disease [11]. Taking a lead from these ongoing researches, we have tried to implement a transformer and CNN-based model for psoriasis lesion segmentation.

In this paper, we have employed TransUNet for psoriasis lesion segmentation, which develops self-attention mechanisms in the context of efficient segmentation prediction. TransUNet is a hybrid structure that leverages the strength of both CNN and the transformer. The global features provided by the transformer and the comprehensive high-resolution spatial information from the pre-trained backbone network make an effective compensation for the loss over the use of individual networks alone. The following points summarize the contribution of our proposed work:

- We have highlighted the major challenges in psoriasis lesion segmentation and its impact on segmentation performance.
- We present a detailed comparative analysis of the various UNet model derivatives for the task of psoriasis lesion segmentation.
- We compare the performance of TransUNet with different standard segmentation models and existing models for psoriasis lesion segmentation.

The organization details of the remaining paper are as follows: Section II deals with the methodology part of our work. In Section III, the details about the experimental setup are given. Section IV deals with results and discussion. Finally, Section V concludes the paper along with the future prospect of the work.

II. METHODOLOGY

In this work, the TransUNet-based model is implemented for psoriasis lesion segmentation from raw color images. The purpose of the work is to explore the efficacy of the transformer as well as CNN for the performance enhancement of lesion segmentation. The implementation procedure consists of a training and testing module where the whole dataset is divided in the ratio of 70:30 (70% of data for training and 30% for testing), the model is trained with raw color images acquired from different psoriasis patients along with their ground truth annotation. In the second module, the best weight obtained from training is used to generate a probability map for the test images. These probability maps are further smoothened to generate the predicted lesion, which is visualized as a binary mask. At last, the segmentation performance metrics are evaluated by comparing the predicted mask with the ground truth annotations.

The architectural detail of the proposed model is shown in Fig. 2, consisting of a transformer with an EfficientNetB1 backbone, skip connection, and decoder, which altogether resembles the UNet framework. The subsection below shows further details of the encoder and decoder section of the proposed model.

A. Trans-Efficient Encoder

Trans-Efficient encoder is a combination of the pre-trained EfficienNetB1 [11] backbone network's locality and the transformer's long-term dependence. EfficienNetB1 [11] is used to generate a feature map from the input image having a size of 512×512. This network does not only go into great detail but does introduce three dimension scaling. It uses a set of ratios to scale all three dimensions i.e. height, width, and resolution. This multidimensional scaling approach of the EfficientNetB1 network is different from the approaches of traditional CNN models. The feature map generated from the transfer learned encoder network of the proposed model is of size 256×256, 128×128, 64×64, and 32×32. Further, the patch embedding is applied to 1×1 patches of the lowest feature map generated by the selected backbone network. By combining the patch embeddings with the location embeddings, the patch spatial information is obtained by Eq. (1), shown below:

$$T_{in} = [s_1E; s_2E; \ldots; s_nE] + E_{pos} \quad (1)$$

where T_{in} corresponds to transformer layer input, s_i is the patch feature obtained from the last layer of the backbone network, E represents linear projection, E_{pos} shows position embedding, n corresponds to patch number, and *;* represents concatenation operation.

T_{in} is processed further through the transformer encoder which consists of X layers of Multi-head Self-attention (MUA) and Multi-layer Perceptron (MLP) blocks. At MLP, we have used the Gelu activation function. Gelu provides the advantage of stochastic regularizers. Hence, the final output for x^{th} layer of transformer block (T_x) can be given by Eq.'s (2) and (3), shown below:

$$T'_x = MUA(L_{nor}(T_{in_{(x-1)}})) + T_{in_{(x-1)}} \quad (2)$$

$$T_x = \mathrm{MLP}(L_{nor}(T'_x)) + T_{in_{(x-1)}} \quad (3)$$

where; L_{nor} represents layer normalization and T'_x refers to the feature representation generated from the MUA block of x^{th} layer transformer.

