iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41990

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

Tongue Segmentation for Disease Diagnosis a Precise and Fast method using Double U-Net Architecture

Vibha Bhatnagar[1], Prashant P Bansod[2], Gauri Gupta[3]
Department of Biomedical Engineering, Shri G. S. Institute of Tech. & Science, 23 Sir M. Visvesvaraya Marg, Indore, Madhya Pradesh 452003, India.
R.G.P.V University

Abstract: ***A robust automatic tongue diagnosis system greatly relies on accurate segmentation of the tongue body from the image. It is a challenge to accurately segment tongue body from an image having close interference of teeth, lips and face. Deep Learning methods such as Deep Convolution Networks like FCN (fully connected Networks), U-NET, Res-Net (Residual Network) have superseded the performance of conventional techniques like Active Contour, Gradient V Flow, Level Set etc. . In this paper looking into the tremendous capability of Deep neural Networks, we have employed Double U-NET for tongue image segmentation of images captured by mobile device and compared the results with that of U-NET, and Res U-NET architectures. Qualitative as well as quantitative analysis of the three reveals a superior performance of Double U-net especially in images with additional dominant features of the face such as lips, teeth, and spaces with the tongue image.***
Keywords: Double U-Net, Residual U-Net, U-Net, tongue segmentation, automatic tongue diagnosis

I. INTRODUCTION

Tongue body segmentation is a major task which directly affects the results from an automatic tongue diagnosis system. In traditional medicine diagnosis is based on analysis of features such as colour, texture, coating, shape etc. from the surface of the tongue. The major challenges associated with tongue body segmentation are - Motion artefacts associated with image capturing, tongue and its surrounding regions with minor colour intensity variations and interference of teeth and lips. Traditional Methods for segmentation roughly falling under basic classes such as thresholding, edge detection, graph theory, level set, active contours gave fairly good accuracy, but had certain limitations associated along-with them.

Deep Convolution Neural Networks present an outstanding ability of feature learning and representation, with viable solutions for automatic, generalizable and efficient semantic image segmentation. The key challenges associated with medical image segmentation are unavailability of large number of annotated and labelled dataset for training the model one of the basic requirements of deep learning models to give accurate results, lack of standard segmentation protocol and huge variations of images among patients. This calls for the quantification of segmentation accuracy to estimate the performance on other applications. With the existing challenges in this domain, a robust and efficient technique for segmenting the body organ or anomaly under investigation was of utmost priority for automation in biomedical applications. Three architectures claiming to be as backbone for medical applications are introduced in brief details.

Olaf Ronnerberg, et.al. [1] designed a network whose architecture visually matched alphabet 'U', hence named it U-NET. Network architecture consisted of a contracting path for contextual features followed by a symmetric expanding path for precise localization. U-Net performed well with small, augmented dataset. Debesh Jha. et.al. [2] proposed an enhanced modified version of U-NET and called it Double U- NET, consisting of two U-net architecture stacked one on top of other. It used pretrained VGG19 as the encoder module and ASPP (Atrous Spatial Pyramid Pooling) slightly differing with original UNET architecture. Weihao Weng [3] designed INET architecture motivated by Inception Model, which can recover low-level semantics by concatenating feature maps of all preceding convolution layers, as well as expediting training by multiple shortcut additions. The use of increasing size kernel in convolution layers was advantageous for feasibility for Biomedical Images.

In this paper Double U- Net model is used to train tongue image dataset and results are compared with the results obtained with U-NET and Res- UNET model. Qualitative and Quantitative comparison are presented. Section 2 contains related work in this field, Section 3 gives Methodology, Section 4 -Results, Section 5- Discussion, Section 6- Conclusion.

