IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 15, 2019, accepted December 7, 2019, date of publication December 18, 2019,
date of current version January 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2960504

Skin Lesion Segmentation in Dermoscopic Images With Ensemble Deep Learning Methods

MANU GOYAL[1], AMANDA OAKLEY[2], PRIYANKA BANSAL[1], DARREN DANCEY[1], AND MOI HOON YAP[1], (Senior Member, IEEE)

[1]Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M15 6BH, U.K.
[2]Waikato Clinical School, The University of Auckland, Hamilton 3204, New Zealand

Corresponding author: Moi Hoon Yap (m.yap@mmu.ac.uk)

ABSTRACT Early detection of skin cancer, particularly melanoma, is crucial to enable advanced treatment. Due to the rapid growth in the number of skin cancers, there is a growing need of computerised analysis for skin lesions. The state-of-the-art public available datasets for skin lesions are often accompanied with a very limited amount of segmentation ground truth labeling. Also, the available segmentation datasets consist of noisy expert annotations reflecting the fact that precise annotations to represent the boundary of skin lesions are 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 fully automated deep learning ensemble methods to achieve high sensitivity and high specificity in lesion boundary segmentation. We trained the ensemble methods based on Mask R-CNN and DeeplabV3+ methods on ISIC-2017 segmentation training set and evaluate the performance of the ensemble networks on ISIC-2017 testing set and PH2 dataset. Our results showed that the proposed ensemble methods segmented the skin lesions with Sensitivity of 89.93% and Specificity of 97.94% for the ISIC-2017 testing set. The proposed ensemble method Ensemble-A outperformed FrCN, FCNs, U-Net, and SegNet in Sensitivity by 4.4%, 8.8%, 22.7%, and 9.8% respectively. Furthermore, the proposed ensemble method Ensemble-S achieved a specificity score of 97.98% for clinically benign cases, 97.30% for the melanoma cases, and 98.58% for the seborrhoeic keratosis cases on ISIC-2017 testing set, exhibiting better performance than FrCN, FCNs, U-Net, and SegNet.

INDEX TERMS Skin cancer, skin lesion segmentation, ensemble segmentation methods, deep learning, melanoma, instance segmentation, semantic segmentation.

I. INTRODUCTION

Cancers of the skin are the most common form of malignancy in humans [1]. The most common malignant skin lesions are melanoma, squamous cell carcinoma and basal cell carcinoma. It is estimated that in 2019, 96,480 new cases will be diagnosed with melanoma and more than 7,000 people will die from the disease in the United States [2], [3]. Early detection of melanoma can save lives.

It can be difficult to differentiate benign lesions from skin cancers. Skin cancer specialists examine their patients' skin lesions using visual inspection aided by hand-held dermoscopy. They may capture digital close-up (macroscopic) and dermoscopic (microscopic) images. Dermoscopy is a means to examine the skin using a bright light and magnification, and employs either polarisation or immersion fluid to reduce surface reflection [4]. In common use for the last 20 years, dermoscopy has improved the diagnosis rate over visual inspection alone [5]. The ABCD criteria help non-dermatologists screen skin lesions to differentiate common benign melanocytic naevi (naevi) from melanoma [6]. Fig. 1 illustrates the ABCD rules for skin lesion diagnosis, where:

1) A: Asymmetry property checks whether two halves of the skin lesion match or not in terms of colour, shape, edges. The skin lesions are divided into two halves based on long axis and short axis as shown in Fig. 1. In the case of melanoma, it is likely to have an asymmetrical appearance.
2) B: Border property. It defines whether the edges of skin lesion are smooth, well-defined or otherwise. In the case of melanoma, edges are likely to be uneven, blurry and jagged.

The associate editor coordinating the review of this manuscript and approving it for publication was Haiyong Zheng.

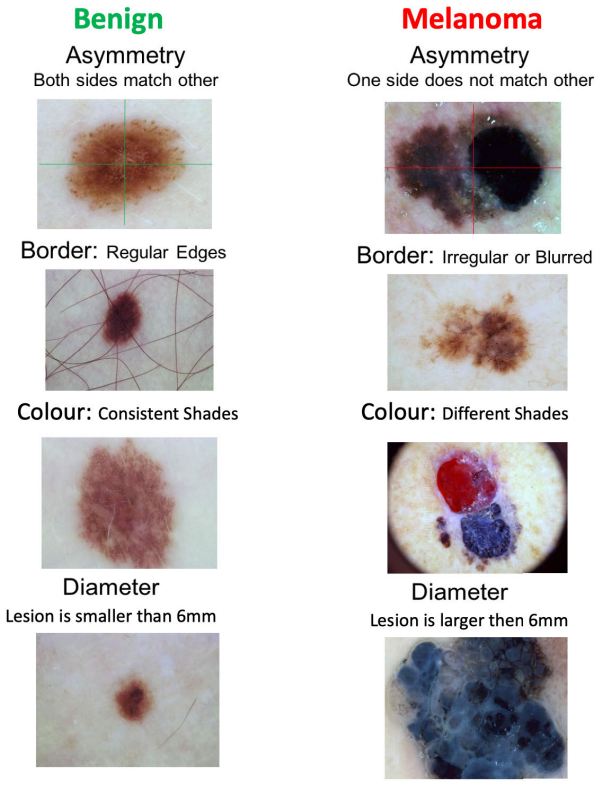


FIGURE 1. Lesion diagnosis by dermatologists. ABCD criteria for lesion diagnosis focuses on finding the certain properties of lesions.

3) C: Colour property. The colour in a melanoma varies from one area to another, and it often has varying shades of tan, brown, red, and black.
4) D: Diameter property. It measures the approximate diameter of the skin lesion. The diameter of a melanoma is generally greater than 6mm (the size of a pencil eraser).

End-to-end computerised solutions that can produce accurate segmentation of skin lesions irrespective of types of skin lesions are highly desirable to mirror the clinical ABCD Rule. For segmentation of medical imaging, *Jaccard Similarity Index (JSI)*, *specificity* and *sensitivity* are deemed as important performance measures for methods. Hence, computerised methods need to achieve high scores in these performance metrics.

The majority of the state-of-the-art computer-aided diagnosis based on dermoscopic images are composed of multi-stages, which include image pre-processing, image segmentation [7], feature extraction [8] and classification [9]. Using hand-crafted feature descriptors, benign naevi tend to have small dimensions and a roundish shape, as illustrated in Fig. 1. (but some naevi are large and unusual shapes). Other feature descriptors used in previous works include asymmetry features, colour features and texture features. Pattern analysis is widely used to describe the dermoscopic appearance of skin lesions, for example, the melanocytic algorithm elaborated by Lio and Nghiem [10]. Ashour *et al.* [11], [12] proposed a histogram-based clustering estimation (HBCE) algorithm to determine the required number of clusters in the neutrosophic c-means clustering (NCM) method to perform segmentation on ISIC-2016 dataset and neutrosophic k-means (ONKM) using genetic algorithm for skin lesion detection in dermoscopy images. Various computer algorithms have been devised to classify lesion types using feature descriptors and pattern analysis based on image processing and conventional machine learning approaches. Two reviews by Korotkov and Garcia [13] and Pathan *et al.* [1] reported that the majority of these used hand-crafted features to classify or segment the lesions. Korotkov and Garcia [13] concluded that there is a large discrepancy in previous research and the computer-aided diagnosis (CAD) systems were not ready for implementation. The other issue was the lack of a benchmark dataset, which makes it harder to assess the algorithms. Pathan *et al.* [1] concluded that the CAD systems worked in experimental settings but required rigorous validation in real-world clinical settings. Because these systems require manual tuning of hyper-parameters and composed of multi-stages.

With the rapid growth of deep learning approaches, many researchers [14]–[16] have proposed using Deep Convolutional Neural Networks (CNN) for melanoma detection and segmentation.

II. DEEP LEARNING FOR SKIN LESION SEGMENTATION

This section reviews the state-of-the-art deep learning approaches for segmentation for skin lesions. Deep learning has gained popularity in medical imaging research including Magnetic Resonance Imaging (MRI) on brain [17], breast ultrasound cancer detection [18]. Recently, U-Net is a popular deep learning approach in biomedical imaging research, proposed by Ronneberger *et al.* [19]. 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.

Researchers have made significant contributions proposing various deep learning frameworks for the detection and segmentation of skin lesions. Yu *et al.* [15] 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 [20], 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 the segmentation stage, they ranked second in segmentation among the 28 teams. The work showed promising results, but the two-stage framework and very deep networks were computationally expensive.

Bi *et al.* [16] 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 [21] of 90.66% but achieved marginal improvement if compared to Team ExB in ISIB 2016 competition with 91.18%.

Yuan *et al.* [14] 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.* [16], they evaluated their results on ISBI 2016 and PH2 dataset. The results were outperformed by 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 and Yap [22] 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 seborrhoeic keratosis rather than a single class of skin lesion.

Vesal *et al.* [23] and Goyal *et al.* [24] proposed two-stage segmentation method which used Faster-RCNN in the first stage and then a modified version of U-Net and deep extreme method respectively as second stage to achieve segmentation results.

Soudani and Barhoumi [25] used two deep learning classification models to recommend the most appropriate segmentation technique on ISIC-2017 dataset. Recently, Al-masni *et al.* [26] proposed a fully resolution convolutional network (FrCN) to learn full resolution features of each pixel of dermoscopic skin lesion images for skin segmentation. They achieved Jaccard Index of 77.11% on ISIC-2017 testing set.

The research showed that deep learning achieved promising results for segmentation and classification of skin lesions. However, there is a huge difference in availability of ground truths with respect to segmentation and classification in public skin lesion datasets. This is mainly because, the expert annotation for segmentation ground truths is very expensive and laborious when compared with classification labels. In ISIC 2018 competition, only a total number of 3694 images were available for the segmentation challenge in comparison to 11720 images in the classification challenge [27], [28]. The expert annotations used in segmentation tasks are not always very accurate which could affect the performance of segmentation algorithms. Some of these examples in ISIC-2017 testing set are shown in Fig. 2. Other issues regarding the expert annotation are, some experts tend to draw a very precise outer boundary of skin lesions while others draw loose outer boundary to represent skin lesions for

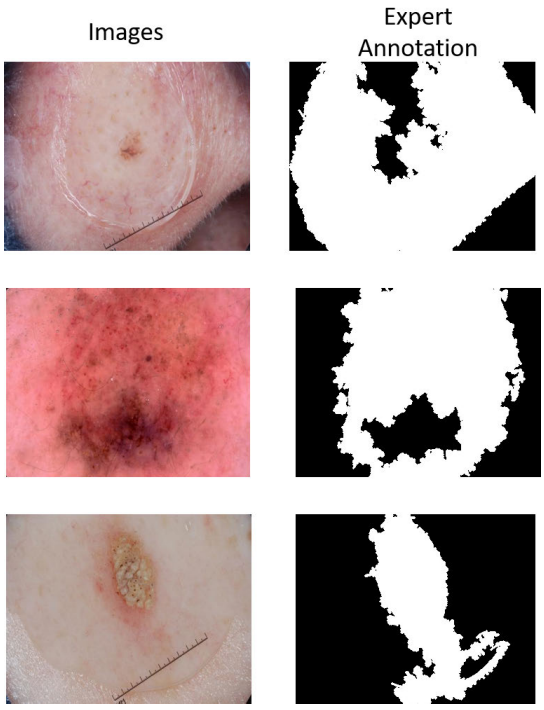


FIGURE 2. Examples of noisy expert annotations from from skin lesion dataset. (Left) Original Images; and (Right) Ground truth in binary masks.

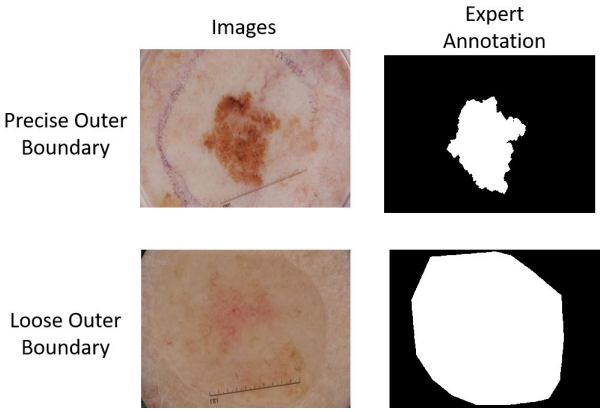


FIGURE 3. Examples of precise and loose boundary representation of skin lesions. (Left) Original Images; and (Right) Ground truth in binary masks.

segmentation ground truths as shown in Fig. 3. Depending on whether the expert annotations precise or loose for the skin lesion dataset, it is very difficult for any deep learning algorithm to produce accurate segmentation results. For precise boundary representation of skin lesions, the algorithm needs to have high *Specificity* score whereas for loose representation, high *Sensitivity* score is desirable. In this paper, we addressed these issues by developing three fully automatic CNN-based ensemble methods to suit both precise and loose lesion boundary segmentation. We trained them on the ISIC-2017 dermoscopic training set and tested the robustness of the ISIC-2017 trained algorithms on ISIC-2017 testing set and another publicly available dataset, the PH2 dataset.

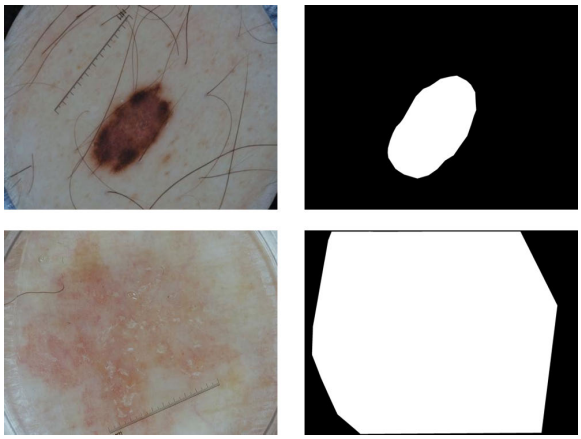


FIGURE 4. Examples of skin lesion images and labels from ISIC-2017 dataset. (Left) Original Images; and (Right) Ground truth in binary masks.

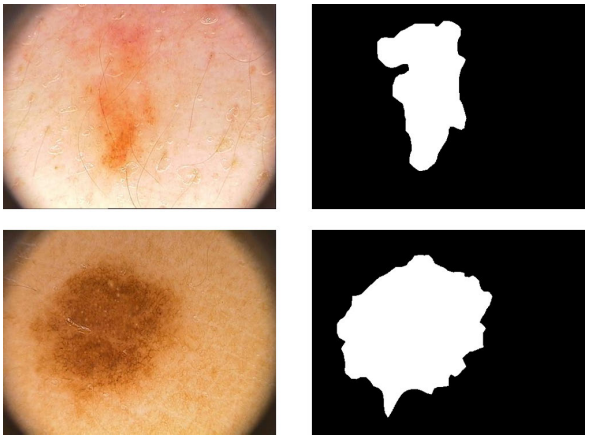


FIGURE 5. Examples of skin lesion images and labels from PH2 dataset. (Left) Original Images; and (Right) Ground truth in binary masks.

III. METHODOLOGY

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

A. SKIN LESION DATASETS

For this work, we used two publicly available datasets for skin lesions, which are ISIC-2017 Challenge (Henceforth ISIC-2017) [27] and PH2 dataset [21]. To improve the performance and reduce the computational cost, we resized all the images to 500 × 375.

1) ISIC-2017 SEGMENTATION DATASET

The International Skin Imaging Collaboration (ISIC) is a driving force by providing the digital skin lesion image datasets with expert annotations from around the world for automated CAD solutions for the diagnosis of melanoma and other cancers. This community also organises yearly skin lesion challenges to attract wider participation of researchers to improve the diagnosis of CAD algorithms and spread awareness regarding skin cancer [27]. The 2017 segmentation competition category consists of 2750 images with 2000 images in the training set, 150 in the validation set and 600 in the testing set. Fig. 4 is an example of loosely drawn boundary by experts to label ground truths in the dataset. Hence the algorithms need to achieve high *sensitivity* to perform well in this testing set. Even though ISIC Challenge 2018 [27] was conducted last year, they did not share the ground truth of their testing set. Therefore, our work was based on the ISIC-2017.

2) PH2 DATASET

PH2 dataset has 200 images in which 160 images are naevus (atypical naevus and common naevus), and 40 images are of melanoma [21]. In this dataset, the ground truths represent the precise and true boundaries of skin lesion (high *specificity*), as shown in the Fig. 5. We used this dataset as additional testing set for deep learning models trained on the ISIC-2017 segmentation training set.

B. ENSEMBLE METHODS FOR LESION BOUNDARY SEGMENTATION

We designed these end-to-end ensemble segmentation methods to combine Mask R-CNN and DeeplabV3+ with pre-processing and post-processing methods to produce accurate lesion segmentation as shown in Fig. 6. This section describes each stage of our proposed ensemble method.

1) PRE-PROCESSING

The ISIC Challenge dataset comprised of dermoscopic skin lesion images taken by different dermatoscopes and camera devices all over the world. Hence, it is important to perform pre-processing for colour normalization and illumination with colour constancy algorithm [31]. We processed the datasets with Shades of Gray algorithm [32] as shown in Fig. 7.

2) DEEPLABV3+

DeeplabV3+ is one of the best performing semantic segmentation networks achieving the test set performance of 89.0% and 82.1% in PASCAL VOC 2012 and Cityscapes datasets respectively [29]. DeeplabV3+ is an encoder-decoder network which makes the use of CNN 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 Fig. 8. To train DeeplabV3+ on skin lesion dataset, we used a pre-trained model on PASCAL VOC 2012 and adjusted the final output of 21 classes to a single class for segmentation of skin lesion [33]. It assigns semantic label lesion to every pixel in a dermoscopic image.

3) MASK R-CNN

Mask R-CNN is a recent deep learning architecture to provide instance segmentation i.e. identifying object outlines at the pixel level [30]. Mask R-CNN is inspired by the Faster R-CNN for object detection by adding a branch for

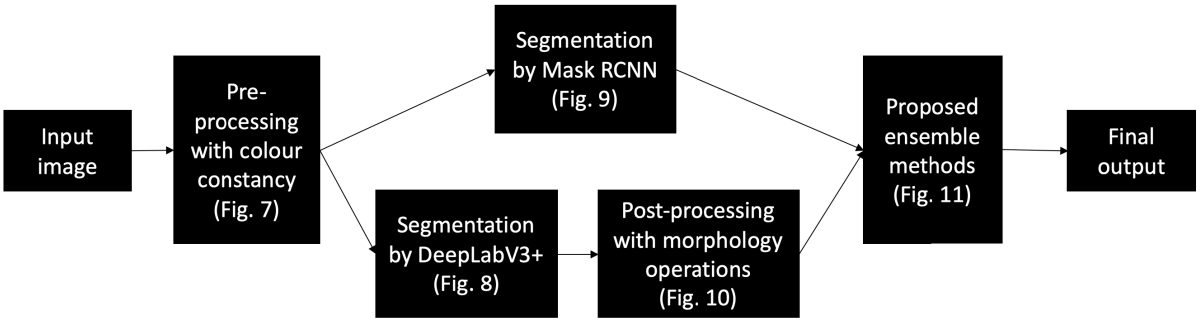


FIGURE 6. Complete flow of our proposed ensemble methods for automated skin lesion segmentation.

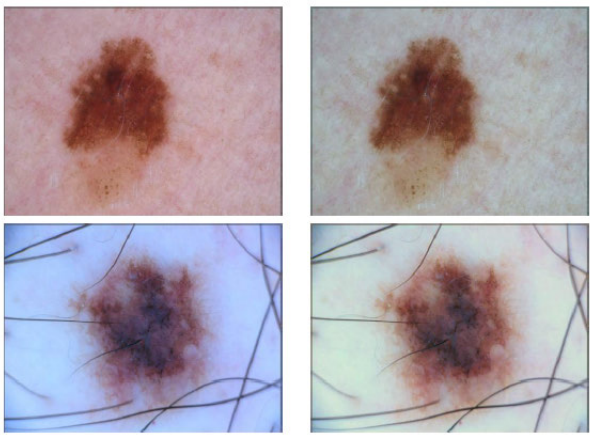


FIGURE 7. Examples of pre-processing stage by using Shades of Gray algorithm. (a) Original images with different background colours; and (b) Pre-processed images with more consistent background colours.

predicting an object mask in parallel with the existing branch for bounding box recognition [30], [34]. The framework is Mask R-CNN is detailed in Fig. 9. We fine-tuned a pre-trained Mask R-CNN with ResNet-InceptionV2 model (henceforth Mask R-CNN) on MS-COCO dataset for a single class as skin lesion for this experiment [35]. In some cases, Mask R-CNN generated more than one output due to the generation of 2k proposals from the Region Proposal Network (RPN) in the initial stage of Mask R-CNN. To limit the single output mask of highest confidence per image, we set the value of Top N proposals to 1 from RPN at test time.

4) POST-PROCESSING

We used basic image processing methods, i.e. morphological operations to fill the region and remove unnecessary artefacts of the results as illustrated in Fig. 10. These issues were only countered by DeeplabV3+ as in the case of Mask R-CNN, we have not had these issues. Hence, post-processing is only used for the semantic segmentation methods like FCNs and DeeplabV3+.

5) ENSEMBLE METHODS

We used two types of ensemble methods called Ensemble-ADD and Ensemble-Comparison. First of all, if there is no prediction from DeeplabV3+, the ensemble methods pick up the prediction of Mask R-CNN and vice versa. Then, Ensemble-ADD combines the results of both Mask R-CNN and DeeplabV3+ to produce the final segmentation mask. Ensemble-Comparison-Large picks the larger segmented area by comparing the number of pixels in the output of both methods. On contrary, Ensemble-Comparison-Small picks the smaller area from the output. The ensemble methods are illustrated in Fig. 11 where (a) shows Ensemble-ADD; (b) shows Ensemble-Comparison-Large; and (c) represents Ensemble-Comparison-Small. For convenience, we used the abbreviation as Ensemble-A for Ensemble-Add, Ensemble-L for Ensemble-Comparison-Large, and Ensemble-S for Ensemble-Comparison-Small. Ideally, Ensemble-S is designed for performing well in *Specificity* i.e. to have precise segmentation masks whereas Ensemble-A and Ensemble-L are designed for high *Sensitivity*.

C. PERFORMANCE METRICS

We evaluated the performance of the segmentation algorithms by using *Jaccard Similarity Index (JSI)*. In addition, we report our findings in *Dice Similarity Coefficient (Dice)*, *Sensitivity*, *Specificity*, *Accuracy* and *Matthew Correlation Coefficient (MCC)* [37].

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

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

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

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

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

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

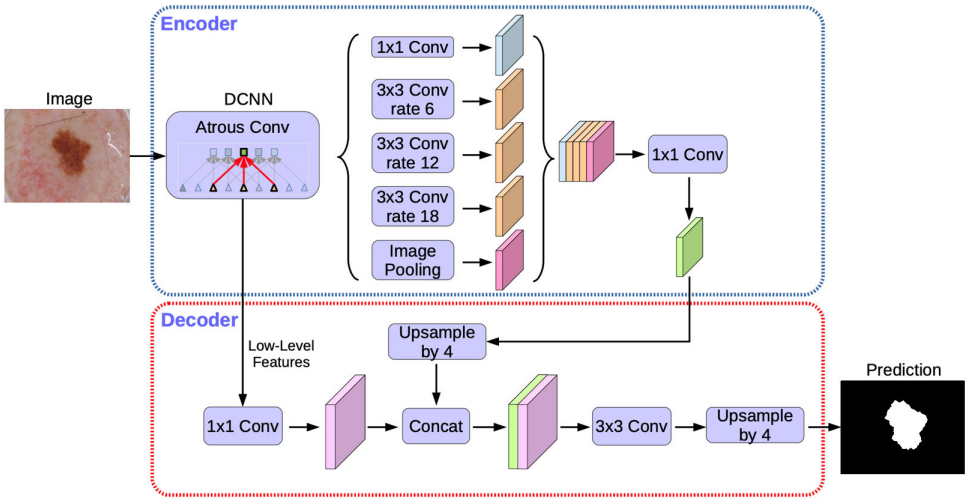


FIGURE 8. Architecture of DeeplabV3+ on skin lesion segmentation [29].

TABLE 1. Performance evaluation of our proposed methods and state-of-the-art algorithms on ISIC Skin Lesion Segmentation Challenge 2017.

Method	Accuracy	Dice	Jaccard Index	Sensitivity	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 [19]	0.901	0.763	0.616	0.672	0.972
SegNet [36]	0.918	0.821	0.696	0.801	0.954
FrCN [26]	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

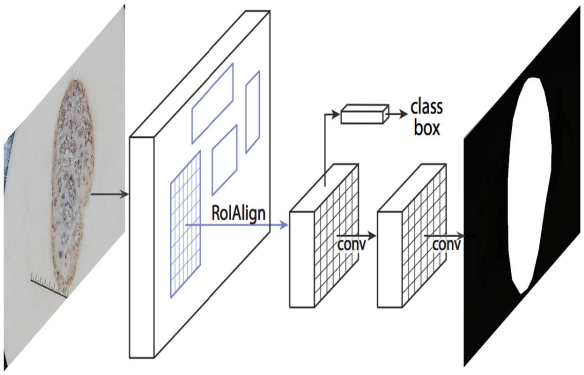
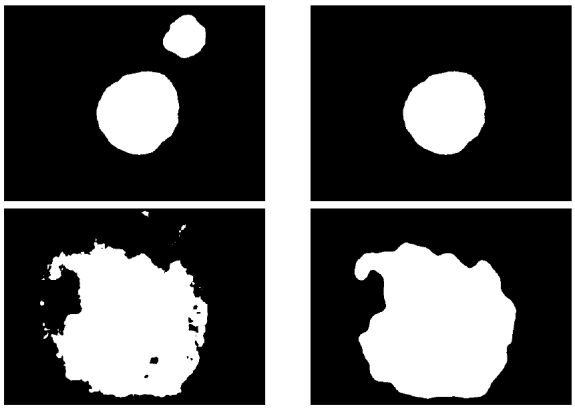


FIGURE 9. Architecture of Mask R-CNN on skin lesion segmentation [30].



(a) Output Segmentation Masks (b) Processed Masks

FIGURE 10. Examples of post-processing stage by using image processing methods: (a) Results from CNN segmentation with artefacts and holes within the lesions; and (b) Post-processed result after morphology operations.

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. *Accuracy* in segmentation methods report the percent of pixels in the image which were correctly classified as in eq. (3). *JSI* and *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

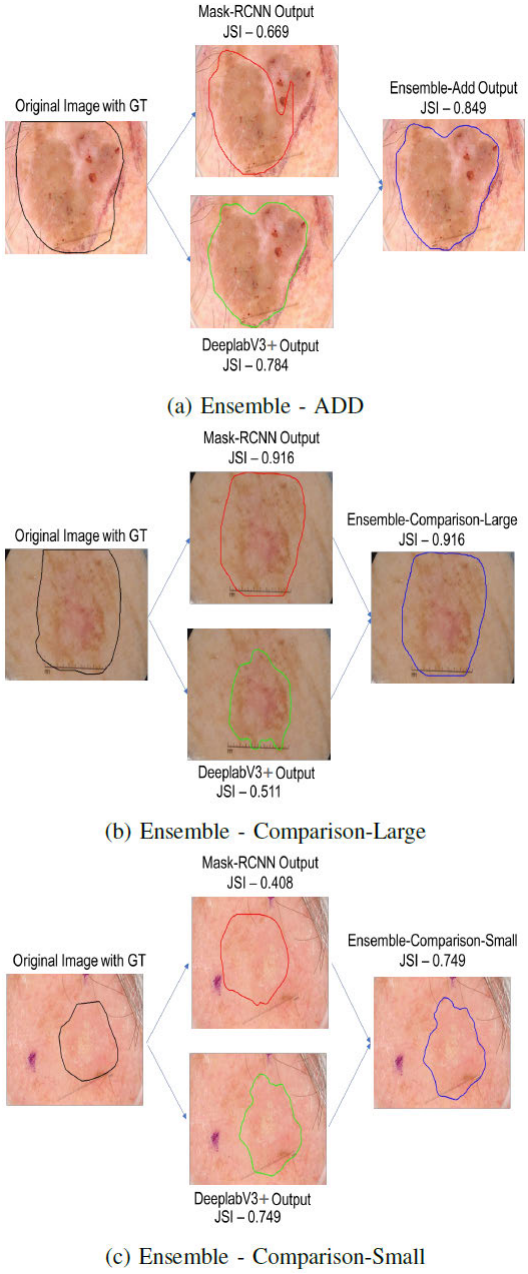


FIGURE 11. **Illustration of ensemble methods: (a) Ensemble-ADD (b) Ensemble-Comparison (Large) (c) Ensemble-Comparison (Small).**

found, as in eq. (4) and (5) respectively. *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. (6).

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 (600 images) and PH2 dataset (200 images) with given expert annotations [21], [27].

We trained 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.

We tested the different hyper-parameters including learning rates (1e-1 to 1e-4) depending on the deep learning models to train the models on ISIC-2017 training and validation segmentation dataset. We used 100 epochs for training the both deep learning architectures. We selected the models on the basis of minimum validation losses for the evaluation. For training DeeplabV3+, we used a CNN architecture of Xception-65 which use depth-wise separable convolutions, we set the `batch_size` of 1, weight for `l2_regularizer` as 0.00004, and `train_crop_size` of 513 × 513. The `base_learning_rate` of 0.001, `decay_factor` of 0.1 and `decay_step` of 2000, and momentum optimizer value is set at 0.9. For Mask R-CNN, we set the weight for `l2_regularizer` as 0.0, initializer that generates a truncated normal distribution with a standard deviation of 0.01. For training, we used a `batch_size` of 1, optimizer as momentum with manual `step_learning_rate` with an initial rate as 0.0003, 0.00003 at step 30000 and 0.000003 at step 45000. The momentum optimizer value is set at 0.9. We did not use any data augmentation techniques to train these models. We compared the predictions of each algorithm with given ground truths of testing sets.

A. COMPARISON WITH ISIC CHALLENGE 2017

Table 1 summarises the performance of our proposed methods when compared to the best method in the ISIC-2017 segmentation challenge and other segmentation algorithms using default competition performance metrics presented in [26]. Our proposed methods achieved the highest scores in the default performance measures in this challenge when compared to the other algorithms. As mentioned earlier for ISIC-2017, the ground truths are annotated loosely by the experts, the performance of algorithms in this dataset depends upon *Sensitivity* score. Our proposed method Ensemble-A achieved highest *Sensitivity* of 89.92% in comparison to other algorithms. Hence, Ensemble-A achieved a *JSI* of 79.34% for ISIC testing set 2017 which outperformed the first positioned algorithm in the competition by 2.8%, second positioned method by 3.1%, U-Net by 17.7%, SegNet by 9.7%, and FrCN by 0.5%. For *Specificity*. Ensemble-S outperformed other algorithms with a score of 97.94% which is marginally better than competition winners and a significant difference of 3.9% from Ensemble-A. Otherwise, Ensemble-A received the highest score in other categories of performance metrics that are *Sensitivity*, *Accuracy*, and *Dice*.

B. JACCARD SIMILARITY INDEX COMPARISON WITH TRAINED DEEP LEARNING MODELS

In Fig. 12, we compared the JSI scores produced by the proposed ensemble methods with DeeplabV3+ and Mask

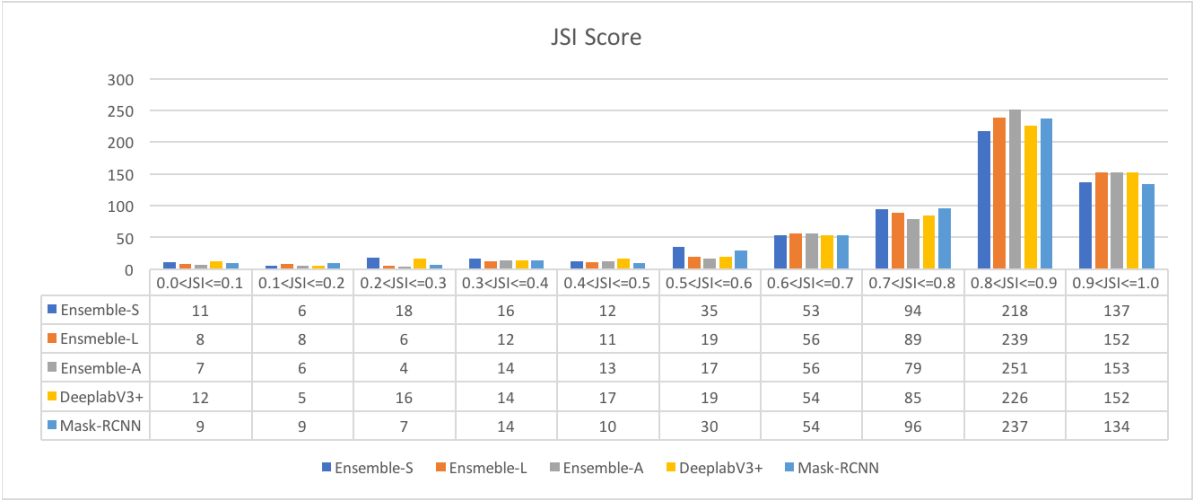


FIGURE 12. Comparison of JSI scores of our proposed methods for skin lesion segmentation of ISIC-2017 testing set.

TABLE 2. 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 Seborrhoeic 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 R-CNN	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**

R-CNN. This figure provides further clarification of the performance of each proposed method regarding the *JSI*. For JSI <= 0.5, Ensemble-A has minimum total number of cases (44), while DeeplabV3+ has a maximum (55). For 0.8 < JSI <= 1, Ensemble-A has a maximum number of cases of 404 out of the 600 images in the testing set, which is at least 13 cases more than the other algorithms. Overall, Ensemble-A performed the best in the evaluation metrics.

C. COMPARISON WITH THE STATE-OF-THE-ART BY LESION TYPES

In ISIC-2017 segmentation task, the participants were asked to segment the boundaries of lesion irrespective of the lesion types. In this section, we compare the accuracy of segmentation results based on three lesion types: Naevus, Melanoma, and Seborrhoeic Keratosis (SK).

In Table 2 and 3, we present the performance of our proposed method with other trained fully convolutional networks. We trained FCNs, DeeplabV3+, Mask R-CNN, and ensemble methods on the ISIC 2017 training set and tested on ISIC 2017 testing set. Since Ensemble-S method compares and picks the smaller area, it performed best in the *Specificity* category with a score of 97.94% by outperforming other algorithms overall and also for each lesion type in Table 2. Hence, Ensemble-S achieved the best position to perform well for evaluation metric *Specificity* as per our claim. For *Sensitivity*, Ensemble-A and Ensemble-L performed the best among other algorithms with a score of 89.93% and 88.70%. Hence, we proved our claim of designing these algorithms for high *Sensitivity*. Ensemble-A is marginally better in the *Accuracy* than other methods. Particularly, in Naevus category, DeeplabV3+ performed better than Ensemble-A with a fine margin.

In Table 3, we reported the performance of algorithms in terms of *Dice*, *JSI*, and *MCC*. Ensemble-A method outperformed all the participating segmentation algorithms for each skin lesion type.

D. COMPARISON ON PH2 DATASET

To test the robustness of our method and cross-dataset performance, we evaluate our proposed algorithms on PH2 dataset.

TABLE 3. 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 Seborrhoeic 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 R-CNN	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**

TABLE 4. Performance evaluation of different segmentation algorithms on PH2 dataset.

User Name (Method)	Accuracy	Dice	Jaccard Index	Sensitivity	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 R-CNN	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

It is worth noted that Ensemble-A produced the best results in ISIC 2017 testing set where as in PH2 dataset, Ensemble-S achieved best scores in PH2 dataset except *Sensitivity* in which Ensemble-A performed best, as shown in Table 4. This again proved our claim of performance of our Ensemble-S method for high *Specificity* and Ensemble-A for high *Sensitivity*. In PH2 dataset, the expert annotations are very precisely drawn for outer boundary of skin lesions, hence, that is why Ensemble-S performed the best on PH2 dataset.

V. CONCLUSION AND FUTURE SCOPE

Robust end-to-end skin segmentation solutions are very important to provide inference according to the ABCD rule system for the lesion diagnosis of melanoma. In this work, we designed the fully automatic ensemble deep learning methods which combine one of the best segmentation methods, i.e. DeeplabV3+ (semantic segmentation) and Mask R-CNN (instance segmentation) to produce notably more accurate results in different skin lesion datasets comprised of noisy expert annotations. According to our claim, Ensemble-L and Ensemble-A performed best in *Sensitivity* whereas Ensemble-S in *Specificity* in both ISIC 2017 testing set and PH2 dataset. We also utilised the pre-processing by using a colour constancy algorithm to normalise 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 ISIC challenge winners with good improvement on popular performance metrics used for segmentation. Further improvement can be made by fine-tuning the hyper-parameters of both networks in our ensemble methods. This study only focuses on the ensemble methods for segmentation tasks on skin lesion datasets. While incorporating more pre-processing techniques such as removing hair follicles and using data-augmentation techniques such as natural data-augmentation [38] can further improve the performance of these algorithms. In this work, we utilised the momentum as our optimization algorithm and cross-entropy as loss function, it would be interesting to see the impact of different optimizers such as Adam, SGD and loss functions such as Dice, JSI, FCE loss with these algorithms. Lastly, it can be 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," *Biomed. Signal Process. Control*, vol. 39, pp. 237–262, Jan. 2018.

[2] National Cancer Institute. (2019). *Cancer Stat Facts: Melanoma of the Skin*. Accessed: Sep. 7, 2019. [Online]. Available: https://seer.cancer.gov/statfacts/html/melan.html

[3] Melanoma Foundation (AIM). (2019). *Melanoma Stats, Facts and Figures*. Accessed: Sep. 7, 2019. [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," *Clin. Dermatol.*, vol. 20, no. 3, pp. 222–227, 2002.

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

[6] N. R. Abbasi, H. M. Shaw, D. S. Rigel, R. J. Friedman, W. H. McCarthy, I. Osman, A. W. Kopf, and D. Polsky, "Early diagnosis of cutaneous melanoma: Revisiting the ABCD criteria," *J. Amer. Med. Assoc.*, vol. 292, no. 22, pp. 2771–2776, Dec. 2004.
[7] M. E. Celebi, H. Iyatomi, G. Schaefer, and W. V. Stoecker, "Lesion border detection in dermoscopy images," *Comput. Med. Imag. Graph.*, vol. 33, no. 2, pp. 148–153, 2009.
[8] C. Barata, M. E. Celebi, and J. S. Marques, "A survey of feature extraction in dermoscopy image analysis of skin cancer," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1096–1109, Jun. 2018.
[9] M. E. Celebi, N. Codella, and A. Halpern, "Dermoscopy image analysis: Overview and future directions," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 474–478, Jan. 2019.
[10] P. A. Lio and P. Nghiem, "Interactive atlas of dermoscopy," *J. Amer. Acad. Dermatol.*, vol. 50, no. 5, pp. 807–808, 2004.
[11] A. S. Ashour, Y. Guo, E. Kucukkulahli, P. Erdogmus, and K. Polat, "A hybrid dermoscopy images segmentation approach based on neutrosophic clustering and histogram estimation," *Appl. Soft Comput.*, vol. 69, pp. 426–434, Aug. 2018.
[12] A. S. Ashour, A. R. Hawas, Y. Guo, and M. A. Wahba, "A novel optimized neutrosophic k-means using genetic algorithm for skin lesion detection in dermoscopy images," *Signal, Image Video Process.*, vol. 12, no. 7, pp. 1311–1318, 2018.
[13] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: A review," *Artif. Intell. Med.*, vol. 56, no. 2, pp. 69–90, 2012.
[14] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1876–1886, Sep. 2017.
[15] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 994–1004, Apr. 2017.
[16] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Trans. Biomed. Eng*, vol. 64, no. 9, pp. 2065–2074, Sep. 2017.
[17] 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, Mar. 2015.
[18] 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 J. Biomed. Health Inform.*, vol. 22, no. 4, pp. 1218–1226, Aug. 2017.
[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Munich, Germany: Springer, 2015, pp. 234–241.
[20] 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)," 2016, *arXiv:1605.01397*. [Online]. Available: https://arxiv.org/abs/1605.01397
[21] 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 *Proc. 35th Annu. Int. Conf. Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 5437–5440.
[22] M. Goyal and M. H. Yap, "Multi-class semantic segmentation of skin lesions via fully convolutional networks," 2017, *arXiv:1711.10449*. [Online]. Available: https://arxiv.org/abs/1711.10449
[23] S. Vesal, S. M. Patil, N. Ravikumar, and A. K. Maier, "A multi-task framework for skin lesion detection and segmentation," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer, 2018, pp. 285–293.
[24] M. Goyal, J. Ng, A. Oakley, and M. H. Yap, "Skin lesion boundary segmentation with fully automated deep extreme cut methods," *Proc. SPIE*, vol. 10953, Mar. 2019, Art. no. 109530Q.
[25] A. Soudani and W. Barhoumi, "An image-based segmentation recommender using crowdsourcing and transfer learning for skin lesion extraction," *Expert Syst. Appl.*, vol. 118, pp. 400–410, Mar. 2019.
[26] 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," *Comput. Methods Programs Biomed.*, vol. 162, pp. 221–231, Aug. 2018.
[27] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, and H. Kittler, "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 *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172.
[28] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, Aug. 2018, Art. no. 180161.
[29] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.
[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[31] J. H. Ng, M. Goyal, B. Hewitt, and M. H. Yap, "The effect of color constancy algorithms on semantic segmentation of skin lesions," *Proc. SPIE*, vol. 10953, Mar. 2019, Art. no. 109530R.
[32] G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *Proc. Color Imag. Conf.*, no. 1, 2004, pp. 37–41.
[33] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.
[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Zurich, Switzerland: Springer, 2014, pp. 740–755.
[36] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[37] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
[38] M. Goyal and M. H. Yap, "Region of interest detection in dermoscopic images for natural data-augmentation," 2018, *arXiv:1807.10711*. [Online]. Available: https://arxiv.org/abs/1807.10711

MANU GOYAL received the M.Tech. degree in computer science and applications from Thapar University, India. He recently successfully defended his Ph.D. thesis entitled "Novel Computerised Techniques for Recognition and Analysis of Diabetic Foot Ulcers." This work was completed when he was with Manchester Metropolitan University. He is currently a Postdoctoral Research Associate with the Geisel School of Medicine, Dartmouth College. His research expertise is in medical imaging analysis, computer vision, deep learning, wireless sensor networks, and the Internet of Things.

AMANDA OAKLEY is currently a Dermatologist and an Adjunct Associate Professor with the Department of Medicine, The University of Auckland. She is best known for DermNet New Zealand, a popular online dermatological resource, and research in teledermatology and early diagnosis in melanoma.

PRIYANKA BANSAL received the M.C.A. degree from Punjabi University, India. She is currently a Research Intern with Manchester Metropolitan University, U.K. This work was completed when she was with Manchester Metropolitan University. Her research interests are in image processing, data analytics, and computer vision.

DARREN DANCEY is currently the Head of the Department of Computing and Mathematics. Since 2011, he has been with University as a Lecturer of computer science, where he held various positions with the Faculty of Science and Engineering. He has a research background in artificial intelligence with a Ph.D. degree in artificial neural networks/deep learning. His main contribution in this area was developing methods to address the so-called black-box problem of artificial intelligence systems to reveal how such systems are making decisions.

MOI HOON YAP (M'11–SM'19) received the Ph.D. degree in computer science from Loughborough University, in 2009. After the Ph.D., she worked as a Postdoctoral Research Assistant at the Centre for Visual Computing, University of Bradford, from April 2009 to October 2011. She is a Reader (Associate Professor) in computer vision with Manchester Metropolitan University and also a Royal Society Industry Fellow with Image Metrics Ltd., from 2016 to 2018. Her research expertise is computer vision, applied machine learning, and deep learning. She serves as an Associate Editor of *Journal of Open Research Software* and a reviewer of the IEEE TRANSACTIONS/Journals (Image Processing, Access, Medical Imaging, Biomedical Health, and Informatics).