Skin Lesion Segmentation using SegNet with Binary Cross-Entropy

Prashant Brahmbhatt[1], Siddhi Nath Rajan[2]

[1,2]IMS Engineering College, Ghaziabad, India

Abstract: In this paper a simple and computationally efficient approach as per the complexity has been presented for Automatic Skin Lesion Segmentation using a Deep Learning architecture called SegNet including some additional specifications for the improvisation of the results. The secondary objective is to keep the pre/post -processing of the images minimal. The presented model is trained on limited images from the PH2 dataset which includes dermoscopic images, manually segmented. It also contains their masks, the clinical diagnosis and the identification of several dermoscopic structures, performed by professional dermatologists. The aim is to achieve a performance threshold Jaccard Index (IOU) 92% after evaluation.

Keywords: Skin Lesion Segmentation, Dermoscopic, Jaccard Index

I. INTRODUCTION

As per the figures provided by the American Cancer Society, only 1% of all the cancers are diagnosed as melanoma cancers but they cause most of the skin cancer deaths. The non-melanoma cancers also cause of a large number of deaths. For the year 2019 the death estimation is 7,230 (4,740 men and 2,490 women) from melanoma [1]. As per the reports of World Health Organization around the world 3 million non-melanoma skin cancers and 1,32,000 melanoma skin cancers are recorded every year [2]. After the initial surgery most people are cured and the life expectancy is greatly increased if the cancer is timely diagnosed. Therefore, the proper diagnosis of the lesion is required for the treatment of a patient. To classify the lesion as melanoma or non-melanoma the images require to be segmented which creates a mask used for cropping the lesion then further identify the features from.

The traditional approach of manual segmentation by the expert dermatologists has been already proposed to be replaced with computerized segmentation techniques in which machine assisted methods are used. The previous manual approach was time-consuming, complex and dependent on the observer and his capabilities. The efficient approach of segmentation can save a lot of time and expertise required to complete the task with sufficient accuracy. The already proposed architectures involve some state-of- the-art neural networks like FCN [3], VGG [4], Fast-RCNN [5] and U-Net [6].

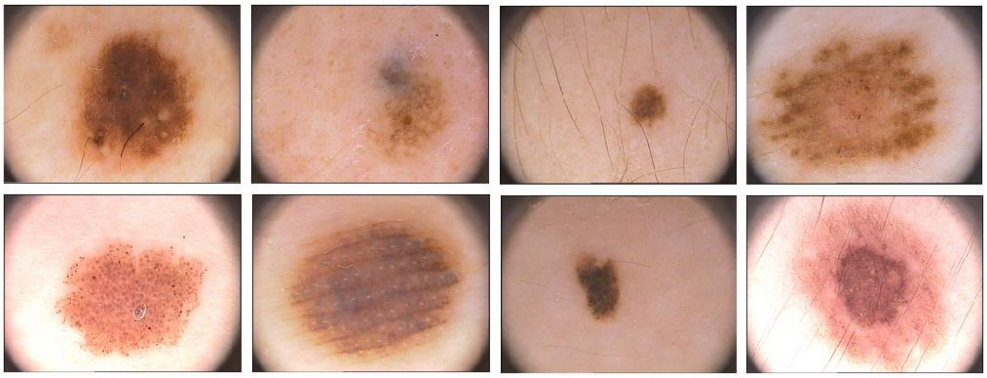


Fig. 1. Sample microscopic images of the skin lesions found on the human skin.

The Model

SegNet [7] is a deep neural network based on the encoder-decoder architecture for pixel-wise segmentation, originally proposed and implemented by members of the Computer Vision and Robotics Group at the University of Cambridge, UK [8]. SegNet is based on the principle of Semantic Pixel-Wise Labeling initially used for the segmentation of street images which include a total of twelve different classes. Each pixel in the image is to be classified among one of the

twelve classes as per the data.

The architecture of SegNet has non-linear layer encoding sequences and decoders corresponding to each layer. There is a final classifier present for the pixel-wise classification. We have used it as the base architecture with varying hyper parameters and structural differences for our lesion segmentation problem.

Network Architecture

The proposed is a 64-layered network excluding the final activation layer. Every sequence of encoder has multiple convolutional layers, batch normalized with ReLU non-linearity which is followed by non-overlapping maxpooling and sub-sampling. At the center of the network there are two dense layers present before the first up-sampling begins.