II. RELATED WORK

Deep learning techniques have shown impressive performance in terms of speed of computation, complexity and segmentation results with respect to the traditional methods. Jiang Li, et.al. [4] segmented the tongue region from the image by enhanced HSV colour model Convolution Neural Network. To obtain clear edges RGB image after being converted to HSV model was passed through Contract Limited Adaptive Histogram Equalization (CLAHE). Their model tested on a dataset of 264 images taken by digital camera and results compared with Snakes Model. Results of comparison showed better performance of their model enhanced HSV -CNN over Snakes, especially processing time drastically decreased to 0.0275 sec from time of 3.1355 sec for Snakes model. Yushan Xue, et.al. [5] used fully convolution networks (FCN) for tongue body segmentation of images of size 379 x 489 captured by customized equipment. Performance comparison of their model of FCN-8s was evaluated with that of Deep Lab V3, Deep Lab V3 Plus Learning Based Matting (LBM) resulting in highest Mean Intersection of Union (mIOU) of 93.7% being achieved with Deep Lab V3 against a value of 90.48% with FCN-8s. Though Fully convolution model was fastest amongst the three, they concluded Deep Lab V3 to have better performance on quantitative analysis.

Bingqian Lin, et.al. [6] proposed a model that did not have tight constraints regarding the image quality regarding its size and illumination during acquisition, in other words did not require any prior pre-processing. The Deep CNN put forward by them was based on RES- Net (Residual Neural Network) with different number of layers. Primarily they tested their model on 2344 images captured by cell phone, with different number of layered Res-Net models and compared to traditional Grab -Cut model. From all the different number layer architecture considered, 50 layered Res- Net model showed superior performance. Wei Yuan, et.al. [7] quoted to have developed light weight with better accuracy model over their counterpart higher performance algorithms. A labelled annotation method was also developed along with for reducing manual work for annotation on dataset. Model was tested on 5616 images taken by digital camera. Researchers claim that their model could contribute extensively to automatic remote applications and performance can be further enhanced by exploring use of sample mining and hyper parameter optimization.

XinLei Li, et.al. [8] designed a light weight Encoder -Decoder Architecture, tested on dataset of 5,600 images (FDU/SHUTCM) generated by them and also on publicly available dataset BIOHIT (12 images) and PolyU/HIT (300 images). Segmentation accuracy achieved for their dataset is quoted to be 99.15%. Model architecture consisted of TIFE (Tongue Image Feature Extraction) to extract features with larger receptive fields without loss of spatial resolution, whereas a Context Model is used to increase the performance by aggregating multiscale contextual information. Decoder is designed as simple yet efficient feature up sampling module fusing different depth features and refining segmentation results along tongue boundary. Misclassification error due to class imbalance is also taken care of in loss module.

Xiadong Huang, et.al. [9] developed an enhanced fully convolution network with encoder-decoder structure. Encoder consisted of Deep Residual Network for dense feature maps followed behind by Receptive Field Block to capture sufficient global contextual information. Decoder module incorporated feature Pyramid Network to fuse multiscale feature maps to acquire sufficient positional information to recover the contour of tongue body. Results delivered sensitivity of 98.97% with average Dice Similarity Coefficient of 97.26% on a dataset comprising of 700 images. Qichao Tang, et.al. [10] compared performance of deep learning-based model Mobile Net V2 in combination with Single Shot Multibox Detector (SSD) with conventional method Haar like feature-based detection algorithm. When applied on a dataset of 798 images deep learning based model showed better sensitivity.

III. METHODOLOGY

In order to get exceptionally good results of segmentation it is very crucial to use low level details while retaining high level semantic information. In the following paragraphs brief overview of the deep learning models considered for comparison on tongue images is given. UNET is the base line model, Res-U-NET and Double U-NET build on the U- Net architecture with enhanced characteristics of Residual architecture and two U-NET architecture combined together with some modification respectively. Followed by subsections giving details of dataset considered experimental setup and parameters on which basis quantitative & qualitative analysis is done.

A. Overview of Model Architectures

U-NET Architecture developed by Olaf Ranneberger, et.al. [1] a deep neural network architecture consisting of a contracting path for context capturing and a symmetric expanding path that enables precise localization. U-NET architecture mainly consists of two parts analysis (encoder) module and synthesis(decoder) module. Training strategy of the proposed architecture relied on data augmentation of annotated samples in efficient manner. That is, it showed good performance even in case of small data set as opposed to the basic requirement of large dataset for effective training of deep networks.

One important characteristic of their model is that while up-sampling there are large number of connected feature channels that enable propagation of contextual information to higher resolution layers. Due to symmetric expansion and contracting path of the network it takes a U-shaped architecture.