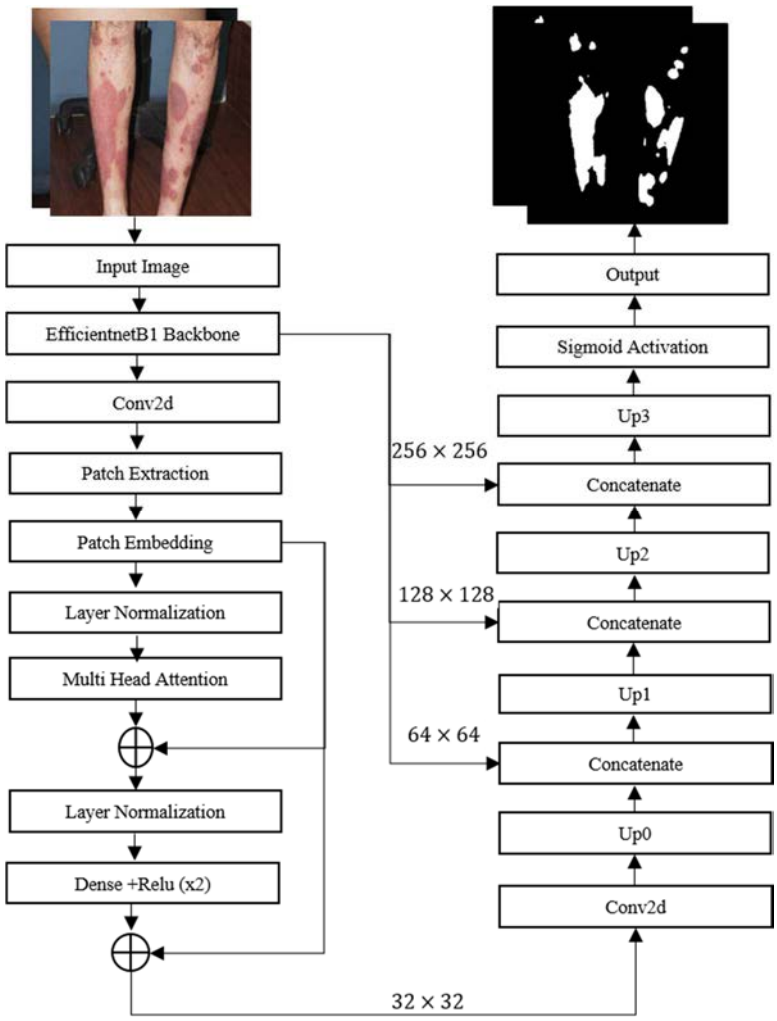


Fig. 2: TransUNet Architecture

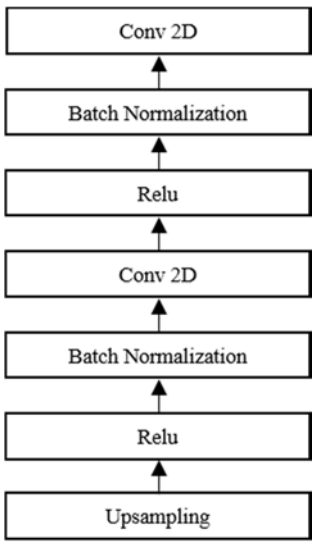


Fig. 3: Details of Up(0-3) blocks

B. Decoder

As illustrated in Fig. 2, the final output of the encoder act as input for the decoder. At each stage, the input feature maps are up-sampled and concatenated with the respective feature maps of EfficientNetB1 that is passed through the long skip connection to the corresponding upsampling blocks. The upsampling blocks (i.e. Up(0-3)) consist of a combination of convolution, Relu, batch normalization, and upsampling blocks as shown in Fig 3. The decoder part resembles the simple UNet decoder framework. However, the difference resided in the fused feature in the decoder from the encoder side. One from the skip-connected local features of the EfficientNetB1 network and the other from the global context features of the transformer. At last, the final probability map is obtained using the sigmoid activation function because of the binary classification task. The output obtained from the last layer of the model is in the form of a probability mask similar to the dimension of the original input image. This mask is finally converted into binary mask form by hard thresholding with a value of 0.8.

