Automatic Lesion Boundary Segmentation in Dermoscopic Images with Ensemble Deep Learning Methods

Manu Goyal, *Student Member, IEEE,* and Moi Hoon Yap, *Member, IEEE,*

***Abstract*—Early detection of skin cancer, particularly melanoma, is crucial to enable advanced treatment. Due to the rapid growth of skin cancers, there is a growing need of computerized analysis for skin lesions. These processes including detection, classification, and segmentation. The state-of-the-art public available datasets for skin lesions are often accompanied with very limited amount of segmentation ground truth labeling as it is laborious and expensive. The lesion boundary segmentation is vital to locate the lesion accurately in dermoscopic images and lesion diagnosis of different skin lesion types. In this work, we propose the use of fully automated deep learning ensemble methods for accurate lesion boundary segmentation in dermoscopic images. We trained the Mask-RCNN and DeeplabV3+ methods on ISIC-2017 segmentation training dataset and evaluate the various ensemble performance of both networks on ISIC-2017 testing set, PH2 dataset. Our results showed that the proposed ensemble method segmented the skin lesions with Jaccard index of 79.58% for the ISBI 2017 test dataset. In comparison to FrCN, FCN, U-Net, and SegNet, the proposed ensemble method outperformed them by 2.48%, 7.42%, 17.95%, and 9.96% for the Jaccard index, respectively. Furthermore, the proposed ensemble method achieved a segmentation accuracy of 95.6% for some representative clinical benign cases, 90.78% for the melanoma cases, and 91.29% for the seborrheic keratosis cases in the ISBI 2017 test dataset, exhibiting better performance than those of FrCN, FCN, U-Net, and SegNet.**

***Index Terms*—Skin lesion, Melanoma, Convolutional neural networks, Transfer learning**

I. Introduction and Background

SKIN cancer is the most common cancer among all other cancers [1]. The malignant skin lesions consist of the melanocytic lesion, i.e. melanoma, and non-melanocytic lesion, i.e. basal cell carcinoma. Although melanoma is the least common type of skin cancer, it is the most aggressive and deadly cancer. In addition, it has the high capacity to invade tissues and other organs [2]. Hence, it is important to have early detection to save a life. According to the prediction of Melanoma Foundation [3], the estimated new cases of melanoma in the United States is 87,110 (200% increased since 1973) with 9,730 predicted deaths.

In current medical practice, skin cancer specialists primarily examine the patients on visual inspection with manual measurements tools called dermoscopy assessment to determine the skin lesions. Relying on self-vigilance and medical inspection by human vision risk life and survival rate as it is difficult to identify the type of lesions by naked eyes. Hence, over the years, different image modalities were used to inspect the skin, including macroscopy (clinical) and microscopy (also known as Dermoscopy). Dermoscopy is a non-invasive imaging that allows visualisation of skin surface by the light magnifying device and immersion fluid [4]. It is one of the most widely used imaging techniques in dermatology and it has increased the diagnosis rate [5].

In current medical practice, dermatologists primarily examine the patients on visual inspection with manual measurements tools called dermoscope to differentiate between benign and malignant skin lesions. Relying on self-vigilance and medical examination by human vision risk life and survival rate as it is difficult to identify the type of skin lesions by naked eyes. Dermoscopy is non-invasive imaging that allows visualization of skin surface by the light magnifying device and immersion fluid. It is one of the most widely used imaging techniques in dermatology, and it has increased the diagnosis rate. Dermatologists rely on ABCD rule for lesion diagnosis to differentiate the benign skin lesions from skin cancer as shown in Fig. 1. The ABCD rule for diagnosis as follows.

1) A: Asymmetry property checks whether two halves of the skin lesion match or not in terms of color, shape, edges. In general practice, the skin lesions are divided into two halves based on long axis and short axis as shown in the Fig. 1. In the case of skin cancer, it is likely to have asymmetrical appearance.
2) B: Border property defines whether the edges of skin lesion are smooth, well defined or not. In the cases of skin cancer, edges are likely to be uneven, blurry, and jagged.
3) C: Color property means that the colors throughout the skin lesions are not same. The color in cancerous skin lesion especially melanoma, vary from one area to the another, and usually has different shades such as tan, brown, red, black.
4) D: Diameter property measures the approximate diameter of the skin lesion. The diameter of skin lesion especially melanoma cases is generally greater than 6mm (the size of pencil eraser).

End-to-end computerized solutions that can produce accurate segmentation of skin lesions irrespective of types of skin lesions are highly desirable to aid ABCD Rule. For seg-

M. Goyal and M.H. Yap are with the School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Manchester, UK.
E-mail: manu.goyal@stu.mmu.ac.uk
Manuscript received xxxxx xx, 2017; revised xxx xx, 201x.