Zhengxin Zhang [11] combined the strengths of U-NET along with Residual learning to develop a new model called Res U-NET for road extraction from aerial images, later also found to be successful with biomedical images. This combination enables to achieve two-fold advantage by using residual blocks for deeper networks with no constant worry of vanishing and exploding gradients, it also facilitates easy training of the network. Second important advantage is by rich skip connections better information propagation is ensured, allowing network with fewer parameters and better performance. The difference between U-Net and Res U-Net architectures is that the plain neural units of U-NET are replaced by Residual Units. Secondly cropping operation is removed rendering a much-sophisticated architecture and better performance.

Double U-NET Architecture proposed by Debesh Jha. et.al. [2], Double UNET differs from UNET as it uses two set of UNET architecture in its single model structure. Model designers have used a modified UNET as first network in Double UNET whereas keeping the second network almost similar to original U-NET.

Network 1 uses pretrained VGG19 followed by 3x3 convolution layer, batch normalization, ReLU activation, squeeze and excite block and final 2x2 max pooling with stride 2 to reduce spatial dimensions. This is one major difference as compared to U-NET encoder. In order to extract high resolution feature maps and achieve superior performance ASPP (Atrous Spatial Pyramid Pooling) is done after the encoder block. Decoder block in both the networks performs 2x2 bilinear up sampling on input features thereby doubling the dimension of the input feature map. Model creators have also used Squeeze and Excite block to reduce redundancy in information and passing only relevant data in encoder of network 1 as well as decoder blocks of both the networks. Output from network 1 is element-wise multiplied by the input to it and this is passed as input to network 2.

Second network is similar to original UNET structure except that it contains ASPP block in between encoder and decoder. The output mask produced by this network 2, i.e., output 2 is concatenated with the output 1 of previous network to see the difference between the two. Skip connections are used in both the parts of the model, network one used skip connections to concatenate encoder to output feature map on the contrary second network uses skip connections both from encoder of first network as well as second network, as a result spatial resolution is maintained and quality of the output feature map is enhanced. The structure and settings of Double U-Net architecture as applied to tongue segmentation is shown in Fig.1. The size of the input image is 576x768 and the output is binary image of the same size.