III. Experimental Details

The dataset used in this work was taken from [7]. It consists of 500 RGB images of size 512×512 along with ground truth annotations validated by expert dermatologists. The complete model is implemented using Python programming language on a workstation having specifications as NVIDIA Quadro P5000 GPU with 16 GB memory and Intel Xeon Silver 4116 CPU @2.10 GHz frequency with 64 GB RAM.

We have optimized the parameter of the basic TransUNet model experimentally for our segmentation task. The details of different model hyperparameters used in this work are provided in Table I. The selection of hyperparameters was done through exhaustive experiments. With the selected model hyperparameters, the loss curve was found to converge better during the training process. The mini-batch size of 20 is selected considering the training speed and computational cost. The initial value of the learning rate is selected as 0.0001. Instead of a constant learning rate, we have adopted the adaptive learning mechanism for our model training that decreases the learning factor when performance stagnates. This also resolves the issue of speed resulting in achieving better performance in fewer epochs, as well as the gradient exploding problem during training. We selected the number of MLP by hit and trial method so that it can provide an optimum balance between the model size and its performance.

Table I: Model Hyperparameters

Parameter	Value
Mini Batch size	20
filters	[8,16,32,64,128]
Initial learning rate	0.0001
Epochs	100
Optimizer	Adam
Mlp_activation	Gelu
Embed_dim	768
Num_mlp	3072

Also, we have used the most lightweight EfficientNetB1 [13] network instead of using ResNet50 which is used as a backbone in basic TransUNet [12]. Table II shows the trainable parameter for TransUNet with different backbones. With the use of EfficientNetB1 as a backbone, the trainable parameter of TransUNet is found to be reduced.

Table II: TransUNet with different backbone.

Model	Total No. of parameters
VGG16	23,440,873
ResNet50	17,367,081
Mobilenetv2	16,122,537
Densenet121	13,375,209
EfficientNetB0	9,494,448
EfficientNetB1	9,855,252

IV. Result and Discussions

To demonstrate the segmentation ability of the proposed method, we have split the dataset into a 70:30 ratio where 70% of images are used for training and 30% is used for testing purposes. The performance of the TransUNet is evaluated based on the most effective segmentation metrics i.e. Jaccard Index (JI) and Dice Score (DS), which handles class-imbalance problems pretty well. The final values obtained by the proposed model for JI and DS metrics are 0.9201 and 0.9571 respectively for the intended task. The satisfactory performance of the model is also reflected in the training and testing curves shown in Fig. 4.