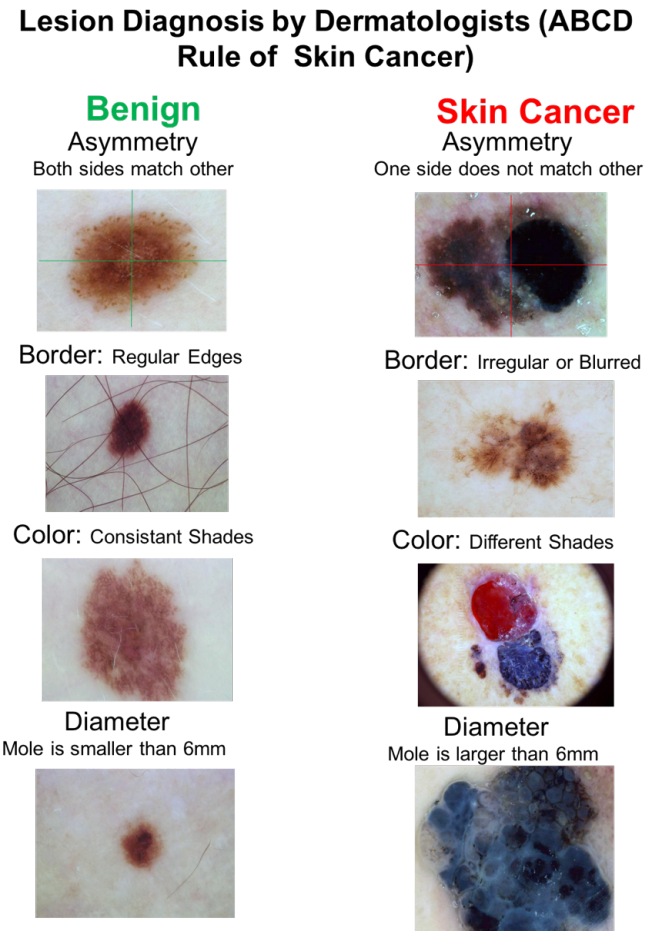


Fig. 1. ABCD Criteria for lesion diagnosis focuses on finding the certain properties of lesions

mentation of medical imaging, *dice*, *specificity* and *sensitivity* are deemed as important performance measures for methods. Hence, computerized methods need to achieve high scores in these performance metrics.

The majority of the state-of-the-art computer-aided diagnosis on dermoscopy images composed of multi-stages, which include image pre-processing, image segmentation, features extraction and classification [6], [1]. Using hand-crafted feature descriptors, the dermatologists are able to differentiate benign lesions based on their shape features as they normally have small dimensions and more circular. Other feature descriptors used in previous works including asymmetry features, color features and texture features. Pattern analysis was widely used to describe the appearance of skin lesions, this including the melanocytic algorithm elaborated by Argenziano et al. [7]. Using the features descriptors and pattern analysis, machine learning algorithms were used to classify the lesions types. There are many research in developing computerized methods based on image processing and conventional machine learning approaches. These research were presented in two survey papers where the majority of the approaches were using hand-crafted features to classify or segment the lesions, the earlier review was in 2012 reviewed by Korotkov et al. [6] and the later was conducted by Pathan et al. [1]. Korotkov et al. [6] concluded that there is a large discrepancy in previous research and the computer-aided diagnosis systems were not ready for implementation. The other issue was the lack of benchmark dataset which makes it harder to assess the algorithms. Pathan et al. [1] concluded that the CAD systems worked for experimental settings but subject to rigorous validation in real-world clinical settings.

However, recent advancement in computer vision algorithms especially with latest deep learning methods, the image recognition has vastly improved in few years. These systems are capable of providing end-to-end solutions with great accuracy in the medical imaging such as Magnetic Resonance Imaging (MRI), dual-energy X-ray absorptiometry, ultrasonography, and computed tomography [8], [9], [10], [11], [12]. There are further advancement in medical imaging field comprises of dermatology, the evaluation of skin such as facial skin in face images and diabetic foot ulcers in foot images is made possible with these algorithms [13], [14], [15], [16]. In near future, these systems on mobile devices can aid dermatologists and patients to improve skin lesion diagnosis [17], [18]. Recently, the *International Skin Imaging Collaboration (ISIC)* started to organize annual challenges for skin lesion segmentation and classification to push researchers in this field to produce more accurate and robust methods for lesion diagnosis [19], [20], [21].

With the rapid growth of deep learning approaches, many researchers [22], [23], [24] have proposed Deep Convolutional Neural Networks for melanoma detection and segmentation.

In this work, we propose the novel ensemble of state-of-the-art methods for segmentation that are Deeplabv3+ (semantic segmentation) and Mask-RCNN (instance segmentation) for accurate lesion boundary segmentation on dermoscopic images of ISIC Challenge 2017 dataset [25], [26], [20]. Then, we test the robustness of trained algorithms on other completely unseen publicly available datasets, i.e. PH2 dataset. We compared the performance of our proposed methods with other popular segmentation methods as well as competition winners. Ensemble methods are very popular to produce best results in the classification tasks. According to our best knowledge, this is the first time, ensemble methods for skin segmentation are proposed. The last year ISIC Challenge 2018 did not make the dataset public [21].

II. Deep Learning for Skin Lesion Segmentation

Deep learning has gained popularity in medical imaging research including Magnetic Resonance Imaging (MRI) on brain [27], breast ultrasound cancer detection [28] and diabetic foot ulcer classification and segmentation [14], [15]. A popular deep learning approach in biomedical imaging research is U-Net, proposed by Ronneberger et al. [29]. U-Net enables the use of data augmentation, including the use of non-rigid deformations, to make full use of the available annotated sample images to train the model. These aspects suggest that the U-Net could potentially provide satisfactory results with the limited size of the biomedical datasets currently available. An up-to-date review of conventional machine learning methods is presented in [1]. This section reviews the state-of-the-art deep learning approaches for segmentation for skin lesions.

Researchers have made significant contributions proposing various deep learning frameworks for the detection of skin lesions.

Yu et al. [23] proposed very deep residual networks of more than 50 layers for two-stage framework of skin lesions segmentation followed by classification. They claimed that the deeper networks produce richer and more discriminative features for recognition. By validating their methods on ISBI 2016 *Skin Lesion Analysis Towards Melanoma Detection Challenge* dataset [30], they reported that their method ranked first in classification when compared to 16-layer VGG-16, 22-layer GoogleNet and other 25 teams in the competition. However, in segmentation stage, they ranked second in segmentation among the 28 teams. Although the work showed promising results, but the two-stage framework and very deep networks are computationally expensive.

Bi et al. [24] proposed a multi-stage fully convolutional networks (FCNs)for skin lesions segmentation. The multi-stage involved localised coarse appearance learning in the early stage and detailed boundaries characteristics learning in the later stage. Further, they implemented a parallel integration approach to enable fusion of the result that they claimed that this has enhanced the detection. Their method outperformed others in PH2 dataset [31] of 90.66% but achieved marginal improvement if compared to Team ExB in ISIB 2016 competition with 91.18%.

Yuan et al. [22] proposed an end-to-end fully automatic method for skin lesions segmentation by leveraging 19-layer DCNN. They introduced a loss function using Jaccard Distance as the measurement. They compared the results using different parameters such as input size, optimisation methods, augmented strategies, and loss function. To fine tune the hyper-parameters, 5-fold cross-validation with ISBI training dataset was used to determine the best performer. Similar to Bi et al. [24], they evaluated their results on ISBI 2016 and PH2 dataset. The results were outperformed the state-of-the-art methods but they suggested that the method achieved poor results in some challenging cases including images with low contrast.

Goyal et al. [17] proposed fully convolutional methods for multi-class segmentation on ISBI challenge dataset 2017. This was a very first attempt to perform multi-class segmentation to distinguish melanocytic naevus, melanoma and seborrhiec keratoses rather than single class of skin lesion.

The research showed that deep learning achieved promising results for skin lesions segmentation and classification. However, these methods did not make their codes available and not validated on the ISBI 2017 dataset, which has 2000 images compared to 900 in ISBI 2016 dataset.

III. Methodology

This section discusses the publicly available skin lesion datasets, the preparation of the ground truth, and the performance measures to validate our results.

A. Publicly Available Skin-Lesion Datasets

For this work, we used three publicly available datasets for skin lesions that are ISBI-2017 Challenge (Henceforth ISBI-2017), and PH2 datasets. PH2 has 200 images in which 160 images are naevus (atypical naevus and common naevus), and 40 images are of melanoma [31]. ISBI-2017 is a subset of ISIC Archive dataset [20]. In segmentation category, it consists of 2750 images with 2000 images in training set, 150 in validation set and 600 in the testing set. Since the ground truths for the challenges only included the segmentation mask, to produce the ground truth for ROI, we circumscribed a rectangle bounding box on the segmentation mask as shown in the Fig. 2. We solely used ISBI-2017 to train the models. We resized all the images to 500 $\times$ 375 to improve the performance and reduce the computational costs.

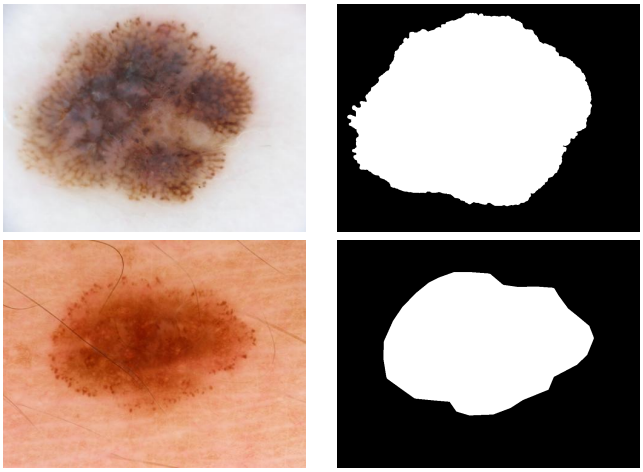


Fig. 2. Examples of Segmentation Masks (a) Original Image (b) Segmentation Masks

B. Ensemble Methods for Lesion Boundary Segmentation

We designed this end-to-end ensemble segmentation method to combine Mask-RCNN and DeeplabV3+ with pre-processing and post-processing method to produce accurate lesion segmentation as shown in the Fig. 3. This section describes each stage of our proposed ensemble method.

1) Pre-Processing: ISIC Challenge dataset comprised of dermoscopic skin lesion taken by different dermoscope and camera devices all over the world. Hence, it is important to perform pre-processing for color normalization and illumination with color constancy algorithm. We processed the dataset with Shades of Gray algorithm [33] as shown in the Fig. 4. We used pre-processing for both Mask-RCNN and DeeplabV3+ methods.

2) DeeplabV3+: We trained DeepLabV3+ with default setting on skin lesion datasets which is one of the best performing semantic segmentation networks [25]. It assigns semantic label lesion to every pixel in an dermoscopic image. DeeplabV3+ is an encoder-decoder network which makes the use of convolutional neural network called Xception-65 with atrous convolution layers to get the coarse score map and then, conditional random field is used to produce final output as shown in the Fig. 5.

3) Mask-RCNN: We fine-tuned Mask-RCNN with ResNet-InceptionV2 (Mask-RCNN) for single class as skin lesion for

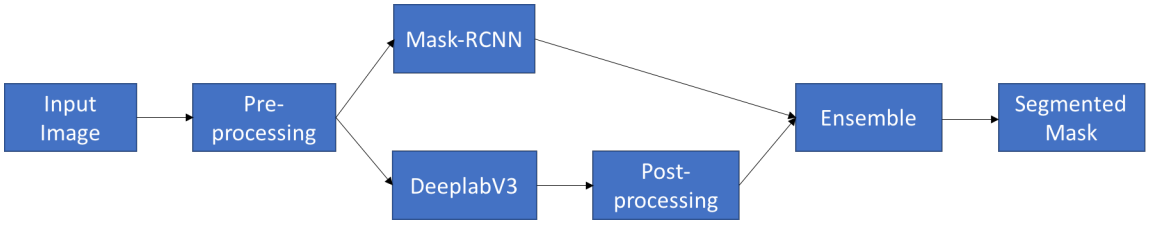
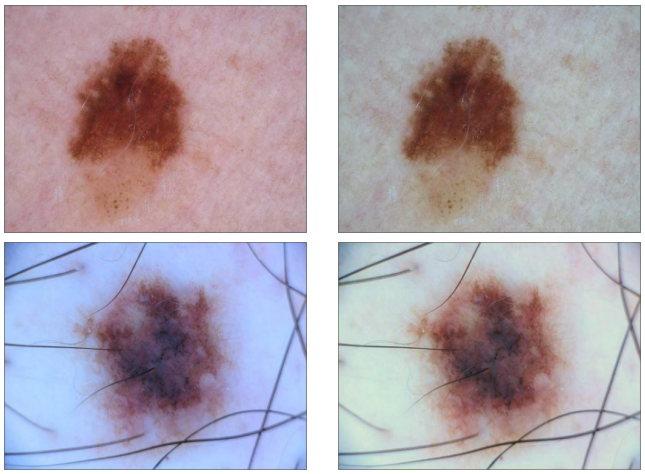
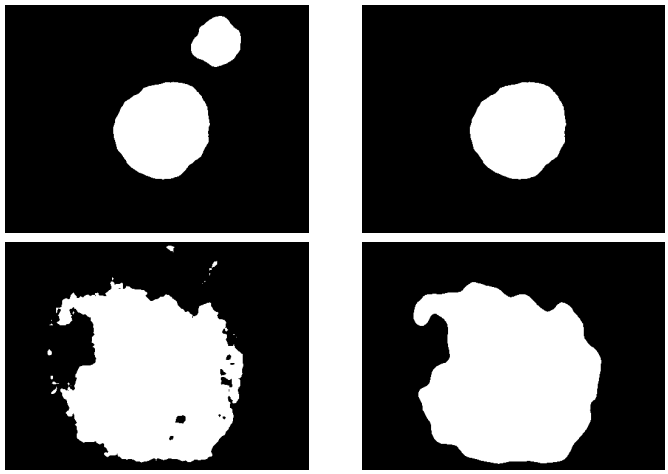


Fig. 3. Complete flow of our proposed ensemble method to produce lesion segmentation



(a) Original Images (b) After Preprocessing

Fig. 4. Examples of pre-processing by Shades of Gray algorithm



(a) Output Segmentation Masks (b) Processed Masks

Fig. 6. Examples of post-processing by Image Processing Methods

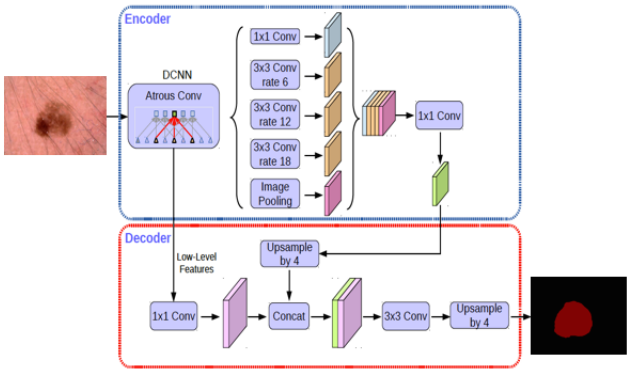


Fig. 5. Detailed architecture of DeeplabV3+ for segmentation on skin lesion dataset [25].

this experiment [26]. In default setting, in some cases, Mask-RCNN generate more than one output. We made changes in the final layer of Mask-RCNN architecture to produce only single output mask of highest confidence per image.

4) Post-processing: We used basic image processing methods such as morphological operations to and fill the region and remove unnecessary artefacts of the results as shown in the Fig. 6. These issues were countered in only DeeplabV3+ as in the case of Mask-RCNN, we have not had these issues. Hence, post-processing is only used for the semantic segmentation methods like FCN and DeeplabV3+.

5) Ensemble: We used two types of ensemble methods called Ensemble-ADD and Ensemble-Comparison. First of all, if there is no prediction from one of the method, ensemble methods picks up the prediction of other method. Then, Ensemble-ADD combines the outputs of both Mask-RCNN and DeeplabV3+ to produce final segmentation mask. Ensemble-Comparison-Large picks the larger segmented area by comparing the number of pixels in output of both methods whereas Ensemble-Comparison-Small picks the smaller area from above mentioned methods. Both ensemble methods are demonstrated by the Fig. 7.