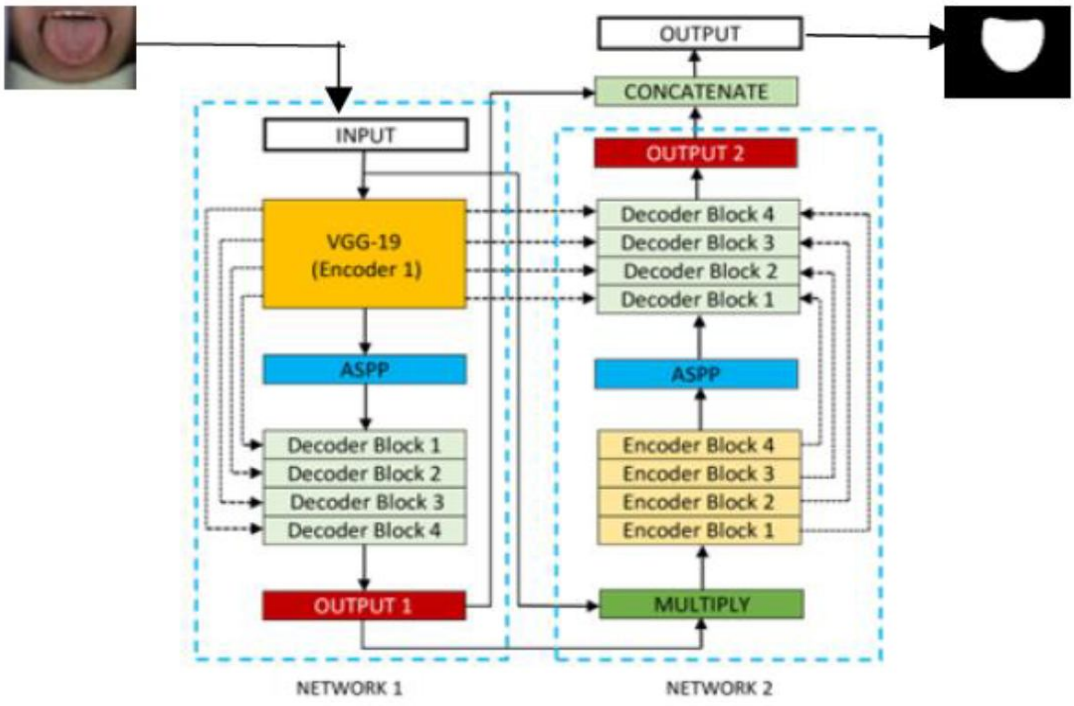


Fig. 1 Block Diagram of Double U-Net Architecture.

B. Experimental Methods

This section elaborates upon the experiment performed using Double U-Net to demonstrate its efficiency and effectiveness. First, the dataset used in the experiment is described followed by some insight into the experimental hardware. Next a qualitative evaluation of Double U-Net model in an intuitive manner on different groups of tongue images is presented. Finally, metrics considered to evaluate the performance of the applied architecture for quantitative analysis are elaborated.

1) *Dataset Description:* The tongue image dataset used for the experiment consists of 300 images captured by different Mobile phone in varied environment and locations. We used Labelme Software to annotate the image and generate a binary mask. Data augmentation of images and mask by centre cropping, horizontal flip and grid distortion was done to finally get augmented dataset of 636 images in total. The model was also tested on publicly available HIT dataset with 300 images of size 768 x576 bitmap images. First, we performed data augmentation on the available 300 images and mask by centre cropping, horizontal flip and grid distortion to generate augmented dataset. Dataset hereby is split into following sets training set consisting of 1188 images and validation and testing dataset 118 images each.
2) *Experimental Setup:* The experiments were performed with hp Pavilion laptop with a 1.60 GHz intel i5 8 th generation processor and 8 GB of Ram. Training of the model was done on Google Colab Python 3 Google Compute Engine backend GPU. Parameters set for training are learning rate set to α = 10-4, number of epochs N=300, Early stopping and Reduce LR on Plateau is also used, and applied ADAM Optimizer. Binary cross entropy loss was considered.
3) *Qualitative Evaluation:* To evaluate the robustness and effectiveness of the applied Double U-Net architecture various groups of tongue images were considered. Tongue images can be divided into various groups such as tongue not completely protruding, tongue with apparent gap in the mouth, tongue with teeth showing and tongue closely surrounded by lips and almost same intensity face area. Fig.2. shows the results of some samples falling in above mentioned groups. Double U-Net achieves a promising segmentation performance as indicative by the qualitative evaluation.