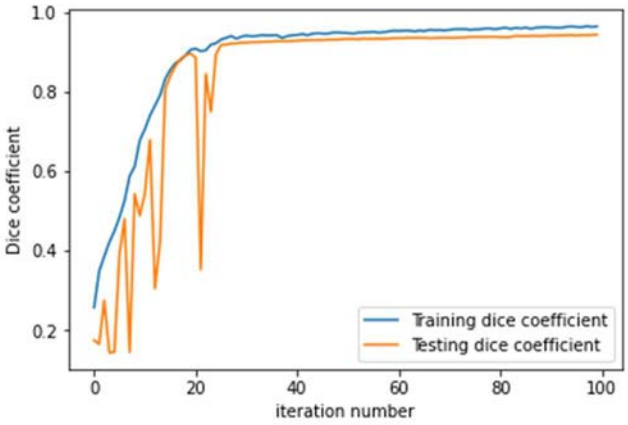


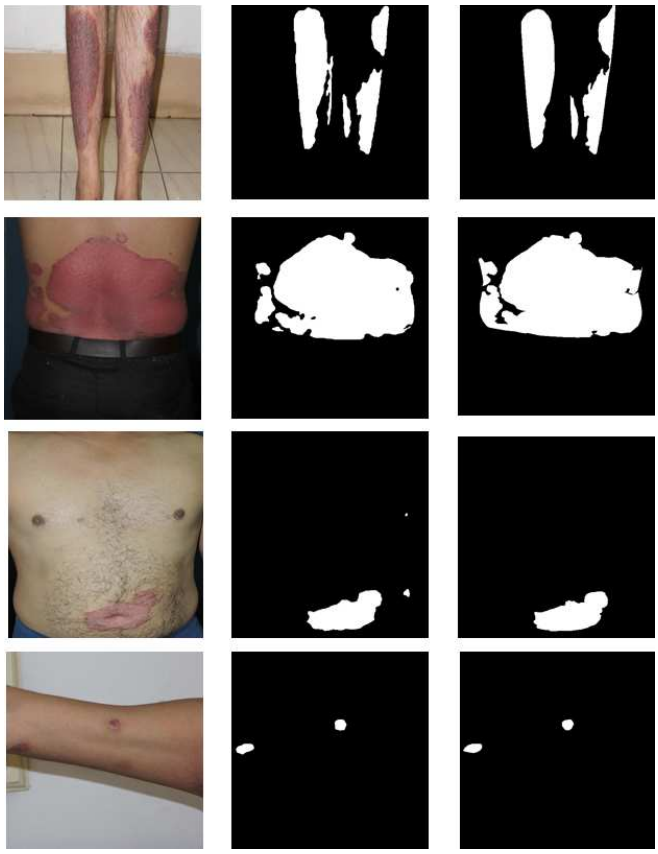
Fig. 4: Training and Testing Curve

The prediction results of our proposed work are shown in Fig. 5, where Fig. 5(a) corresponds to the raw RGB images from the dataset. As can be seen, these images comprise various artifacts like skin hair, clothes, varying backgrounds, illumination, etc. Fig. 5(b) corresponds to the predicted output, and Fig. 5(c) represents the ground truth annotations validated by an expert dermatologist. From the analysis of predicted output, we can say that our proposed model efficiently does the segmentation on raw images without the need for any preprocessing steps.

To evaluate the efficacy of the TransUNet model, we have compared it with the different variants of the UNet model listed in Table III. The performance of all these models is obtained by fine-tuning the pre-trained EfficientNetB1 that is adopted in the encoder part as the backbone. From the different UNet variants, the standard UNet model shows better performance than its other variants and r2-UNet has the smallest parameters but its performance is less than 90% for both the evaluation parameter i.e. DS and JI. From Table III, it can be observed that there is a considerable amount of improvement in the segmentation performance with the proposed hybrid model (i.e. TransUNet) over other listed variants of UNet models.

Table III: Performance of proposed model with different variants of the UNet model

Model (with EfficientB1 as a backbone)	Total no. of parameters	DS	JI
UNet [15]	2,010,900	0.9235	0.8512
UNet++ [16]	2,094,927	0.9182	0.8484
r2U-Net [17]	1,603,857	0.8892	0.7789
TransUNet (Proposed model)	9,855,252	0.9571	0.9201



(a) Raw RGB Images (b) Predicted Output (c) Ground truth annotation

Fig. 5: Prediction results

We have also made a comparative analysis of our segmentation model with the existing state-of-the-art models that are tabulated in Table IV. From Table IV, we can state that the performance of our TransUNet model is far better than FCN [14] and SegNet [18] in terms of both model parameters and segmentation performance. PsLSNet [5] is the mini U-net model with less number of trainable parameters but the overall performance of this model is about 5.8% of DS and 10.3% of JI less compared to the proposed one. The proposed model shows little improvement in the segmentation performance over the ResUNet model [7]. But, this little improvement is a much more significant one as the proposed model can detect nipples as non-lesions which is one of the major drawbacks of ResUNet model observed in [7]. Also, this improvement is achieved from a relatively 55.32% lighter model over ResUNet model. Therefore, the proposed model can differentiate between the lesion and non-lesions (i.e. healthy skin, background, nipple areas) efficiently and effectively over the existing models.