C. Performance Metrics

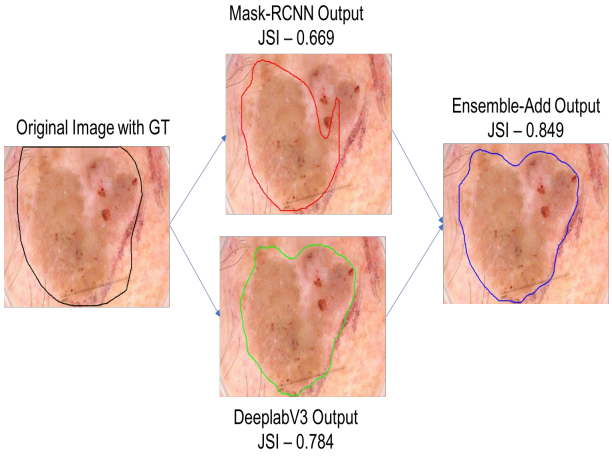
In medical imaging, *Sensitivity* and *Specificity* are the standard evaluation metrics and where as for segmentation evaluation, *Dice Similarity Coefficient (Dice)* is popularly used by researchers [34], [35]. We report our findings in *Jaccard Similarity Index (JSI)*, *Dice*, *Sensitivity*, *Specificity*, *Accuracy* and *Matthew Correlation Coefficient (MCC)* [36] as our evaluation metrics for segmentation.

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

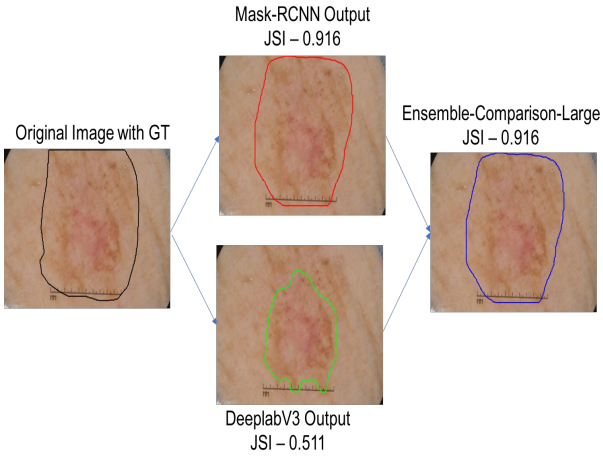
$$Specificity = \frac{TN}{FP + TN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

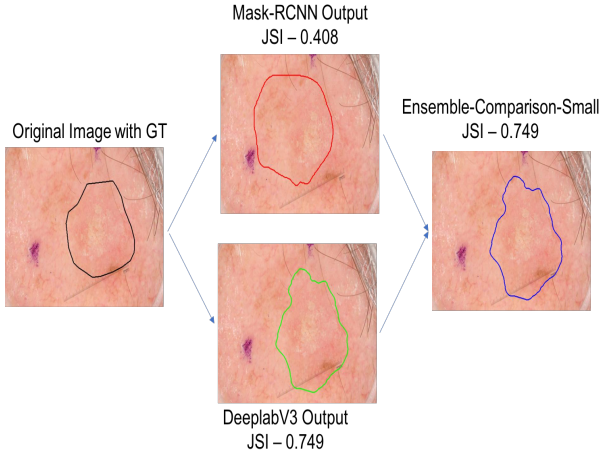
$$JSI = \frac{TP}{(TP + FP + FN)} \tag{4}$$



(a) Ensemble - ADD



(b) Ensemble - Comparison-Large



(c) Ensemble - Comparison-Small

Fig. 7. Examples of ensemble methods: (a) Ensemble-ADD (b) Ensemble-Comparison (Large) (c) Ensemble-Comparison (Small)

$$Dice = \frac{2 * TP}{(2 * TP + FP + FN)} \tag{5}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

Sensitivity is defined in eq (1), where *TP* is *True Positives* and *FN* is *False Negatives*. A high *Sensitivity* (close to 1.0) indicates good performance in segmentation which implies all the lesions were segmented successfully. On the other hand, *Specificity* (as in eq. (2)) indicates the proportion of *True Negatives (TN)* of the non-lesions. A high *Specificity* indicates the capability of a method in not segmenting the non-lesions. *JSI* and *Dice Similarity Index (Dice)* is a measure of how similar both prediction and ground truth are, by measuring of how many *TP* found and penalising for the *FP* that the method found, as in eq. (3). *MCC* has a range of -1 (completely wrong binary classifier) to 1 (completely right binary classifier). This is a suitable measurement for the performance assessment of our segmentation algorithms based on binary classification (lesion versus non-lesions), as in eq. (4).

IV. Experiment and Results

This section presents the performance of our proposed methods and various state-of-the-art segmentation methods on ISIC 2017 testing set that consists of 600 dermoscopic images and PH2 dataset consists of 200 images [20], [31].

We used ISBI-2017 dataset to train all the networks on a GPU machine with the following specification: (1) Hardware: CPU - Intel i7-6700 @ 4.00Ghz, GPU - NVIDIA TITAN X 12Gb, RAM - 32GB DDR5 (2) Software: Tensor-flow.

In Table I, and II, we presented the performance of our proposed method with other trained fully convolutional networks using popular segmentation evaluation metrics that are *Sensitivity*, *Specificity*, and *Accuracy*, *Dice*, *Jaccard Similarity Index*, and *Matthew's correlation coefficient*. We trained fully convolutional networks, DeeplabV3+, Mask-RCNN, and ensemble methods on the ISIC 2017 training set and tested on ISIC 2017 testing set as shown in Table I and II. In Table III, we compared our results with competition winners and other segmentation algorithms presented in [37] with default competition performance metrics.