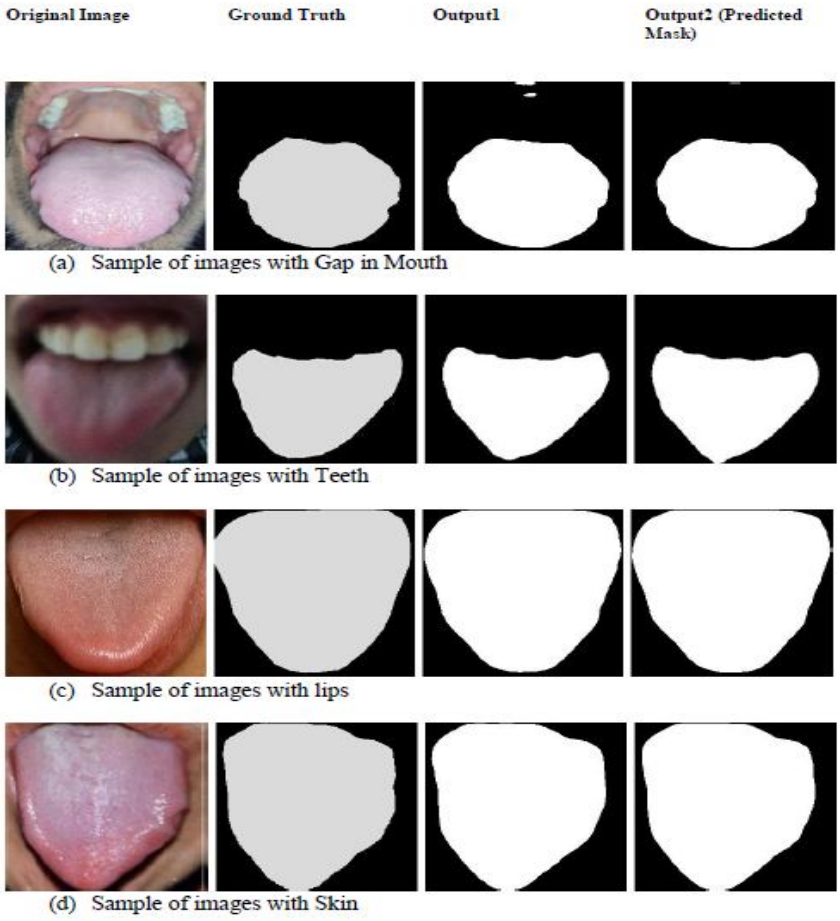


Fig. 2 Segmentation results of some samples of various image groups for Double U-Net

C. Quantitative Evaluation Metrics

Evaluation Metrics considered are IOU, Precision, Recall, and Dice Coefficient. A brief introduction to the metrics considered is given in brief. Then comparison of results achieved by Double U- Net with U-Net and Res U-Net is done based on above metrics as well as qualitative results in the subsequent section.

IOU -Intersection -Over- Union also known as Jaccard Index is the most common commonly used straightforward and extremely effective metric used in semantic segmentation. IOU is the area of overlap between the predicted segmentation 'B' and the ground truth 'A 'divided by the area of union between the predicted segmentation and the ground truth

$$IOU=(|A\cap B|)/(|A\cup B|)$$

The Purity of our positive detections relative to the ground truth describes Precision

$$Precision=TP/|TP+FP|$$

Where TP = True Positive

FP = False Positive

FN = False Negative

TN = True Negative

The completeness of our positive predictions relative to the ground truth is effectively described by Recall. Basically, it gives an idea of all of the annotated masks in our ground truth, how many were captured as positive predictions.

$$Recall=TP/|TP+TN|$$

The Dice coefficient is very similar to the IOU. They are positively correlated. Dice Score serves as means of validating an algorithm by calculating how similar the objects are. It is not only a measure of how many positives you find, but it also penalizes for the false positives that the method finds, similar to precision

$$Dice\ Coeffecient=(2\times TP)/(2TP+FP+FN)$$

IV. RESULTS

The comparison results of the three models on train and validation dataset, with the same dataset and hyperparameter settings gave results as shown in Table I. Results indicate that all three models are at par but Double U- Net shows minor improvement in validation metrics. Qualitative comparison of the models revealed that Double U-Net (Fig.2) surely indicated an upper hand with some typical images with lips, skin, gap, or teeth areas in close proximity.

TABLE I
Evaluation Metrics for the three models considered

Method	Training				Validation			
	IOU	Precision	Recall	Dice Coeff	IOU	Precision	Recall	Dice Coeff
Double U-Net	0.731592	0.995516	0.992643	0.844974	0.726155	0.992614	0.986376	0.841355
U-Net	0.711364	0.993138	0.98834	0.831316	0.693874	0.976928	0.972125	0.819224
Res U-Net	0.72753	0.993848	0.987717	0.84226	0.707698	0.983475	0.968638	0.828786

It is clearly visible from the plots of Dice Coefficient / Loss, for150 epochs for all the methods presented in Fig.3. that in case of Double U- Net attains a value of 84.5%. double U Net proves to be slightly better performer giving a precision of 99.3% at 150 epochs