Table IV: Comparison with state-of-the-art methods

Model	**DS**	**JI**	**Training parameters (in million)**
FCN [14]	0.6482	0.5101	134.33
SegNet [18]	0.8035	0.7023	29.46
PsLSNeT [5]	0.8993	0.8174	0.49
ResUNet [7]	0.9481	0.9011	22
TransUNet (Proposed model)	0.95715	0.92012	9.83

V. CONCLUSION

In this paper, we have implemented a blend of CNN and transformer for the complex task of psoriasis lesion segmentation. The TransUNet architecture can automatically segment the psoriasis lesion with high segmentation performance without the need for manual pre-processing or human-engineered features. The efficacy of the proposed work is evaluated for the 70:30 hold-out data division protocol. The Dice Score of 0.9572 and Jaccard Index of 0.9201 shows better model performance in comparison with other segmentation models for the intended task. The effectiveness of the suggested method is confirmed by a comprehensive comparison study with existing methods in the literature.

However, there is still considerable scope for improvement, as can be seen from segmentation results that a minor portion of the nipple is being misclassified as a psoriasis lesion, and missed classification of the lesion in case of shadowing effect near the region of the body curve. The model parameters are comparatively much higher than the existing derivatives of UNet compared in this work. Therefore, the hypothesis testing of transformer techniques with different variants of the UNet model will need to be explored enormously as the further scope of research. Also, this work may lead to the development of a more efficient and lightweight segmentation model for psoriasis lesion segmentation with high-resolution images in future research.

ACKNOWLEDGMENT

Authors acknowledge the Science and Engineering Research Board (SERB), Government of India, for financial support vide Reference No. EEQ/2021/000129 to carry out this work.

REFERENCES

[1] Yang, Yiguang, et al. "A convolutional neural network trained with dermoscopic images of psoriasis performed on par with 230 dermatologists." Computers in Biology and Medicine 139 (2021): 104924.

[2] Shanthi, T., R. S. Sabeenian, and R. Anand. "Automatic diagnosis of skin diseases using convolution neural network." Microprocessors and Microsystems 76 (2020): 103074.

[3] Padilla, Dionis, et al. "Differentiating atopic dermatitis and psoriasis chronic plaque using convolutional neural network mobilenet architecture." 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM). IEEE, 2019.

[4] Shrivastava, Vimal K., et al. "Exploring the color feature power for psoriasis risk stratification and classification: a data mining paradigm." Computers in biology and medicine 65 (2015): 54-68.

[5] Dash, Manoranjan, et al. "PsLSNet: Automated psoriasis skin lesion segmentation using modified U-Net-based fully convolutional network." Biomedical Signal Processing and Control 52 (2019): 226-237.

[6] Raj, Ritesh, Narendra D. Londhe, and Rajendra Sonawane, "Automatic Psoriasis Lesion Segmentation from Raw Color Images using Deep Learning," 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020

[7] Raj, Ritesh, Narendra D. Londhe, and Rajendra Sonawane. "Automated psoriasis lesion segmentation from unconstrained environment using residual U-Net with transfer learning." Computer Methods and Programs in Biomedicine 206 (2021): 106123.

[8] Parmar, Niki, et al. "Image transformer." International conference on machine learning. PMLR, 2018.

[9] Wu, Huisi, et al. "FAT-Net: Feature adaptive transformers for automated skin lesion segmentation." Medical Image Analysis 76 (2022): 102327.

[10] Wang, Jiacheng, et al. "Boundary-aware transformers for skin lesion segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021.

[11] Gulzar, Yonis, and Sumeer Ahmad Khan. "Skin Lesion Segmentation Based on Vision Transformers and Convolutional Neural Networks—A Comparative Study." Applied Sciences 12.12 (2022): 5990.

[12] Chen, Jieneng, et al. "TransUNet: Transformers make strong encoders for medical image segmentation." arXiv preprint arXiv:2102.04306 (2021).

[13] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." International conference on machine learning. PMLR, 2019..

[14] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[15] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[16] Zhou, Zongwei, et al. "UNet++: A nested u-net architecture for medical image segmentation." Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Cham, 2018. 3-11.

[17] Alom, Md Zahangir, et al. "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation." arXiv preprint arXiv:1802.06955 (2018).

[18] Karthik, R., et al. "A deep supervised approach for ischemic lesion segmentation from multimodal MRI using Fully Convolutional Network." Applied Soft Computing 84 (2019): 105685.