The defining characteristic of SegNet is the use of max-pooling indices in the decoders to perform up-sampling of low-resolution feature maps. This leads to retaining of the important detailed features in the image and non-useful features are dropped. SegNet provides smooth images without any post-processing technique involved.

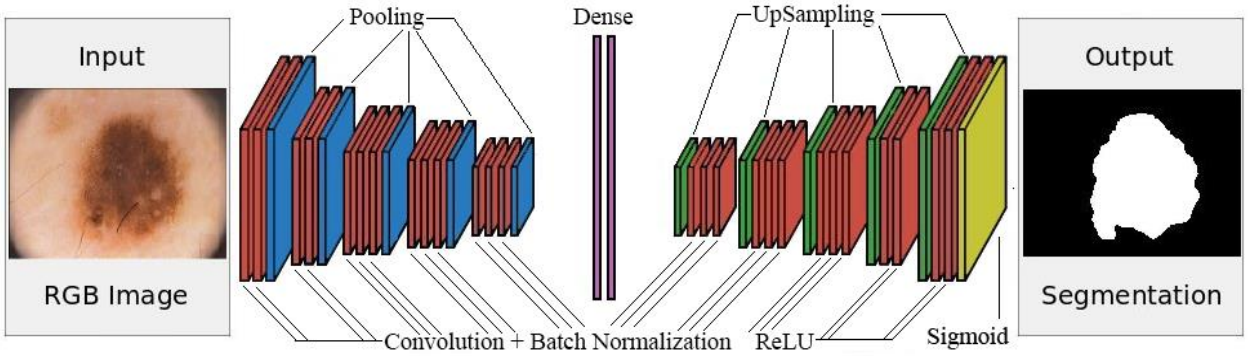


Fig. 2. The proposed network architecture shows all the layers of the network

Table 1. The detailed network architecture with output dimensions at each sequence of layers

Sequences	Filter	Output Dim.	Sequences	Filter	Output Dim.
Input		192 x 256 x 3	u-1		
conv-1	(3,3)	192 x 256 x 64	deconv-1		
bn-1			bn-14	(3,3)	24 x 32 x 512
conv-2	(3,3)	96 x 128 x 128	deconv-2		
bn-2			bn-15		
p-1			deconv-3	(3,3)	24 x 32 x 512
			bn-16		
conv-3	(3,3)	96 x 128 x 128			
bn-3			u-2		
conv-4	(3,3)	48 x 64 x 256	deconv-4		
bn-4			bn-17	(3,3)	24 x 32 x 156
p-2			deconv-5		
			bn-18		
conv-5	(3,3)	48 x 64 x 256	deconv-6	(3,3)	48 x 64 x 256
bn-5			bn-19		
conv-6					
bn-6			u-3		
conv-7	(3,3)	24 x 32 x 512	deconv-7		
bn-7			bn-20	(3,3)	48 x 64 x 256
p-3			deconv-8		
			bn-21		
conv-8	(3,3)	12 x 16 x 512	deconv-9	(3,3)	96 x 128 x 128
bn-8			bn-22		
conv-9					

bn-9			u-4	(3,3)	96 x 128 x 64
conv-10			deconv-10		
bn-10			bn-23		
p-4			deconv-11	(3,3)	96 x 128 x 64
			bn-24		
conv-11	(3,3)	6 x 8 x 512			
bn-11			u-5	(3,3)	192 x 256 x 64
conv-12			deconv-12		
bn-12			bn-24		
conv-13			deconv-13	(3,3)	192 x 256 x 1
bn-13			bn-25		
p-5					
Dense-1		6 x 8 x 1024			
Dense-2		12 x 16 x 1024	Output		192 x 256

Loss Function

The loss function used here is the binary cross-entropy. The cross-entropy is a function which measures how far away from the true value the prediction is for each of the classes and then averages the errors class wise to obtain the final loss.

In this problem, there lies only two classes for each pixel, either black or white (0 or 1) as per the mask. So, here binary cross-entropy is used as the loss function rather than the categorical cross-entropy originally proposed.
The binary cross-entropy is in the below form:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=0}^{N}(y * \log(\hat{y}_i) + (1-y) * (1-\hat{y}_i))$$

Training

For the training procedure has been carried out on 75% of the images available out of the 200 images in the PH2 dataset [9]. Although, the actual number of images will be more than 150 because of the new transformed images that will be added after image augmentation process. As per the architecture of the network the total parameters to be trained are 33,377,795 out of 33,393,669 whereas the non-trainable parameters are 15,874. The implementations are in Keras and the environment used is the IPython notebook provided by Kaggle [10].

Image Augmentation

The procedure of image augmentation on training images has been used for increasing the robustness of the model and reducing the chances of overfitting. It will also increase the data images that are available in the dataset. The two simple techniques that used are image rotation and horizontal flipping [11]. In the image rotation all the images in the training set are rotated with a range [-40◦, +40◦] [3] and flipped along the horizontal axis only.

All the above transformations are exactly performed on the corresponding masks of the images as well to maintain the correct orientation of feature images with their truth masks. After the augmentation the transformed images are included in the training set which increases our training set from 150 to 450. Out of these 450 images, 90 have been excluded from training set for the formation of a validation set.

Optimizer

The employed optimizer is the SGD (Stochastic Gradient Descent) for the network. The learning rate which is an important hyper parameter in the optimization is set to 0.001 which is one of the many values generally used for learning rate parameter.

The momentum is also used which is an approach that provides an update rule motivated from the physical perspective of optimization. The advantages of using momentum with SGD are that a small change results in large speed up in learning process. Analogically, the velocities are stored for all the parameters, and used for making the updates. The value of momentum used in for optimization is 0.9.

Batch Normalization

Batch Normalization [12] is the technique of speeding up the learning process of the neural network by normalizing the values in the hidden layers similar to the principle behind the normalization of the features in the data or activation values. In proposed network the batch normalization layer is present after every convolution layer with a total of 25 batch normalization layers in the entire network architecture.

II. EXPERIMENTAL DESIGN AND RESULTS

Databases

The used dataset is the PH2 dermoscopic dataset which contains 200 dermoscopic images and their label masks. Each one is an RGB image and the fixed dimension of each image is 572 x 765. The dataset has been provided publicly for experimental and studying purposes, to facilitate research on both segmentation and classification algorithms of dermoscopic images. The database is acquired at the Dermatology Service of Hospital Pedro Hispano, Matosinhos, Portugal [9].

For the training purpose the dimensions of each image have been reduced to 192 x 256 before feeding it into the network. It largely reduces the parameters to be trained in the network as well as the training time and complexity without significantly affecting the results.

Performance Evaluation

The generated binary mask images in the output of the network are evaluated on different mathematical measures in comparison with the ground truth lesion masks as provided in the dataset. The accuracy is measured pixel-wise. The used measures are as below:

- Intersection Over Union: The Jaccard index, also known as Intersection over Union. The Jaccard similarity coefficient is a statistical similarity measure to check the diversity among the sample sets. The IOU gives the similarity among sets and the formula is the size of the intersection over the size of the union of the sets.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

- Dice Coefficient: The Dice score is like precision. It measures the positives as well as it applies penalty to the false positives given by the model. It is more similar to precision than accuracy.

$$Dice = \frac{2 \times TP}{(TP + FP) + (TP + FP)}$$

- Precision: Precision is a measure which is more focused towards catching the false positives in the results of the model.

$$Precision = \frac{TP}{TP + FP}$$

- Recall: Recall is a measure which is targeted towards the actual or the true positives yielded by the model output. In the scenarios where the cost of the False Negatives is greater than recall is the better metric to choose the best model among the possible ones.

$$Recall = \frac{TP}{TP + FN}$$

- Accuracy:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Key Component Validation

The network after training on the set of 360 images and validating on the 90 images produced the results that are observed after 100 epochs.

The training curves of the network corresponding to the training set as well as the validation set are also plotted. The curves include the loss curve and the accuracy curve with respect to the epochs along the horizontal axis. Initially for

the training set the loss is above 0.735 which gradually declined and reached 0.115. For the validation set it began from 0.707 and ended up at 0.1595. The accuracy for the training set increased from 0.500 to 0.978 and that of the validation set from 0.312 to 0.955.

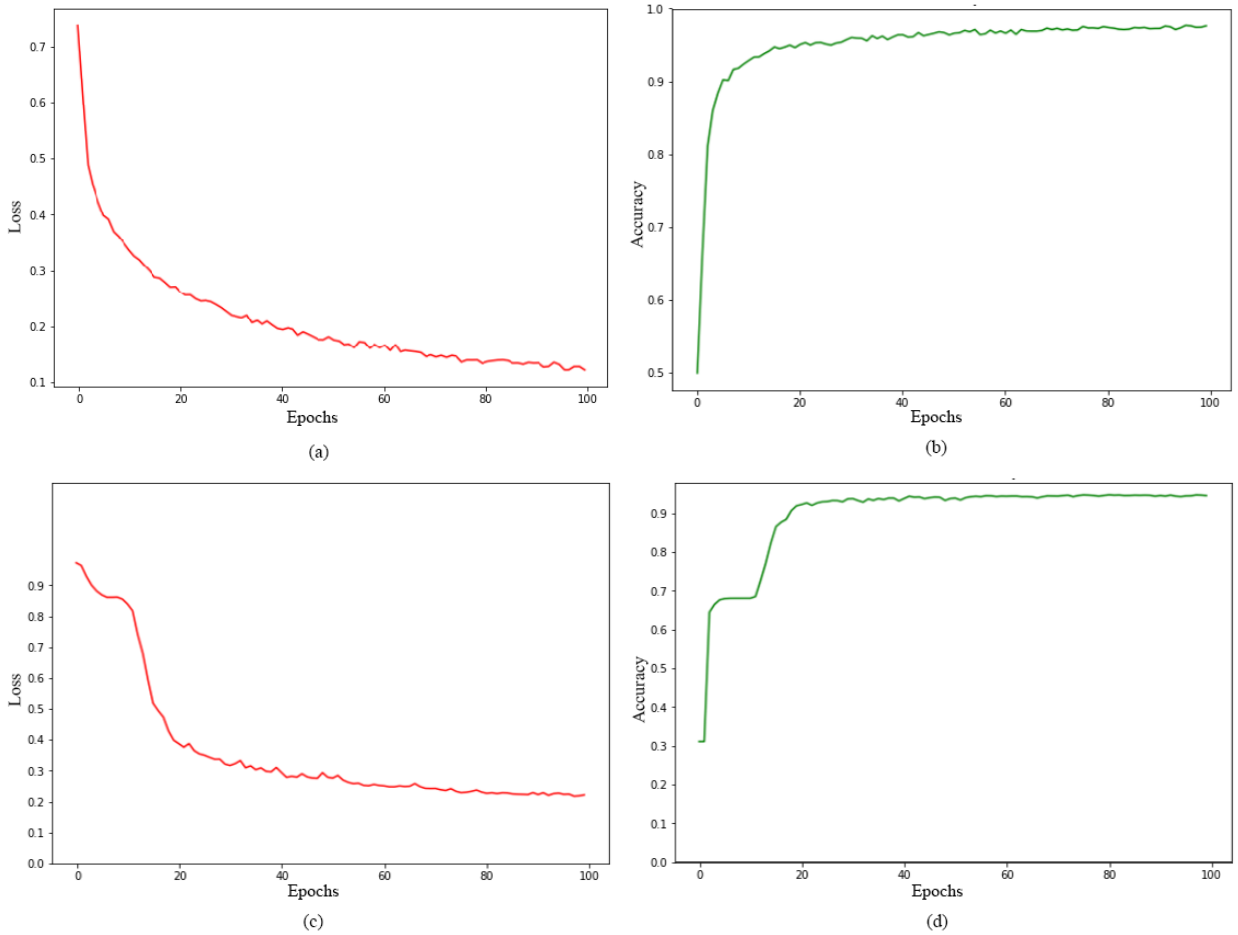


Fig. 3. The training and validation curves (a) Loss curve on training set; (b) Accuracy curve on training set; (c) Loss curve on validation set; (d) Accuracy curve on validation set

Table 2. Performance statistics after training for 1 epoch

Dataset	JA	DI	PR	RE	AC	Loss
Test	67.71	38.35	6	1.5	60.86	69.21
Validation	67.66	38.17	7.47	1.98	61.41	69.21
Training	67.86	38.72	6.68	1.65	60.2	69.23

Table 3. Performance statistics after training for 100 epochs

Dataset	JA	DI	PR	RE	AC	Loss
Test	93.61	80.37	88.99	92.13	93.96	18.75
Validation	95.09	82.1	91.84	94.19	95.53	15.95
Training	97.02	85.16	95.97	97.32	97.87	11.5

Prediction on PH2 Database

The final part is to save the trained model and make predictions. The predictions are compared visually to the actual ground truth lesion mask images. The predicted outputs are initially slightly blurry at the edges and do not give a precise prediction towards the boundaries however, it still performs considerably well. To have a clear prediction at the boundary thresholding of the pixel values is performed as a simple post-processing technique.

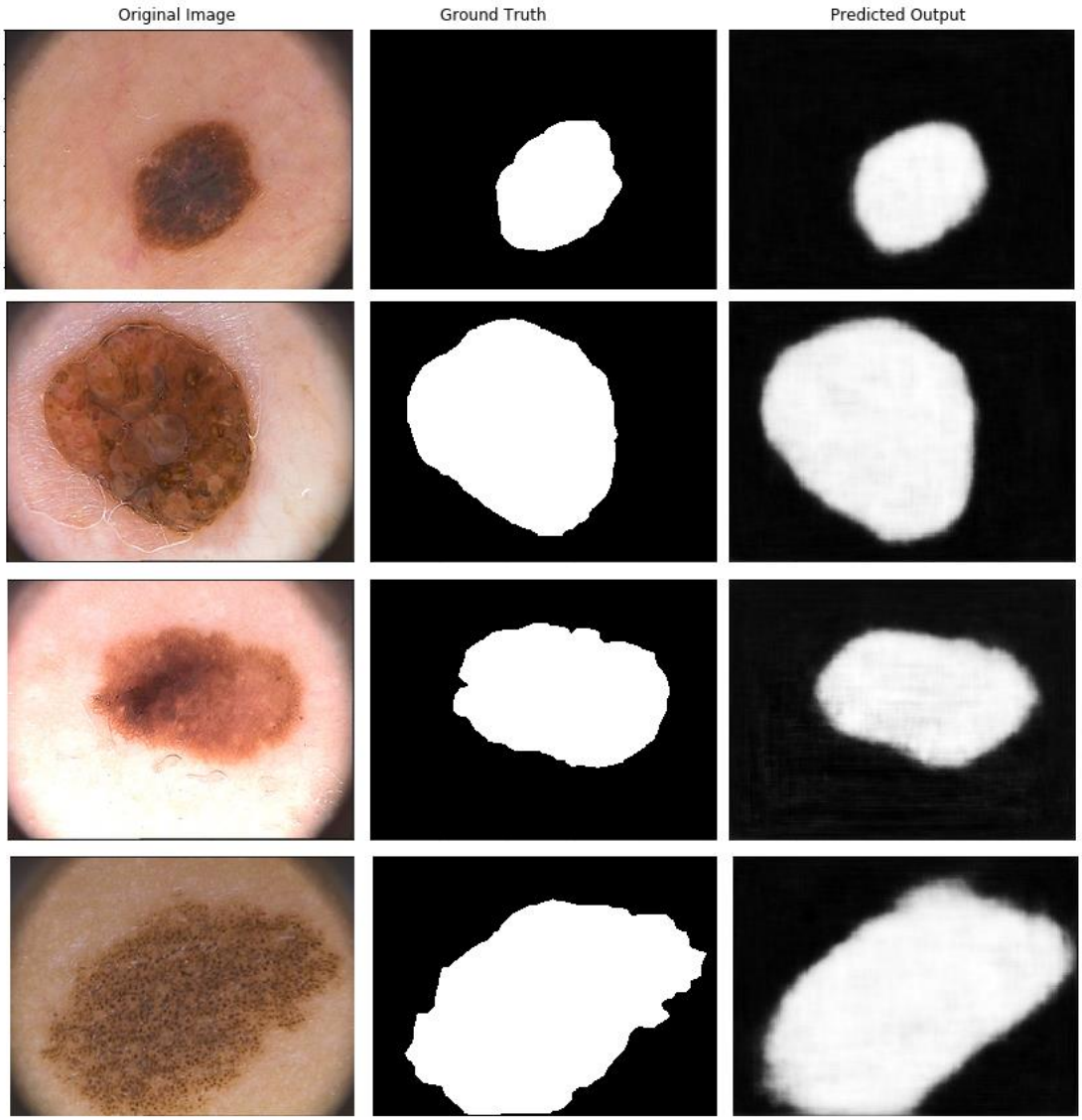
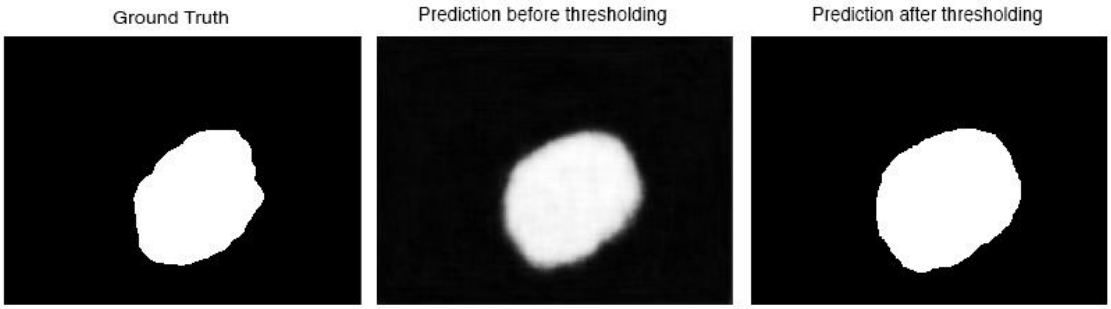


Fig. 4. The comparison of the predictions with original feature images and the ground truths

The predicted images have every pixel value lying in the range [0,1]. To get the clear boundaries the pixel values can be converted to belong in either the black or white class (0 or 1) based on a decided threshold which eliminates any pixel values in the gray shade. The threshold that we have used is 0.2, every pixel having a prediction value of greater than 0.2 will be rounded up to 1 and vice-versa. After performing this technique, the sharp boundaries around lesion in the predicted masks can be observed.



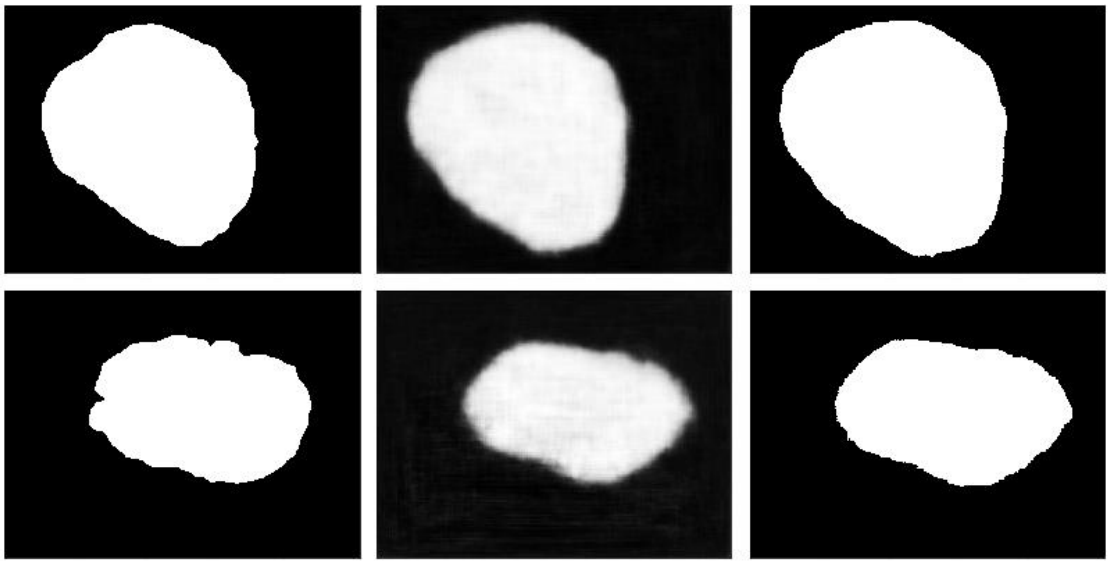


Fig. 5. The comparison of ground truths and original predictions with processed predictions

III. DISCUSSION

The proposed approach and specifications in this paper are only applied to the PH2 dataset and could be used on the ISBI 2016 [13] and ISIC 2017 [14] dataset to fairly compare it with the current state-of-the-art architectures. The advanced approaches of image augmentation such as varying the brightness and contrast or using affine, random crops [11] and levelsets [15] could improve the robustness but it would increase the computational complexity for much larger datasets like ISIC 2017. The larger datasets will significantly increase the training time as they may require more than 100 epochs. The proposed network in this paper took 9.25 seconds per step over 100 epochs resulting in the total duration of 15.416 minutes.

The thresholding value used is chosen merely based on the experimental results and has no theoretical principle behind the exact value. The different image sizes could also be experimented which will increase the time of the training by a large amount but could be needed for the larger datasets with more than 2000 images as cited previously.

IV. CONCLUSION

In this paper the authors proposed the use of the SegNet architecture for solving the skin lesion segmentation and successfully provided with the results of the experiment on the PH2 dataset with sufficient accuracy.

Two techniques of image augmentation, image rotation and horizontal flipping on the training dataset are performed before feeding it to the network for training. After the training process the model was evaluated on several measures for statistical values. The predictions produced from the model on test images were post-processed using the thresholding technique to remove the blurry boundaries around the predicted lesions.

REFERENCES

[1] "Melanoma: Statistics | Cancer.net," American Society of Clinical Oncology (ASCO), 2019. [Online]. Available: American Society of Clinical Oncology (ASCO). [Accessed July 2019].

[2] WHO, "WHO | Skin Cancer," World Health Organisation, 2019. [Online]. Available: https://www.who.int/uv/faq/skincancer/en/index1.html. [Accessed July 2019].

[3] Y. Yuan, M. Chao, Y.C. Lo, "Automatic Skin Lesion Segmentation Using Deep Fully Convolutional Networks With Jaccard Distance," IEEE Transactions on Medical Imaging, 2017.

[4] Lopez, A. Romero and Giro-i-Nieto, Xavier and Burdick, Jack and Marques, Oge, "Skin lesion classification from dermoscopic images using deep learning techniques," in Biomedical Engineering (BioMed), 2017 13th IASTED International Conference on, IEEE, 2017, pp. 49--54.

[5] R. Girshick, "Fast R-CNN," in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015.

[6] Md. Z.Alom, M. Hasan, C. Yakopcic, T. M. Taha, V.K. Asari, "Recurrent Residual Convolutional Neural," Cornell University, 29 May 2018. [Online]. Available: https://arxiv.org/abs/1802.06955v5. [Accessed July 2019].

[7] A.Kendall, V. Badrinarayanan, R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[8] A.Kendall, V. Badrinarayanan, R. Cipolla, "SegNet," University of Cambridge, 2015. [Online]. Available: http://mi.eng.cam.ac.uk/projects/segnet/. [Accessed July 2019].

[9] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, J. Rozeira, "PH2 - A dermoscopic image database for research and benchmarking," 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5437-5440, 2013.

[10] B. Hammer, A. Goldbloom, "Kaggle," Alphabet Inc., April 2010. [Online]. Available: https://www.kaggle.com/. [Accessed October 2019].

[11] F. Perez, C. Vasconcelos, S. Avila, E. Valle, "Data Augmentation for Skin Lesion Analysis," Cornell University, 05 September 2018. [Online]. Available: https://arxiv.org/abs/1809.01442. [Accessed July 2019].

[12] S. Ioffe, C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," 11 February 2015. [Online]. Available: https://arxiv.org/abs/1502.03167. [Accessed July 2019].

[13] D. Gutman, N. C. F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, A. Halpern, "Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC)," May 2016. [Online]. Available: https://arxiv.org/abs/1605.01397.

[14] D. Gutman, N. C. F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, A. Halpern, "Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)," 2017. [Online]. Available: https://challenge.kitware.com/#challenge/583f126bcad3a51cc66c8d9a.

[15] K. H. Cha, L. Hadjiiski, R. K. Samala, H. P. Chan, E. M. Caoili, R. H. Cohan, "Urinary Bladder Segmentation," Med Phys, vol. 43, no. 4, p. 1882–1896., 2016.