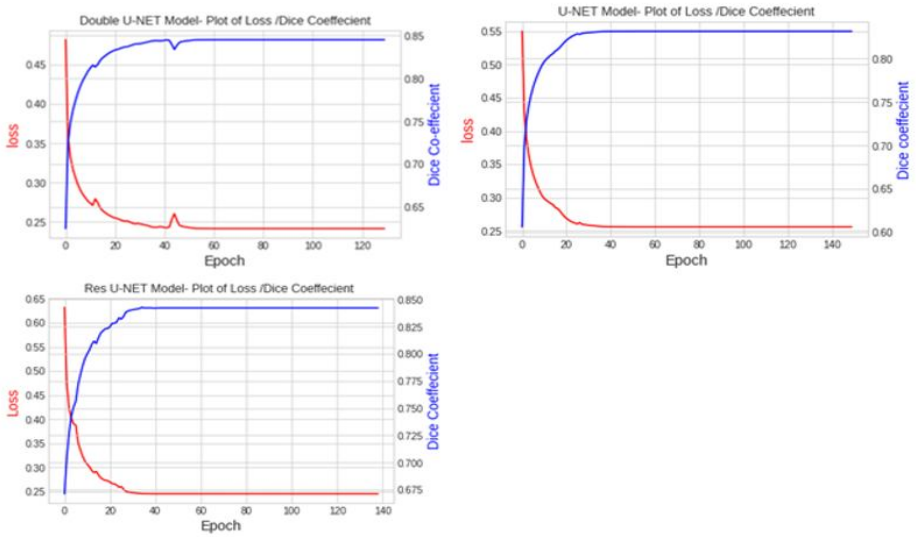


Fig. 3 Segmentation results of some samples of various image groups for Double U-Net

D. Robustness Testing

To further evaluate the robustness of Double U- Net model we conducted comparative experiments on new testing dataset. The new tongue image dataset taken from Harvard Dataverse, collected for study conducted in Yueyang Integrated Hospital of Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine consists of 96 tongue images taken in multiple lighting environments /locations.

Data Augmentation resulted in total 846 images by applying Vertical and Horizontal flip, Random Rotation, Optical Flip and Elastic Transformation. The comparative analysis for the three models under consideration resulted in proving the effectiveness of Double U-Net over U-Net and Res-U-Net. Similar experimental setup was utilized, Double-U-Net achieved an overall accuracy of 68.5% whereas Res- U-Net showed 48.3% accuracy and U-Net an accuracy of 48.7%.

Training time required for Double U-Net model was around 2.8 hrs. for 35 epochs, is indeed greater than the other two models under consideration due to greater number of trainable parameters (29,290,274), whereas U-Net taking minimum time of 52 minutes for 35 epochs and (412,865) trainable parameters. Res U-Net took 1.8 hrs. with (8,220,993) parameters to be trained. We expect to improve the training time further by investigating the effects of pre-processing the input images by enhancement techniques before applying to Double -U-Net architecture for future studies. Some typical image dataset with teeth, lips and poor-quality image are illustrated in the fig .4.

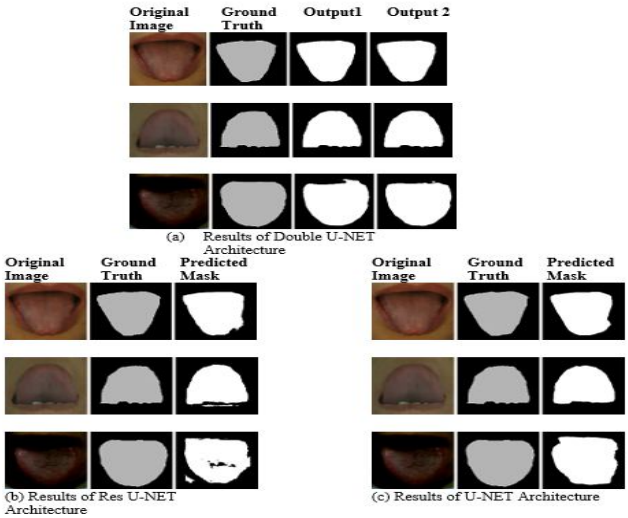


Fig. 4. Comparative Results of the three architectures on Dataverse dataset

V. DISCUSSION

Double U- Net Architecture shows robust performance as compared to U-Net and Res U-Net on all the datasets, specifically in case of dataverse as well as mobile captured dataset where images are not captured in standard conditions. Quantitative as well qualitative results show higher performance indices for double U net architecture. Qualitative results show that Double U-Net is capable of producing better segmentation mask even for the challenging images, thus suggesting robustness of the model. Limitation of Double U-Net is that it uses more parameters as compared to U-Net and Res U-Net, which leads to increase in training time as compared to other two models.