Our proposed method Ensemble-Add achieved *Jaccard Similarity Index* of 79.34% for ISIC testing set 2017 which outperformed the FCN-16s, U-Net, SegNet, and FrCN by approximately 6%, 20%, 12%, and 5% respectively. Ensemble-S outperformed other algorithms in terms of textitSpecificity with the score of 97.94% where as Ensemble-A received highest score in *Sensitivity* and other performance measures. In Fig. 8, we compared the JSI scores produced by the proposed methods.

In 2017, International Skin Imaging Collaboration, launched the ISBI challenge to get the best performance measures in classification and segmentation task. In the lesion segmentation, participants were asked to segment the boundaries of lesion irrespective of original class. The default performance measures of winner algorithms along with other significant segmentation algorithms presented in [37] and our proposed

TABLE I
PERFORMANCE EVALUATION OF OUR PROPOSED METHODS AND STATE-OF-THE-ART SEGMENTATION ARCHITECTURES ON ISIC 2017 TESTING SET (SEN DENOTES *Sensitivity*,SPE IS *Specificity*, ACC IS *Accuracy*, AND SK DENOTES SEBORRHEIC KERATOSIS)

Method	Naevus			Melanoma			SK			Overall		
	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC
FCN-AlexNet	82.44	97.58	94.84	72.35	96.23	87.82	71.70	97.92	89.35	78.86	97.37	92.65
FCN-32s	83.67	96.69	94.59	74.36	96.32	88.94	75.80	96.41	89.45	80.67	96.72	92.72
FCN-16s	84.23	96.91	94.67	75.14	96.27	89.24	75.48	96.25	88.83	81.14	96.68	92.74
FCN-8s	83.91	97.22	94.55	78.37	95.96	89.63	69.85	96.57	87.40	80.72	96.87	92.52
DeeplabV3+	88.54	97.21	**95.67**	77.71	96.37	89.65	74.59	98.55	90.06	84.34	97.25	93.66
Mask-RCNN	87.25	96.38	95.32	78.63	95.63	89.31	82.41	94.88	90.85	84.84	96.01	93.48
Ensemble-S	84.74	**97.98**	95.58	73.35	**97.30**	88.40	71.80	**98.58**	89.91	80.58	**97.94**	93.33
Ensemble–L	90.93	95.74	95.51	83.40	95.00	90.61	85.81	94.74	91.34	88.70	95.45	93.93
Ensemble-A	**92.08**	95.37	95.59	**84.62**	94.20	**90.85**	**87.48**	94.41	**91.72**	**89.93**	95.00	**94.08**

TABLE II
PERFORMANCE EVALUATION OF OUR PROPOSED METHODS AND STATE-OF-THE-ART SEGMENTATION ARCHITECTURES ON ISIC 2017 TESTING SET (DIC DENOTES *Dice Score*,JSI IS *Jaccard Similarity Index*, MCC IS *Mathews Correlation Coefficient*, AND SK DENOTES SEBORRHEIC KERATOSIS)

Method	Naevus			Melanoma			SK			Overall		
	DIC	JSI	MCC	DIC	JSI	MCC	DIC	JSI	MCC	DIC	JSI	MCC
FCN-AlexNet	85.61	77.01	82.91	75.94	64.32	70.35	75.09	63.76	71.51	82.15	72.55	78.75
FCN-32s	85.08	76.39	82.29	78.39	67.23	72.70	76.18	64.78	72.10	82.44	72.86	78.89
FCN-16s	85.60	77.39	82.92	79.22	68.41	73.26	75.23	64.11	71.42	82.80	73.65	79.31
FCN-8s	84.33	76.07	81.73	80.08	69.58	74.39	68.01	56.54	65.14	81.06	71.87	77.81
DeeplabV3+	88.29	81.09	85.90	80.86	71.30	76.01	77.05	67.55	74.62	85.16	77.15	82.28
Mask-RCNN	88.83	80.91	85.38	80.28	70.69	74.95	80.48	70.74	76.31	85.58	77.39	81.99
Ensemble-S	87.93	80.46	85.58	78.45	68.42	73.61	76.88	66.62	74.05	84.42	76.03	81.51
Ensemble-L	88.87	81.69	85.93	83.05	74.01	77.98	81.71	72.50	77.68	86.66	78.82	83.14
Ensemble-A	**89.28**	**82.11**	**86.33**	**83.54**	**74.53**	**78.08**	**82.53**	**73.45**	**78.61**	**87.14**	**79.34**	**83.57**

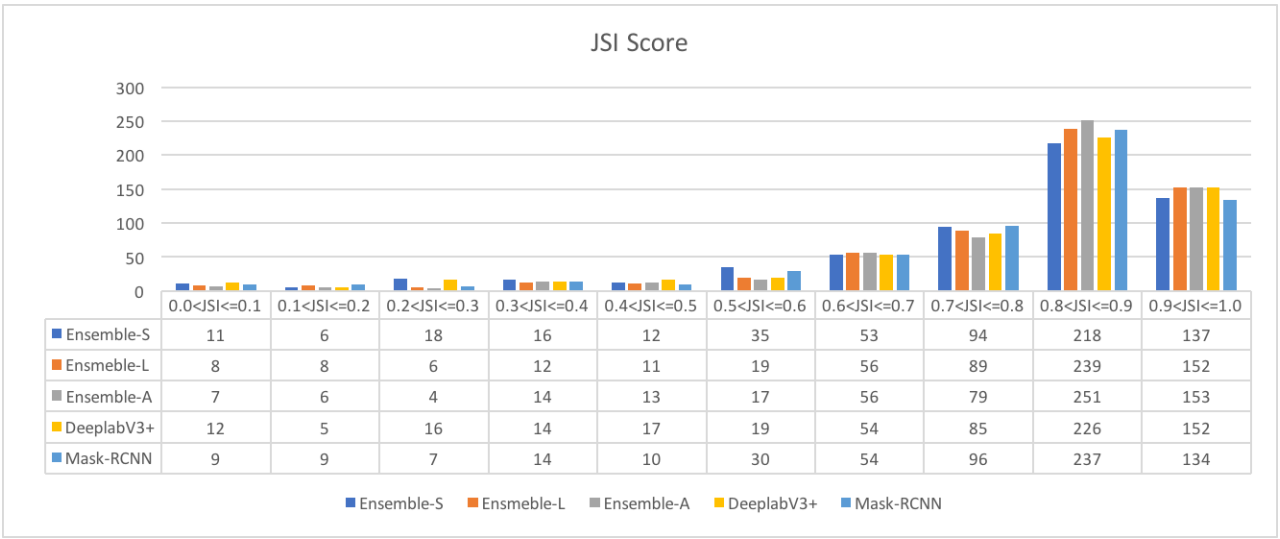


Fig. 8. Comparison of JSI scores of proposed methods for skin lesion segmentation of ISIC Challenge 2017 testing set

TABLE III
PERFORMANCE EVALUATION OF OUR PROPOSED METHODS AND STATE-OF-THE-ART ALGORITHMS ON ISIC SKIN LESION SEGMENTATION CHALLENGE 2017

User Name (Method)	Accuracy	Dice	Jaccard Index	Senstivity	Specificity
First: Yading Yuan (CDNN Model)	0.934	0.849	0.765	0.825	0.975
Second: Matt Berseth (U-Net)	0.932	0.847	0.762	0.820	0.978
U-Net [29]	0.901	0.763	0.616	0.672	0.972
SegNet [38]	0.918	0.821	0.696	0.801	0.954
FrCN [37]	0.940	0.870	0.771	0.854	0.967
Ensemble-S (Proposed Method)	0.933	0.844	0.760	0.806	**0.979**
Ensemble-L (Proposed Method)	0.939	0.866	0.788	0.887	0.955
Ensemble-A (Proposed Method)	**0.941**	**0.871**	**0.793**	**0.899**	0.950

TABLE IV
PERFORMANCE EVALUATION OF DIFFERENT SEGMENTATION ALGORITHMS ON PH2 DATASET

User Name (Method)	Accuracy	Dice	Jaccard Index	Senstivity	Specificity
FCN-16s	0.917	0.881	0.802	0.939	0.884
DeeplabV3+	0.923	0.890	0.814	0.943	0.896
Mask-RCNN	0.937	0.904	0.830	0.969	0.897
Ensemble-S (Proposed Method)	**0.938**	**0.907**	**0.839**	0.932	**0.929**
Ensemble-L (Proposed Method)	0.922	0.887	0.806	0.980	0.865
Ensemble-A (Proposed Method)	0.919	0.883	0.800	**0.987**	0.851

methods are described in Table III. Again, our proposed methods received highest scores in the default performance measures in this challenge when compared to the other algorithms.

It is clear from Table III, the performance of our proposed ensemble methods is better than the competition winners.

To check the further robustness of our method, we used our proposed algorithms trained on ISIC 2017 training set to test PH2 dataset. It is worth noted that Ensemble-A produced the best results in ISIC 2017 testing set where as in PH2 dataset, it's Ensemble-S achieved better score in PH2 dataset in Table IV.

V. CONCLUSION

Robust and end-to-end skin segmentation solutions are very important to aid dermatologists to provide inference according to the ABCD rule system for lesion diagnosis of skin cancers especially melanoma. In this work, we proposed the fully automatic ensemble deep learning methods which combine one of the best segmentation methods that are DeeplabV3+ (semantic segmentation) and Mask Rcnn (instance segmentation) to produce significant accurate results than single segmentation CNN methods on both ISIC 2017 testing set and PH2 dataset. We also utilized the pre-processing by using color constancy algorithm to normalize the data and then, morphological image functions for post-processing to produce segmentation results. Our proposed method outperformed the other state-of-the-art segmentation methods and 2017 challenge winner with significant margin on popular performance metrics used for segmentation. The further improvement can be made by further tweaking the hyper-parameters of both networks we utilized in ensemble methods. This study only focuses on the ensemble methods for segmentation tasks on skin lesion datasets. It can be further tested on the other publicly available segmentation datasets in both medical and non-medical domains.

REFERENCES

[1] S. Pathan, K. G. Prabhu, and P. Siddalingaswamy, "Techniques and algorithms for computer aided diagnosis of pigmented skin lesions???a review," *Biomedical Signal Processing and Control*, vol. 39, pp. 237–262, 2018.

[2] National Cancer Institute, "Cancer stat facts: Melanoma of the skin," 2017, last access: 26/10/17. [Online]. Available: https://seer.cancer.gov/statfacts/html/melan.html

[3] Melanoma Foundation (AIM), "Melanoma stats, facts and figures," 2017, last access: 27/10/2017. [Online]. Available: https://www.aimatmelanoma.org/about-melanoma/melanoma-stats-facts-and-figures/

[4] G. Pellacani and S. Seidenari, "Comparison between morphological parameters in pigmented skin lesion images acquired by means of epiluminescence surface microscopy and polarized-light videomicroscopy," *Clinics in dermatology*, vol. 20, no. 3, pp. 222–227, 2002.

[5] J. Mayer, "Systematic review of the diagnostic accuracy of dermatoscopy in detecting malignant melanoma." *The Medical Journal of Australia*, vol. 167, no. 4, pp. 206–210, 1997.

[6] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: a review," *Artificial intelligence in medicine*, vol. 56, no. 2, pp. 69–90, 2012.

[7] G. Argenziano, H. P. Soyer, V. De Giorgio, D. Piccolo, P. Carli, M. Delfino, A. Ferrari, R. Hofmann-Wellenhof, D. Massi, G. Mazzocchetti *et al.*, "Interactive atlas of dermoscopy," 2000.

[8] E. Ahmad, M. Goyal, J. S. McPhee, H. Degens, and M. H. Yap, "Semantic segmentation of human thigh quadriceps muscle in magnetic resonance images," *arXiv preprint arXiv:1801.00415*, 2018.

[9] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.

[10] M. H. Yap, M. Goyal, F. Osman, E. Ahmad, R. Martí, E. Denton, A. Juette, and R. Zwiggelaar, "End-to-end breast ultrasound lesions recognition with a deep learning approach," in *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 10578. International Society for Optics and Photonics, 2018, p. 1057819.

[11] M. H. Yap, M. Goyal, F. M. Osman, R. Martí, E. Denton, A. Juette, and R. Zwiggelaar, "Breast ultrasound lesions recognition: end-to-end deep learning approaches," *Journal of Medical Imaging*, vol. 6, no. 1, p. 011007, 2018.

[12] S. Walsh, L. Calandriello, M. Silva, and N. Sverzellati, "A deep learning algorithm for classifying fibrotic lung disease on high resolution computed tomography," in *A23. ILD: DIAGNOSIS*. American Thoracic Society, 2018, pp. A7645–A7645.

[13] J. Alarifi, M. Goyal, A. Davison, D. Dancey, R. Khan, and M. H. Yap, "Facial skin classification using convolutional neural networks," in *Image Analysis and Recognition: 14th International Conference, ICIAR 2017, Montreal, QC, Canada, July 5–7, 2017, Proceedings*, vol. 10317. Springer, 2017, p. 479.

[14] M. Goyal, N. D. Reeves, A. K. Davison, S. Rajbhandari, J. Spragg, and M. H. Yap, "Dfunet: Convolutional neural networks for diabetic foot ulcer classification," *arXiv preprint arXiv:1711.10448*, 2017.

[15] M. Goyal, M. H. Yap, N. D. Reeves, S. Rajbhandari, and J. Spragg, "Fully convolutional networks for diabetic foot ulcer segmentation," in *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*. IEEE, 2017, pp. 618–623.

[16] M. Goyal, N. Reeves, S. Rajbhandari, and M. H. Yap, "Robust methods for real-time diabetic foot ulcer detection and localization on mobile devices," *IEEE journal of biomedical and health informatics*, 2018.

[17] M. Goyal and M. H. Yap, "Multi-class semantic segmentation of skin lesions via fully convolutional networks," *arXiv preprint arXiv:1711.10449*, 2017.

[18] M. Goyal, J. Ng, and M. H. Yap, "Multi-class lesion diagnosis with pixel-wise classification network," *arXiv preprint arXiv:1807.09227*, 2018.

[19] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2015, pp. 118–126.

[20] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, "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)," *arXiv preprint arXiv:1710.05006*, 2017.

[21] ——, "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)," in *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. IEEE, 2018, pp. 168–172.

[22] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE Transactions on Medical Imaging*, 2017.

[23] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE transactions on medical imaging*, vol. 36, no. 4, pp. 994–1004, 2017.

[24] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic image segmentation via multi-stage fully convolutional networks," *IEEE Transactions on Biomedical Engineering*, 2017.

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[27] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.

[28] M. H. Yap, G. Pons, J. Martí, S. Ganau, M. Sentís, R. Zwiggelaar, A. K. Davison, and R. Martí, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE journal of biomedical and health informatics*, 2017.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[30] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and 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)," *arXiv preprint arXiv:1605.01397*, 2016.

[31] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "Ph 2-a dermoscopic image database for research and benchmarking," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. IEEE, 2013, pp. 5437–5440.

[32] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *arXiv preprint arXiv:1803.10417*, 2018.

[33] G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *Color and Imaging Conference*, vol. 2004, no. 1. Society for Imaging Science and Technology, 2004, pp. 37–41.

[34] D. Zikic, Y. Ioannou, M. Brown, and A. Criminisi, "Segmentation of brain tumor tissues with convolutional neural networks," *Proceedings MICCAI-BRATS*, pp. 36–39, 2014.

[35] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[36] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

[37] M. A. Al-masni, M. A. Al-antari, M.-T. Choi, S.-M. Han, and T.-S. Kim, "Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks," *Computer methods and programs in biomedicine*, vol. 162, pp. 221–231, 2018.

[38] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *arXiv preprint arXiv:1505.07293*, 2015.