VI. CONCLUSIONS

In this paper we proposed a segmentation method based on Double U-Net. No pre or post processing was done on the dataset. Similar to U-Net prediction model which can work with small data set, Double U-Net performed rationally well with the small dataset under consideration. Quantitative evaluation no doubt showed a cut throat competition amongst the three architectures considered, Double U-Net proved to outperform in cases of challenging images. We believe that the segmentation results can be improved by incorporating some prepossessing on the input image so as to highlight the tongue boundary more effectively and by integration of different CNN Architectures along with pretrained networks for feature extraction. In future we plan to build a light weight model which could work equally well with all images irrespective of the acquisition system and ambience during image capture, for enabling automation of tongue analysis system for disease diagnosis.

VII. ACKNOWLEDGMENT

I acknowledge the support of Shri. Govindram Seksaria Institute of Technology & Science, for providing support with TEQIP Project and advanced technical facilities of CIDI (Centre for Innovation and Design Incubation).

REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation”, Computer Vision and Pattern Recognition (cs.CV), Cornell University, https://arxiv.org/abs/1505.04597v1.

[2] Debesh Jha, Michael A. Riegler, Dag Johanseny, Pal Halvorsen, Havard D. Johanseny, Simula Met, “Double U-Net: A Deep Convolutional Neural Network for Medical Image Segmentation”, Electrical Engineering and Systems Science - Image and Video Processing, Cornell University, https://arxiv.org/abs/2006.04868v2.

[3] Weihao Weng, Xin Zhu, “INet: Convolutional Networks for Biomedical Image Segmentation “DOI: 10.1109/ACCESS.2021.3053408, Volume 9,2021, IEEE Access.

[4] Jiang Li, Baochuan Xu, Xiaojuan Ban, Ping Tai, and Boyuan Ma “A Tongue Image Segmentation Method Based on Enhanced HSV Convolution Neural Network”, © Springer International Publishing AG 2017, Y. Luo (Ed.): CDVE 2017, LNCS 10451, pp. 252–260, 2017, DOI: 10.1007/978-3-319-66805-5_32.

[5] Yushan Xue, Xiaoqiang Li, Pin Wu1, Jide Li, Lu Wang,and Weiqin Tong,” Automated Tongue Segmentation in Chinese Medicine Based on Deep Learning”, © Springer Nature Switzerland AG 2018, ICONIP 2018, LNCS 11307, pp. 542–553, 2018, doi.:10.1007/978-3-030-04239-4_49.

[6] Bingqian Lin, Yanyun Qu_, Junwei Xie, Cuihua Li,” DEEPTONGUE: TONGUE SEGMENTATION VIA RESNET”, 978-1-5386-4658-8/18/$31.00 ©2018 IEEE, ICASSP 2018.

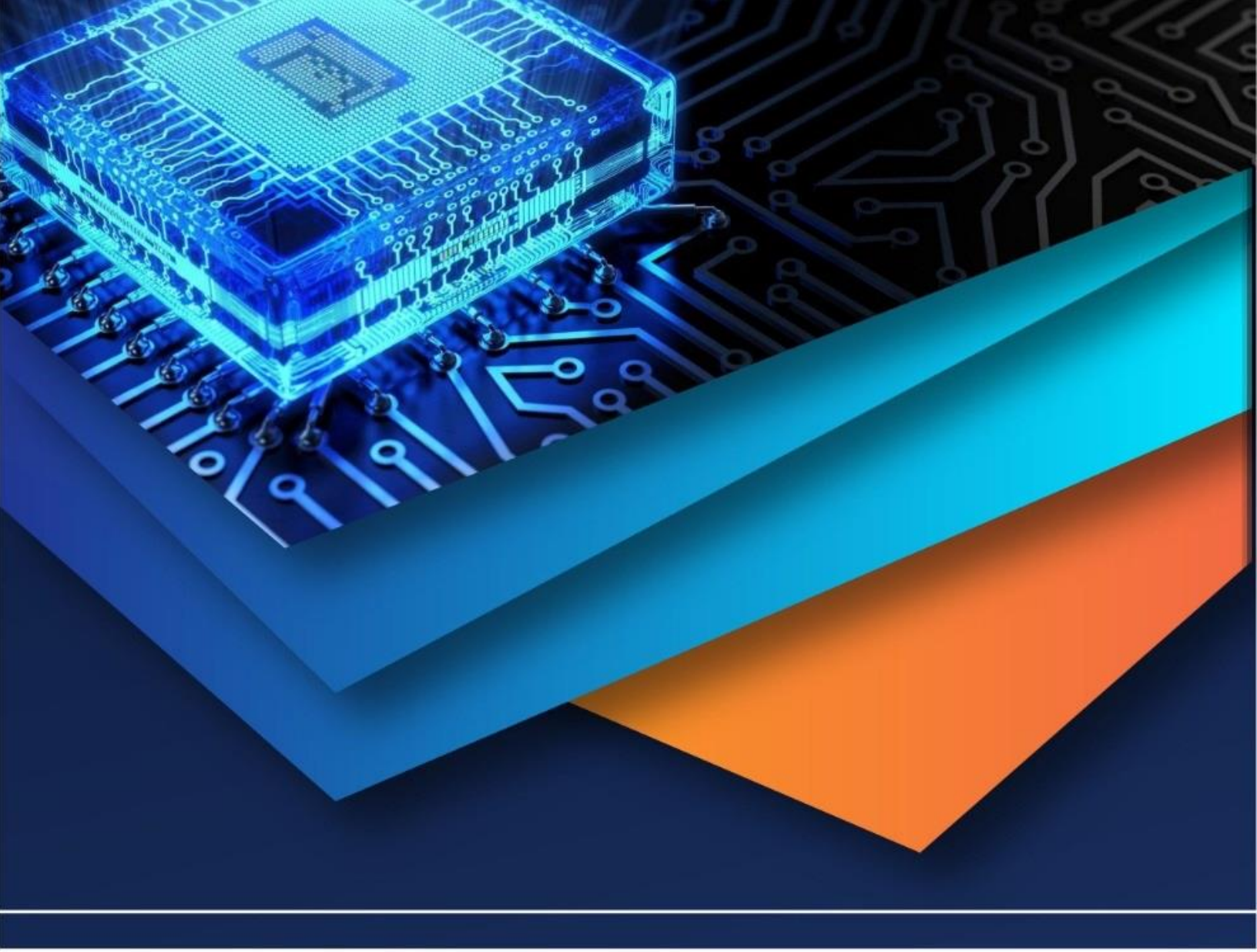
[7] Wei Yuan, Changsong Liu, “Cascaded CNN for Real-time Tongue Segmentation Based on Key Points Localization”, 2019 the 4th IEEE International Conference on Big Data Analytics,doi: 978-1-7281-1282-4/19/$31.00 ©2019 IEEE.

[8] Xinlei Li, Dawei Yang, Yan Wang, Shuai Yang, Lizhe Qi, Fufeng Li, Zhongxue Gan, Wenqiang Zhang,” Automatic Tongue Image Segmentation For Real-Time Remote Diagnosis”, 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),doi: 978-1-7281-1867-3/19/$31.00 ©2019 IEEE.

[9] Xiaodong Huang, Hui Zhang, Li Zhuo , Xiaoguang Li, and Jing Zhang, “TISNet-Enhanced Fully Convolutional Network with Encoder-Decoder Structure for Tongue Image Segmentation in Traditional Chinese Medicine”, Computational and Mathematical Methods in Medicine,Volume 2020, Article ID 6029258, doi.:10.1155/2020/6029258.

[10] Qichao Tang, Tingxiao Yang, Yuichiro Yoshimura,Takao Namiki, Toshiya Nakaguchi, “Learning-based tongue detection for automatic tongue color diagnosis system”, Artificial Life and Robotics (2020) 25:363–369,doi:10.1007/s10015-020-00623-5.

[11] Zhengxin Zhangy, Qingjie Liuy_, Yunhong Wang, “Road Extraction by Deep Residual U-Net”, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, https://arxiv.org/pdf/1711.10684.pdf



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY