C2

by Adi Wibowo

Submission date: 03-Apr-2023 07:34AM (UTC+0700)
Submission ID: 2053899581
File name: 1-s2.0-S2352914821001295-main.pdf (11.27M)
Word count: 9750
Character count: 54875

Informatics in Medicine Unlocked 25 (2021) 100640

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Lightweight encoder-decoder model for automatic skin lesion segmentation

Adi Wibowo [a,*], Satriawan Rasyid Purnama [a], Panji Wisnu Wirawan [a], Hanif Rasyidi [b]

[a] Department of Computer Science, Informatics, Diponegoro University, Semarang, Indonesia
[b] College of Engineering & Computer Science, Australian National University, Canberra, Australia

ARTICLE INFO

Keywords:
Skin lesion segmentation
Encoder-decoder
MobileNet
U-net
Random augmentation
Stochastic weight averaging

ABSTRACT

Accurate skin lesion segmentation (SLS) is an important step in computer-aided diagnosis of melanoma. Automatic detection of skin lesions in dermoscopy images is challenging because of the presence of artifacts and as lesions can have heterogeneous texture, color, and shape with fuzzy or indistinct boundaries. In this study, automatic SLS was performed using a lightweight encoder-decoder, MobileNetV3-UNet, which can achieve high accuracy with low resources. A comprehensive analysis was performed to improve the accuracy of the method in SLS. The semantic segmentation method consists of an encoder-decoder architecture, data augmentation, learning schemes, and post-processing methods. To enhance the SLS, we modified the decoder with the bidirectional ConvLSTM layer from the BCDU-Net and separable blocks from the separable-UNet architecture. Random augmentation was used to improve image diversity in the training dataset to avoid overfitting. Furthermore, a learning scheme based on stochastic weight averaging (SWA) was used to obtain better generalization by averaging multiple local optima. Our method was evaluated using three publicly available datasets, such as ISIC-2017, ISIC-2018, and PH2. We obtained dice coefficient and Jaccard index of 87.74%, 80.25%; 91.01%, 83.44%; and 95.18%, 91.08% for ISIC-2017, ISIC-2018, and PH2, respectively. The experimental results proved that the modified MobileNetV3-UNet method can outperform several state-of-the-art methods.

1. Introduction

Melanoma, a type of skin cancer, has a high mortality rate due to its highly metastatic nature [1]. Although it accounts for ~1% of skin cancer cases, most skin cancer deaths are from melanoma. It was expected that by 2021, 106 110 new melanoma cases will be diagnosed in the United States resulting in 7180 deaths [2]. The estimated five-year survival rate for melanoma is over 99% when diagnosed early, and ~14% when detected at an advanced stage [3]. Therefore, early detection is essential for treatment and prevention of metastasis, which improves prognosis.

Experts widely use dermoscopy to detect melanoma at an early stage. Dermoscopy is a noninvasive imaging technique that helps clinicians perform direct microscopy to observe diagnostic features in pigmented skin lesions [4]. This technique uses optical magnification, fluid immersion, and cross-polarized lighting to translate the epidermal layer, which improves diagnostic accuracy of melanoma, in comparison to conventional methods like asymmetry border color diameter (ABCD) technique [5]. However, diagnosis made by human vision requires considerable time, complex screening, and could be erroneous [6].

The development of computer-aided diagnosis (CAD) systems has aided early detection and analysis of pigmented skin lesions from dermoscopy images [6–9], while reducing time, cost, and subjectivity. Further, automatic segmentation of skin lesions using dermoscopy images can improve skin disease classification [10]. Using this approach, dermatologists can examine pigmented skin lesions and accurately localize cancerous areas. Because lesions can have fuzzy and indistinct boundaries, heterogeneous texture, color, shape, and other artifacts (Fig. 1), automatic segmentation remains challenging [11–13].

Recently, deep learning based on convolutional neural networks (CNNs) has gained prominence in machine learning and computer vision, particularly in semantic image segmentation [14]. The model adopts an encoder-decoder structure predicts pixel-to-pixel segmentation [15]. In the encoder, the input spatial resolution is reduced by downsampling, and low-resolution feature mappings (computationally efficient) that increase pixel-level discrimination are generated. Subsequently, the feature representations are upsampled to retrieve the full-resolution segmentation map in the decoder.

* Corresponding author.
E-mail addresses: bowo.adi@live.undip.ac.id (A. Wibowo), srp21.if@gmail.com (S.R. Purnama), panji@lecturer.undip.ac.id (P.W. Wirawan), hanif.rasyidi@anu.edu.au (H. Rasyidi).

https://doi.org/10.1016/j.imu.2021.100640
Received 26 February 2021; Received in revised form 31 May 2021; Accepted 13 June 2021
Available online 19 June 2021

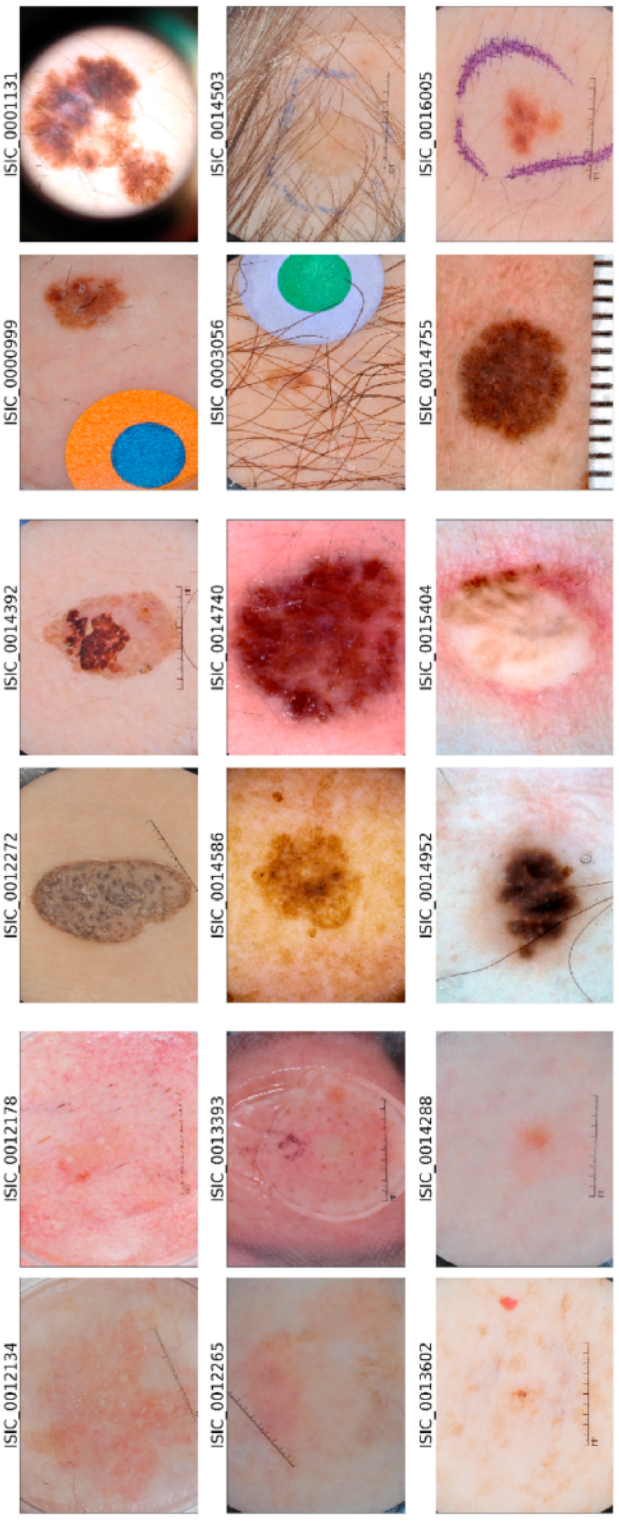

Fig. 1. Dermoscopy images that present challenges in automatic segmentation of skin lesions. (a) Fuzzy and indistinct lesion boundaries; (b) Heterogeneous texture, color, and shape of the lesions; (c) Artifacts. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1
ISIC-2017, ISIC-2018, and PH2 dataset specifications.

Dataset	ISIC-2017	ISIC-2018	PH2
Source	ISIC	ISIC	Hospital Pedro Hispano, Portugal
Total number of images	2750	2594	200
Size (Training/Validation)	2000/150	1815/259	160/40
Size (Test)	600	520	Averaged from the 5-fold cross validation
Image size (pixel)	556 × 679 to 4499 × 6748	556 × 679 to 4499 × 6748	577 × 769

Table 2
Random augmentation parameters.

Line	Operation	Parameter	Type of Augmentation	Probability
1	Horizontal Flip	–	Spatial	0.5
2	Shift, Scale, Rotate	Scale limit = 0.5, rotate limit = 0, shift limit = 0.1	Spatial	1.0
3	PadIfNeeded	Border mode = 4	Spatial	0.5
4	Random Crop	–	Spatial	1.0
5	CLAHE	Clip limit = 4.0, tile grid size = (8, 8) limit = 0.2	Pixel	0.9
	(Random Brightness, Gamma)	Gamma limit = (80, 120)		
6	Sharpen	Alpha = (0.2, 0.5), lightness = (0.5, 1.0)	Pixel	0.9
	Blur	Blur limit = 3		
	Motion Blur	Blur limit = 3		
7	Random Contrast	Hue limit = 0.2	Pixel	0.9
	Hue, Saturation, Value	Hue limit = 20, saturation limit = 30, value limit = 20		

U-Net [16] leverages data augmentation and is specially designed for medical imaging tasks with small datasets. Here, skip connections in the deep network architecture accelerate convergence and are essential for addressing vanishing gradient problems. The long skip connections can accurately capture the context to determine localized lesions in a symmetric expanding path. In a comparative study on real-time semantic segmentation [17], MobileNet was found to be more accurate than other encoders like ShuffleNet and ResNet 18. MobileNetV3 [18] is an improvement over MobileNetV2, and, with a lightweight architecture, is often the choice for segmentation. Additionally, short skip connections associated with the inverted residual bottleneck in the encoder effectively accelerate convergence of the learning process, specifically in deep network architectures with minimal parameters [19]. Due to the flexibility of the encoder-decoder concept, the deep network architecture can be efficiently used for biomedical image segmentation. Further, deep learning models for feature extraction from images may include transfer learning from pre-trained ImageNet weights [20]. Multiple studies on skin lesion segmentation (SLS) [21–23] have shown that network performance improves when pre-trained ImageNet weights are used as initial network weights.

The main objectives of this study are listed as follows:

(1) To utilize a lightweight encoder-decoder based on MobileNetV3 and U-Net for automatic SLS and improve the performance of the network architecture.
(2) To introduce modifications into the encoder and decoder and compare them with the standard U-Net architecture.
(3) To treat the variability in the visual appearance of skin lesions by combining several augmentation methods in SLS [24].
(4) To improve the segmentation map during testing using techniques like stochastic weight averaging (SWA) learning schema and filling in the hole (FITH) post-processing method.

2. Material and methods

2.1. Dataset

The publicly available datasets, International Skin Imaging Collaboration (ISIC) 2017 [25], 2018 [26], and the PH2 database [27] were used to evaluate the proposed SLS method. These datasets contain dermoscopy images with ground truth masks annotated by expert dermatologists. Description about the ISIC-2017, ISIC-2018, and PH2 datasets are provided in Table 1. The ISIC-2017 skin lesion challenge dataset contained training (2000), validation (150), and test (600) images. The image size varied from 556 × 679 pixels to 4499 × 6748 pixels. The ISIC-2018 skin lesion challenge dataset comprised 2594 images for training. This dataset was sequentially (not randomly) separated into training (1815), validation (259), and test sets [28]. The image size varied from 556 × 679 pixels to 4499 × 6748 pixels. The PH2 dataset had 200 dermoscopy images with 40 unique images in each fold for five-fold cross validation. Training was performed with four-folds of the data, while the rest was used for testing and validation. All images had an approximate size of 577 × 769 pixels.

2.2. Pre-processing and post-processing

2.2.1. Image resizing

To adjust variations in image size within the data sets (ISIC-2017, PH2), the images and their corresponding ground truths were resized to 192 × 256 pixels (height × width). Generally, a 3:4 (height: width) ratio is preferred for images [10,21]. For ISIC-2018, the images were resized to 256 × 256 pixels [28,29].

2.2.2. Image augmentation

The augmentation method is applied to images using the Albumentation library [30]. Two types of augmentation: pixel-level, which transforms images at the pixel-level (e.g., color), and spatial level, which transforms images at the spatial level (e.g., rotation), can be utilized. During segmentation, pixel transformation was applied to images, and spatial transformation was performed on images and ground truth masks. Variations in ambient conditions during dermoscopic screening or sampling can result in background illumination, or other extrinsic differences. Augmentation methods that normalize images, such as color constancy [24], can be applied before training, but overfitting can occur during deep learning. Here, variations in images were enlarged by random augmentation, and parameters were fine-tuned during training to obtain a more robust model. Pixel-level augmentations, such as random brightness, gamma, blur, sharpen, contrast, hue, saturation, value, and contrast limited adaptive histogram equalization (CLAHE), were utilized. The spatial-level augmentations, such as horizontal flipping, random crop, shifting, scaling, and rotation, were used randomly to create spatial variability. We have defined a sequence of operations that are executed based on a probability, as shown in Table 2. Some example images and masks before and after the random augmentation procedure are shown in Fig. 2.

2.2.3. Image normalization

The images and ground truth masks have a pixel size of 8-bit, and each pixel has a value between 0 and 255. Normalization was applied to each pixel in the images by dividing the input image by 255, and the normal pixel value range changed to 0–1. Specifically, the ground truth mask becomes binary (0 for background and 1 for foreground) by rounding up or ceiling.

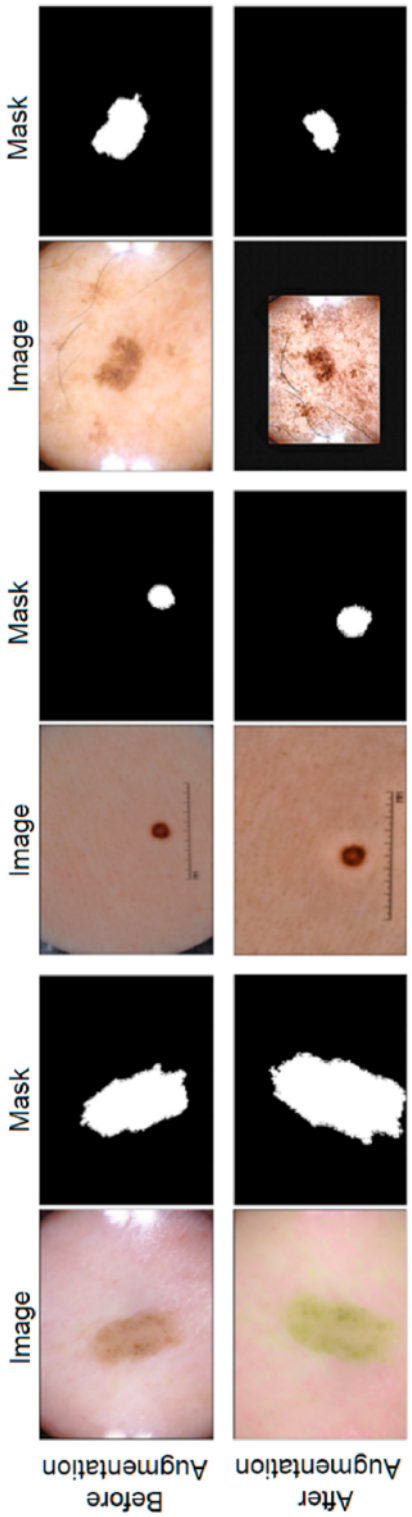

Fig. 2. Example images and ground truth masks before and after the random augmentation procedure.

Table 3
Pre- and post-processing performed on the ISIC-2017, ISIC-2018, and PH2 datasets.

Dataset	Methods	ISIC-2017	ISIC-2018	PH2
Pre-processing (training)	Resizing	192x256x3	256x256x3	192x256x3
	Augmentation	Yes	Yes	Yes
	Normalization	0–1	0–1	0–1
Pre-processing (validation or testing)	Resizing	192x256x3	256x256x3	192x256x3
	Augmentation	No	No	No
	Normalization	0–1	0–1	0–1
Post-processing (training)	FITH	No	No	No
Post-processing (validation)	FITH	No	No	No
Post-processing (testing)	FITH	Yes	No	No

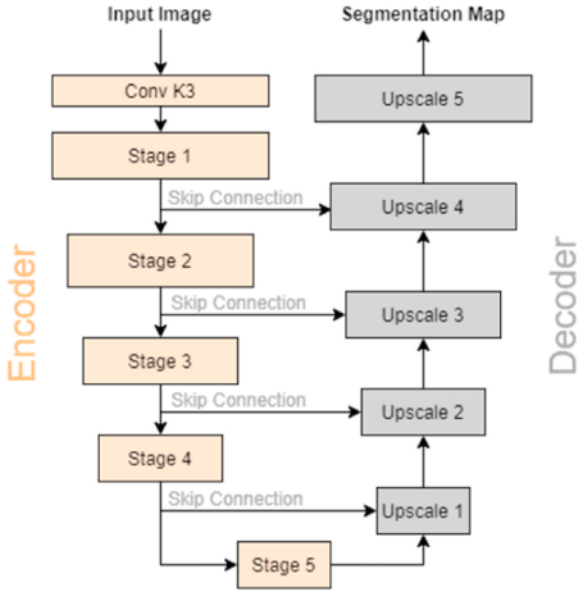

Fig. 3. Proposed decoupled encoder and decoder modules in an end-to-end semantic segmentation architecture.

2.2.4. Post-processing

A commonly used post-processing algorithm for segmentation is FITH. We used the FITH method described in Ref. [21] to improve segmentation. FITH processes holes in the segmentation output to properly define the lesion boundary. The pre- and post-processing details for ISIC-2017, ISIC-2018, and PH2 datasets are listed in Table 3.

2.3. Network structure

The proposed model consists of a decoupled encoder and decoder module, which were combined in an end-to-end semantic segmentation architecture based on U-Net. The U-Net architecture that utilizes skip connections can make the model more robust. The encoder was modified using the lightweight MobileNetV3 feature extraction model. Subsequently, we studied the effect of the short skip connection (inverted residual bottleneck) and the NAS module on the encoder. In the proposed architecture, the skip connection connects the encoder and decoder at four stages, and in the last stage, the encoder and decoder are directly connected (Fig. 3).

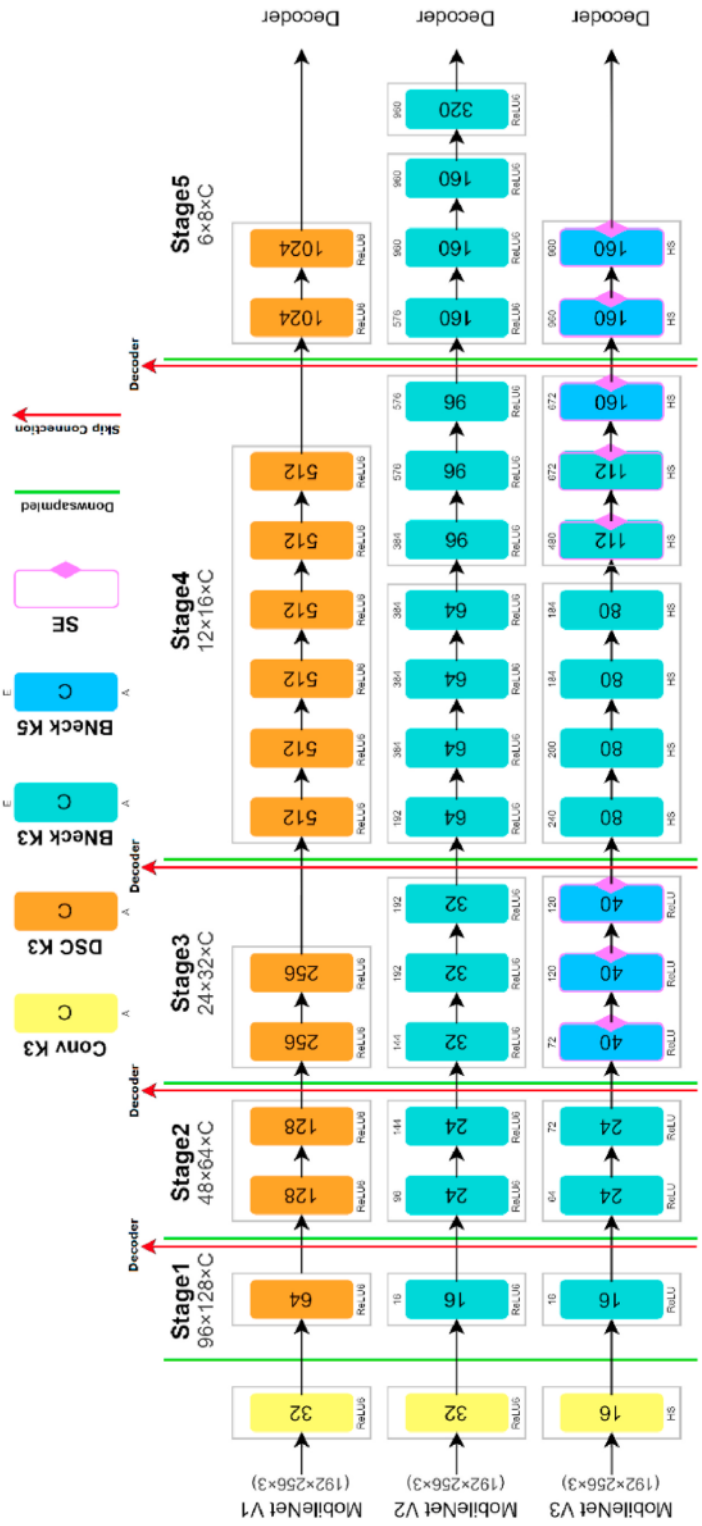

Fig. 4. Detailed encoder architectures, and differences between MobileNetV1, MobileNetV2, and MobileNetV3.

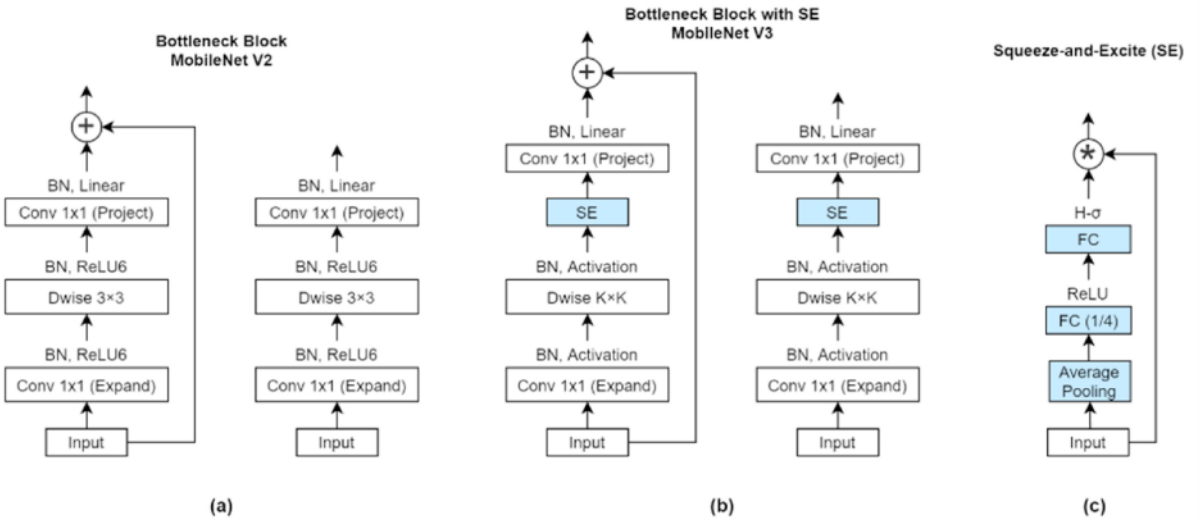

Fig. 5. (a) MobileNetV2 bottleneck block (Inverted Residual Bottleneck), (b) MobileNetV3 bottleneck block, and (c) squeeze-and-excite module in MobileNetV3. The left unit of the bottleneck is used for short skip connection, and the right unit is used before downsampling, which are without the short skip connection.

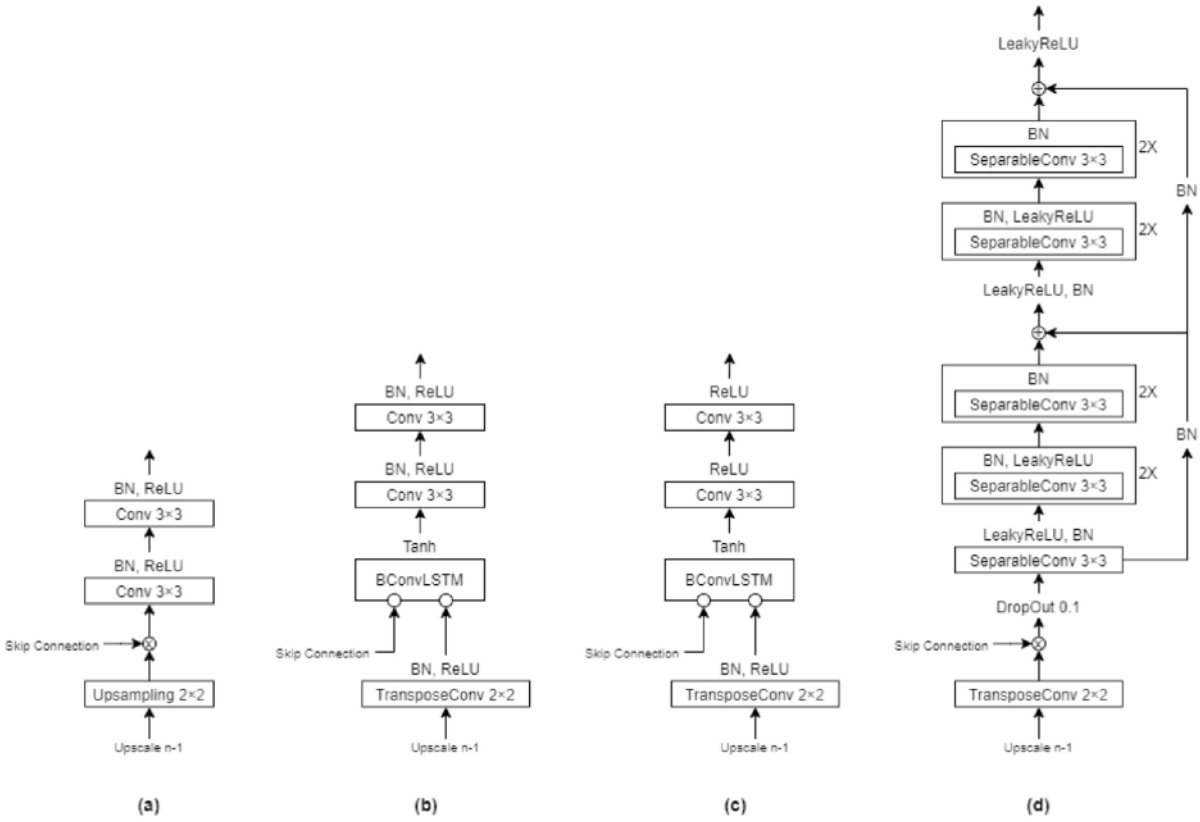

Fig. 6. Blocks used at each stage in the decoder. (a) Standard U-Net, (b) UNet-LSTM, (c) BCDU, (d) Separable-UNet.

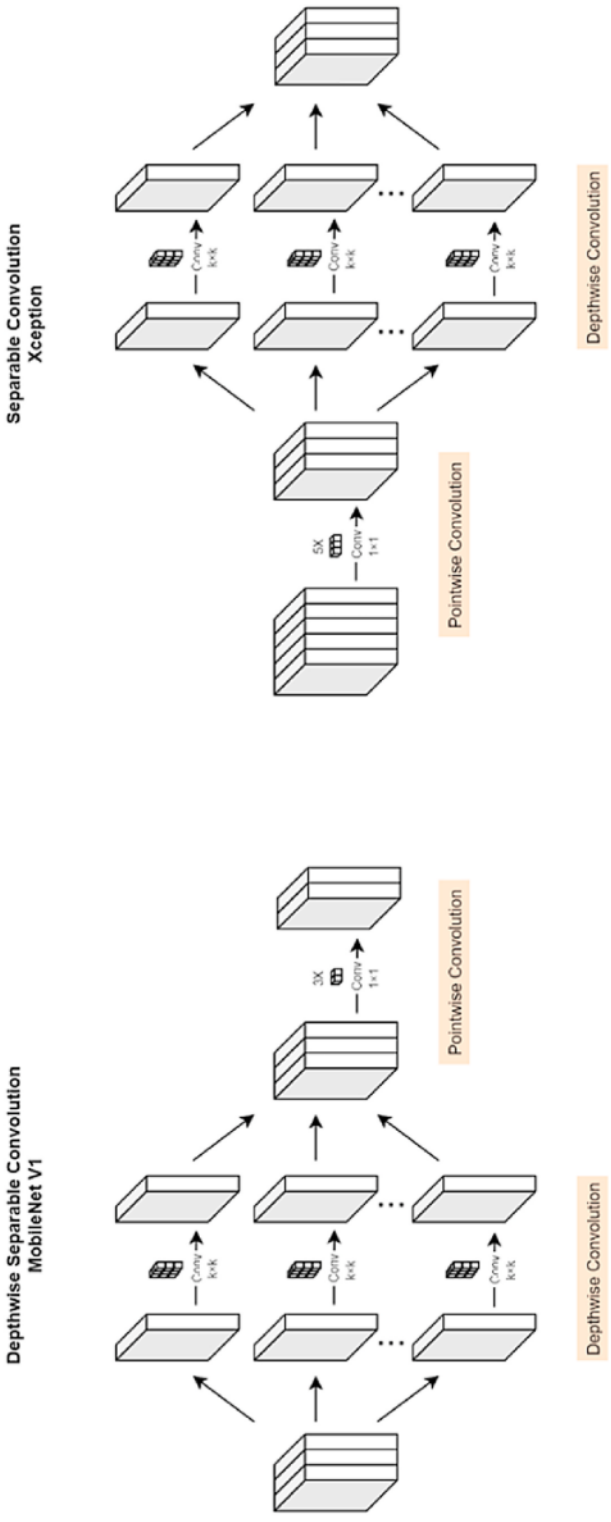

Fig. 7. Different structures of depthwise separable convolution (a) and separable convolution (b).

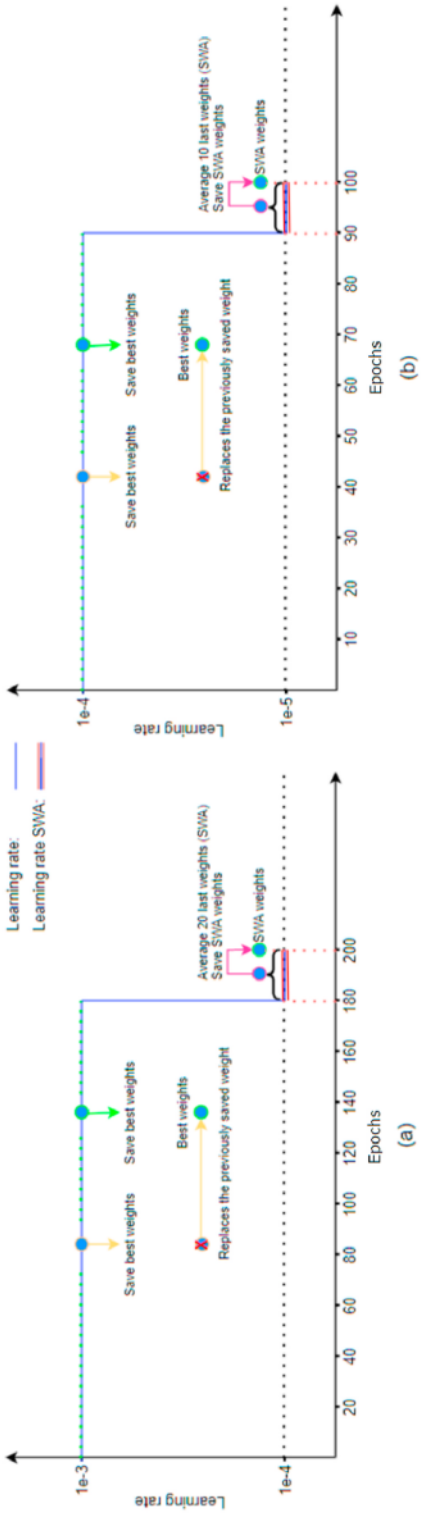

Fig. 8. (a) Learning schema for ISIC-2017 and PH2 dataset and (b) ISIC-2018 dataset.

Table 4
Parameters for training and testing using the ISIC-2017, ISIC-2018, and PH2 datasets.

Dataset	ISIC-2017	ISIC-2018	PH2
Batch size	8	8	8
Epochs	200	100	200
Learning rate	0.001	0.0001	0.001
Learning scheme	SWA & best validation	SWA & best validation	SWA & best validation
SWA epochs	181–200	91–100	181–200
SWA learning rate	0.0001	0.00001	0.0001
Pre-trained weights	ImageNet	ImageNet	ImageNet
Loss function	Jaccard loss	Jaccard loss	Jaccard loss

2.3.1. Encoder

To obtain semantic information from the original image, the proposed encoder uses the MobileNet architecture for feature extraction. The MobileNetV1, MobileNetV2, and MobileNetV3 architectures have been outlined in Fig. 4. Thus, the encoder serves as a feature extractor in semantic segmentation architecture. High accuracy and lightness are essential characteristics for extractor. The skip connections that connect the encoder and decoder are standard when the network stages reduce the spatial dimension (Fig. 4). We compared the effects of several modules available in each MobileNet version. MobileNetV1 architecture [31] uses depthwise separable (orange block in Fig. 4) and pointwise convolution instead of the standard convolution (yellow block in Fig. 4). Depthwise and pointwise convolution, respectively, were followed by batch normalization and rectified linear unit (ReLU6) activation function.

MobileNetV2 [32] uses depthwise separable convolution, residual connections with linear bottlenecks (green block in Fig. 4), and an inverted residual structure. Here, the layer structure is more efficient owing to its low-level problem properties (Fig. 5a).

MobileNetV3 [18] is composed of MobileNetV2 structure, MnasNet's inverted residual bottleneck layer [33], and light network architecture search modules based on squeeze-and-excitation (SE) in the bottleneck structure. SE can improve accuracy by increasing the number of parameters and reducing visible latency. The bottleneck of MobileNetV3 (kernel size of 5×5) along with an SE module is depicted in Fig. 4 (blue block), Fig. 5b, and Fig. 5c. The SE layer was enhanced with a modified swish nonlinearity (purple border in Fig. 4). Further, SE and swish nonlinearity that use a sigmoid were replaced with a hard-sigmoid (H_σ). Because the sigmoid function is computationally heavy compared to ReLU6, hard-sigmoid, as shown below, was utilized.

$$H\_\sigma(x) = \frac{ReLU6(x+3)}{6} \tag{1}$$

2.3.2. Decoder

In the decoder, we used the U-Net-based architecture. The decoder produced segmentation maps at full resolution. With the skip connection feature at each stage, a decoder can be modified with several extension modules or modifying blocks to improve its performance. Extensions and modifications were explored to determine the advantages of each module. We utilized multiple decoders (Fig. 6), including standard U-Net (no additions or modifications), UNet-LSTM (BCDU–Net's [28] decoder with batch normalization in each convolution), BCDU (BCDU–Net's [28] decoder), and separable-UNet (Separable-UNet's [21] decoder).

2.3.2.1. Standard U-Net. The block is shown in Fig. 6a. The concatenation layer restored image features that were lost when passing through various convolutional layers until they are sufficiently deep. The next layer is twice the 3×3 convolution layer, followed by an upsampling operation to be processed in the next block. Each layer was

Table 5
Comparison of different encoder networks using the ISIC-2017 test dataset.

Model	Convolution layer	ACC	DIC	JAI	SEN	SPE	Computation time (s)	Model parameters
VGG16-UNet	Standard convolution	0.9248	0.8409	0.7595	0.7864	**0.9716**	0.0277	2.38×10^7
ResNet50-UNet	Residual network	0.9304	0.8608	0.7816	0.8200	0.9700	0.0230	3.26×10^7
MobileNetV1-UNet	Depthwise separable	0.9352	0.8690	0.7929	0.8509	0.9637	**0.0157**	8.34×10^6
MobileNetV2-UNet	Inverted residual	0.9366	0.8714	0.7941	0.8506	0.9657	0.0197	$\mathbf{8.05 \times 10^6}$
MobileNetV3-UNet	Inverted residual +SE	**0.9381**	**0.8774**	**0.8025**	**0.8624**	0.9636	0.0199	8.27×10^6

Table 6
Comparison of different decoder networks using the ISIC-2017 test dataset.

Model	ACC	DIC	JAI	SEN	SPE	Computation time (s)	Model parameters
MobileNetV3-BCDU	0.9314	0.8612	0.7796	0.8429	0.9613	0.0274	2.16×10^7
MobileNetV3-UNet-LSTM	0.9336	0.8618	0.7838	0.8195	**0.9722**	0.0272	1.81×10^7
MobileNetV3-Separable-UNet	**0.9385**	0.8753	0.7961	0.8523	0.9678	0.0243	$\mathbf{6.29 \times 10^6}$
MobileNetV3-UNet	0.9381	**0.8774**	**0.8025**	**0.8624**	0.9636	**0.0199**	8.27×10^6

Table 7
Comparison between different augmentation strategies using the ISIC-2017 test dataset and MobileNetV3-UNet.

Augmentation strategy	Spatial augmentation		Pixel augmentation				JAI
	Distortion only	Without distortion	Color jitter	CLAHE	Blur	Sharpen with gamma	
None	–	–	–	–	–	–	0.7770
Spatial only	–	✓	–	–	–	–	0.7691
Spatial augmentation with distortion [21]	✓	✓	–	–	–	–	0.7420
Spatial augmentation without distortion [57]	–	✓	✓	–	–	–	0.7782
CLAHE only	–	–	–	✓	–	–	0.7689
Blur only	–	–	–	–	✓	–	0.7742
Sharpen with gamma	–	–	–	–	–	✓	0.7805
Proposed	–	✓	✓	✓	✓	✓	**0.8025**

Table 8
Comparison of different decoder networks with and without augmentation of images in the ISIC-2017 test dataset.

Model	JAI	
	None	Augmentation
MobileNetV3-BCDU	0.7121	0.7796
MobileNetV3-UNet-LSTM	0.7436	0.7838
MobileNetV3-Separable-UNet	0.7682	0.7961
MobileNetV3-UNet	**0.7770**	**0.8025**

Table 9
Comparison of test results with training schemes for each model for ISIC-2017 dataset.

Training Schema	Model	JAI
Standard learning with best validation	MobileNetV3-BCDU	0.7806
	MobileNetV3-UNet-LSTM	0.7824
	MobileNetV3-Separable-UNet	0.7834
	MobileNetV3-UNet	**0.7923**
SWA (constant, last 20 epochs)	MobileNetV3-BCDU	0.7796
	MobileNetV3-UNet-LSTM	0.7838
	MobileNetV3-Separable-UNet	0.7961
	MobileNetV3-UNet	**0.8025**

Table 10
Model (MobileNetV3-UNet) performance with and without post-processing.

Post-processing	ACC	DIC	JAI	SEN	SPE
None	0.9381	0.8751	0.8023	0.8588	**0.9648**
FITH	**0.9381**	**0.8774**	**0.8025**	**0.8624**	0.9636

supplemented with batch and ReLU normalization to accelerate network convergence, training, and nonlinearity. Batch normalization values were calculated using the following equation:

$$y_i = \gamma \frac{x_i - \lambda_B}{\sqrt{\sigma_B^2 + \varepsilon}} + \beta \quad (2)$$

2.3.2.2. UNet-LSTM. Based on the BCDU-Net architecture [28], the bidirectional ConvLSTM module was implemented. Here, the decoder was combined with the feature map extracted from the encoder via the skip connection and the previous decoder block. We used the bidirectional ConvLSTM layer as the standard U-Net decoder extension, namely UNet-LSTM (Fig. 6b). We wanted to determine the effect of the bidirectional ConvLSTM layer on the segmentation capacity. To equalize the number of feature map channels processed by the bidirectional ConvLSTM layer, the upsampling layer was replaced with a transpose convolution. Bidirectional ConvLSTM uses two ConvLSTMs to process the input data from the skip connection and upsamples the decoding path into two-way forward and reverse paths. Then, it makes decisions for those inputs by handling the dependency of the data in both directions. In the standard ConvLSTM, only the forward-direction dependencies are processed. However, all the information in a sequence should be considered, and accounting for backward dependencies is effective. Further, analyzing both forward and backward dependencies from a temporal perspective improves predictive performance [34]. The final output considered bidirectional spatio-temporal information in the presence of a hyperbolic tangent, which was used to combine the outputs from the forward and reverse states in a non-linear manner.

2.3.2.3. BCDU. The entire block of the original BCDU with four skip connections (Fig. 6c) was utilized in the proposed architecture (Fig. 3). In contrast to UNet-LSTM, batch normalization was performed after

Table 11
Comparison of model performances using the ISIC-2017 dataset.

Method	ACC	DIC	JAI	SEN	SPE	Computation Time	Model Parameters
Res-UNet [22]	–	0.8580	0.7720	–	–	–	–
DCL-PSI [52]	0.9408	0.8566	0.7773	0.8620	0.9671	–	–
SU-SWA [21]	0.9431	0.8693	0.7926	0.8953	0.9632	0.0440-s	1.94×10^7
Ensemble-A [53]	0.9410	0.8710	0.7930	**0.8990**	0.9500	–	–
DAGAN [11]	0.935	0.859	0.771	0.835	**0.976**	–	–
ASCU-Net [12]	0.926	0.830	0.742	0.825	0.953	–	–
DL-AuxiliaryTask [13]	**0.9432**	0.8713	0.7946	0.8876	0.9651	–	–
MobileNetV2-UNet (Proposed)	0.9366	0.8714	0.7941	0.8506	0.9657	**0.0197-s**	8.05×10^6
MobileNetV3-Separable-UNet (Proposed)	0.9385	0.8753	0.7961	0.8523	0.9678	0.0243-s	$\mathbf{6.29 \times 10^6}$
MobileNetV3-UNet (Proposed)	0.9381	**0.8774**	**0.8025**	0.8624	0.9636	0.0199-s	8.27×10^6

Table 12
Comparison of model performances using the ISIC-2018 dataset.

Method	ACC	DIC	JAI	SEN	SPE	Computation time	Model parameters
U-Net [16]	0.890	0.647	0.549	0.708	0.964	–	–
Attention U-Net [54]	0.897	0.665	0.566	0.717	0.967	–	–
R2U-Net [54]	0.880	0.679	0.581	0.792	0.928	–	–
Attention R2U-Net [54]	0.904	0.691	0.592	0.726	0.971	–	–
BCDU-Net (d = 3) [28]	0.937	0.851	–	0.785	**0.982**	0.0328-s	2.07×10^7
Double-UNet [23]	–	0.8962	0.8212	–	–	–	–
MobileNetV3-BCDU (Proposed)	0.9466	0.9060	0.8281	0.8903	0.9695	0.0274-s	2.16×10^7
MobileNetV3-UNet-LSTM (Proposed)	0.9456	0.9050	0.8265	0.8970	0.9654	0.0272-s	1.81×10^7
MobileNetV3-Separable-UNet (Proposed)	**0.9485**	0.9073	0.8315	0.9011	0.9643	0.0243-s	$\mathbf{6.29 \times 10^6}$
MobileNetV3-UNet (Proposed)	0.9479	**0.9098**	**0.8344**	**0.9089**	0.9638	**0.0199-s**	8.27×10^6

Table 13
Comparison of model performances using the PH2 dataset.

Method	ACC	DIC	JAI	SEN	SPE	Computation time	Model parameters
Res-UNet [22]	–	0.924	0.854	–	–	–	–
Ensemble-S [53]	0.938	0.907	0.839	0.932	0.929	–	–
DCL-PSI [52]	0.9661	0.9413	0.8605	0.9711	0.9585	–	–
SU-SWA [21]	0.9669	0.9413	0.8940	0.9651	0.9526	0.0440-s	1.94×10^7
ASCU-Net [12]	0.943	0.909	0.800	0.926	0.945	–	–
MobileNetV3-UNet (Proposed)	**0.9870**	**0.9518**	**0.9108**	**0.9892**	**0.9789**	**0.0199-s**	$\mathbf{8.27 \times 10^6}$

transpose convolution. Additionally, a convolutional layer with three channel outputs was implemented before the final convolution layer. The original BCDU block was utilized to determine the effect of batch normalization in each convolution and an additional convolution layer.

2.3.2.4. Separable-UNet. Separable convolutional blocks are used as base blocks in the Xception architecture [35]. Xception performs depthwise separable convolution, which involves depthwise convolution followed by a pointwise convolution, namely separable convolution (Fig. 7). Separable-UNet [21] is a U-Net-based network architecture. The standard convolution layer was replaced by a separable convolutional block (SCB) layer based on Xception. We used the SCB of separable-UNet in our decoder block, as shown in Fig. 6d. A separable convolutional block can improve discrimination between pixel-level representations and reduce the computational complexity of the decoder.

2.4. Stochastic weight averaging

The use of model weights during validation is a common training scheme. However, if the training dataset is unbalanced, ambiguous, and small, models can be overfitted with solutions on the local surface. SWA [36] incorporates weights in the training scheme for the last few epochs by calculating average values for the weights. This method addresses the optimal local problem by averaging several weights (ensemble of points in the weight space) in stable conditions, which is an improvement over the traditional ensemble. The global optimum solution is obtained by widening the local surface point, and the midpoint is considered as the global optimum. This improves SWA and the learning rate. In this method, training can be performed at a cyclic learning rate with a cosine annealing scheduler, wherein several weight points are averaged at the end of cycles as optima. Additionally, when training is performed with a small constant learning rate, the terminal weight points from a constant learning rate scheduler are averaged. Here, we implemented SWA with a constant learning rate, as shown in previous studies on SLS [21].

2.5. Training and testing

Initially, a dataset is used for training, and a second dataset is used for validation. The first approach monitored the validation data scores, and the highest scores were stored as the model weights during training. However, SWA does not monitor the validation score; the average model weights determined in the terminal epochs are used as the final weights. Thus, the standard method of learning was combined with SWA validation method during training, and the results were evaluated. The learning schema for all the datasets is shown in Fig. 8.

All the models were implemented using the Keras framework [37]. Training was performed on an i7 processor with NVIDIA RTX 2060. Details about the training and testing procedure performed details using

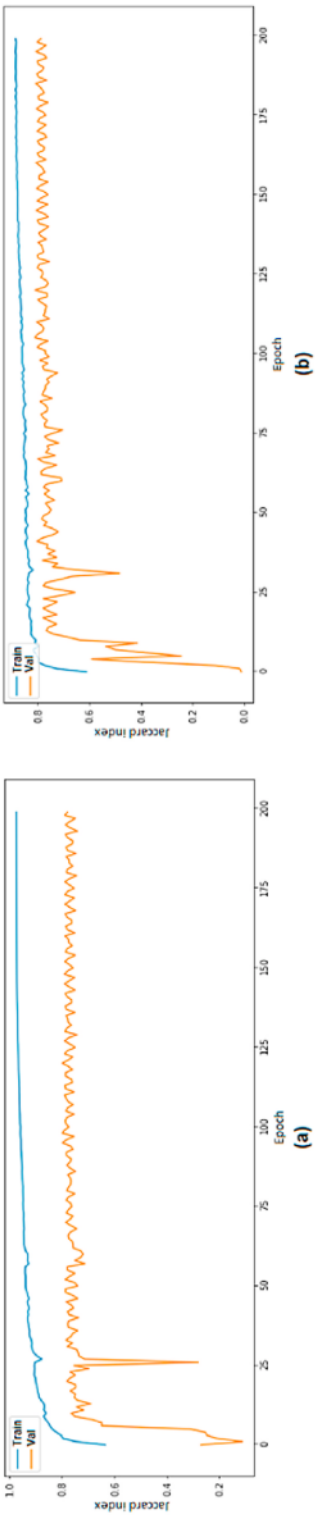

Fig. 9. Training and validation Jaccard index (*JAI*) of MobileNetV3-UNet without augmentation (a) and with augmentation (b).

the ISIC-2017, ISIC-2018, and PH2 datasets are shown in Table 4. The total epochs for ISIC-2017 and PH2 were 200, while for ISIC-2018 it was 100. The training process used adaptive moment estimation (Adam) [38] to optimize the model with a learning rate of 1e-3 for ISIC-2017 and PH2, and 1e-4 for ISIC-2018. The batch size value was 8. Image sizes for ISIC-2017 and PH2 were 192 × 256 × 3, while for ISIC-2018 it was 256 × 256 × 3. Further, pre-trained model weights on ImageNet were utilized. For the loss function, the Jaccard loss $L_{jaccard}$ [10], which is a complement of the Jaccard index (JAI), was utilized. If G is the ground truth and P is the segmentation result of the model, then $L_{jaccard}$ is determined by:

$$L_{jaccard} = 1 - JAI = 1 - \frac{|G \cap P|}{|G| + |P| - |G \cap P|} \tag{3}$$

2.6. Performance evaluation

Five common evaluation metrics [21–23] were used to evaluate the proposed segmentation method: accuracy (*ACC*), sensitivity (*SEN*), specificity (*SPE*), Dice coefficient (*DIC*), and Jaccard index (*JAI*). The evaluation metrics were formulated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$SEN = \frac{TP}{TP + FN} \tag{5}$$

$$SPE = \frac{TN}{TN + FP} \tag{6}$$

$$DIC = \frac{2*TP}{2*TP + FN + FP} \tag{7}$$

$$JAI = \frac{TP}{TP + FN + FP} \tag{8}$$

where TP, TN, FN, and FP denote true positive, true negative, false negative, and false positive, respectively. TP represents the number of correctly segmented lesion pixels, while FN represents unsegmented lesions. TN represents the number of unsegmented non-lesion pixels, while segmented non-lesion pixels were FP.

3. Theory

Several traditional unsupervised and supervised methods for segmenting skin lesions using dermoscopy images are available. The unsupervised methods include thresholding [39,40] region-merging [9,41], energy functions [42,43], and clustering [44,45]. Unsupervised methods are advantageous because data labelling is not required; however, large and distinct datasets are difficult to handle. It is challenging to deal with fuzzy pigment boundaries and complicated skin conditions with the unsupervised methods. Moreover, these methods involve many intermediate steps that depend on the data [13]. Supervised methods focus on pixel feature or region extraction and classification of lesions and normal tissues. Xie et al. [46] extracted RGB color features using a self-generating neural network classifier and a genetic algorithm. He et al. [47] determined texture features with Gabor and gray level co-occurrence matrix (GLCM) features and an SVM classifier. However, this traditional supervised method relies on low-level features, such as color and texture, and cannot capture semantic information from a high-level image. Further, its performance is based on multiple parameters and data pre-processing steps, which makes it highly complex [21,48], and limits its generalization ability. However, complex pre-processing and semantic segmentation can be handled with the deep learning CNNs.

In the deep learning methods, which involve pre-processing and semantic segmentation in a series, the encoder-decoder architecture that

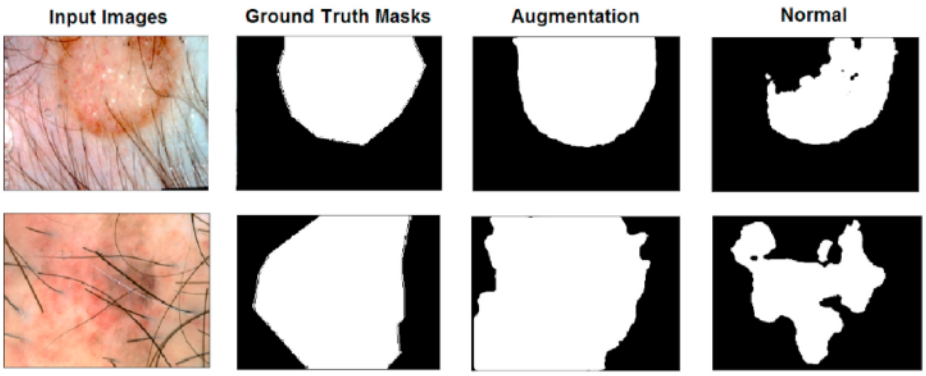

Fig. 10. Segmentation obtained from MobileNetV3-UNet model training. From left to right, input images, ground truth masks, augmentation, and normal or without augmentation (all were without post-processing) are shown.

predicts pixel segmentation has shown considerable success. An encoder extracts the feature map (downsampling), which is similar to an image classification mechanism without flattening. A decoder adjusts the resolution (upsampling) of the feature map to obtain an image of the original size. The fully convolutional network (FCN) was the first encoder-decoder architecture model that solves the SLS problem developed with a very deep residual 50-layer [49]. Then, the FCN was updated with a new loss function based on Jaccard distance [10], multi-scale [50], and multi-stage [51]. However, their studies produced inconsistent segmentation issues on different types of skin lesions that overfit the dominant non-melanoma studies with a low result on melanoma segmentation. This problem was addressed using a stepwise integration method with an ensemble of three FCN models [52]. However, this method still has challenges in dealing with complex lesion images and increasing computational complexity during training. Goyal et al. [53] used non-FCN models with an ensemble of mask R–CNN and DeeplabV3+. However, this method requires high complexity. Another non-FCN is the pyramid pooling network method with a multi-task learning approach [13]. This architecture consists of a feature extractor and a pyramid pooling module (PPM) connected to a parallel cross-connection layer (CCL) architecture. This method uses two multi-scale feature aggregation (MSFA) modules to aggregate information from feature maps of different scales. In addition, it edged prediction as an auxiliary task to help in the segmentation task. However, segmentation and edge detection results are slightly blurred at the lesion boundary, which can reduce the segmentation performance. To solve the problems in biomedical image segmentation, such as a limited dataset and a high level of image difficulty, U-Net is quite robust in studying the features at each depth level.

Several U-Net improvement methods for SLS have focused on improving efficiency and feature discrimination, that is, modifying the encoder-decoder block part and adding modules. In the modification encoder-decoder, Tang et al. [21] used Xception as an encoder and replaced the decoder block using a separable convolutional block layer. Zafar et al. [22] used ResNet50 as an encoder to produce a deeper architecture. The adding module mechanism in the U-Net involved sequential modules to combine spatial-temporal information and attention modules to increase important information. Alom et al. [54] used a recurrent convolutional neural network and recurrent residual convolutional neural network as sequential modules in U-Net. To capture important spatial and temporal features from the upscaling and skip connection layer, Azad et al. [28] proposed a bidirectional convolutional LSTM module and added three densely connected convolutions to mitigate the problem of learning redundant features in successive convolutions for the encoder. Tong et al. [12] proposed the attention gate spatial and channel attention U-Net (ASCU-Net) that uses triple attention mechanism to capture the contextual information, spatial correlation between features, and a more relevant field of view of the target. A combination of the modification encoder-decoder block and adding module was proposed by Lei et al. [11], which integrated the dense convolution U-Net (UNet-SCDC) and the dual discrimination module in the generative adversarial network (GAN) mechanism. The optimization has been performed without considering computational efficiency, and it is challenging to perform on mobile or real-time devices. An interesting approach is MobileNet, which supports mobile computer-aided devices [55].

MobileNet for a semantic segmentation model was performed for the first time by Siam et al. [17]. The results show that MobileNetV1 can obtain the highest performance with minimum computational complexity as an encoder. MobileNetV1 introduced depthwise convolution to reduce the number of parameters, followed by 1×1 pointwise convolution to aggregate the feature information of each channel in each pixel; this is called depthwise separable convolution. The evolution of MobileNet is quite notable; MobileNetV2 modifies the idea of the classic bottleneck structure from the ResNet architecture and connects shortcuts between linear bottlenecks. The inverted residual bottleneck mainly improves the accuracy and optimizes the complexity model. The latest MobileNetV3 [18] was enhanced through the network architecture search, and the sigmoid activation function was optimized to be hard-sigmoid to reduce computational complexity and increase accuracy. MobileNetV3 has different channel expansion levels at each bottleneck block.

Inspired by the encoder-decoder architecture on U-Net and the efficiency of the MobileNetV3 model, we propose an automatic SLS architecture by combining MobileNetV3 as an encoder with U-Net as a decoder to increase the efficiency of an SLS. MobileNetV3 is a solution for efficient architecture. BCDU and Separable-UNet also inspired us to explore modifications to the decoder by adding some temporal modules, such as the BCDU, and applying block modifications, such as the one in Separable-UNet. From this exploration, the MobileNetV3-UNet, MobileNetV3-Separable-UNet, MobileNetV3-BCDU, and MobileNetV3-LSTM-UNet architectures were obtained.

4. Results

4.1. Experimental results

We compared the critical component effect in the proposed SLS method, which comprised the encoder and decoder network, random augmentation, learning scheme, and post-processing method. A comparative experiment was performed using the ISIC-2017 dataset, which is a challenging SLS dataset with a training, validation, and test

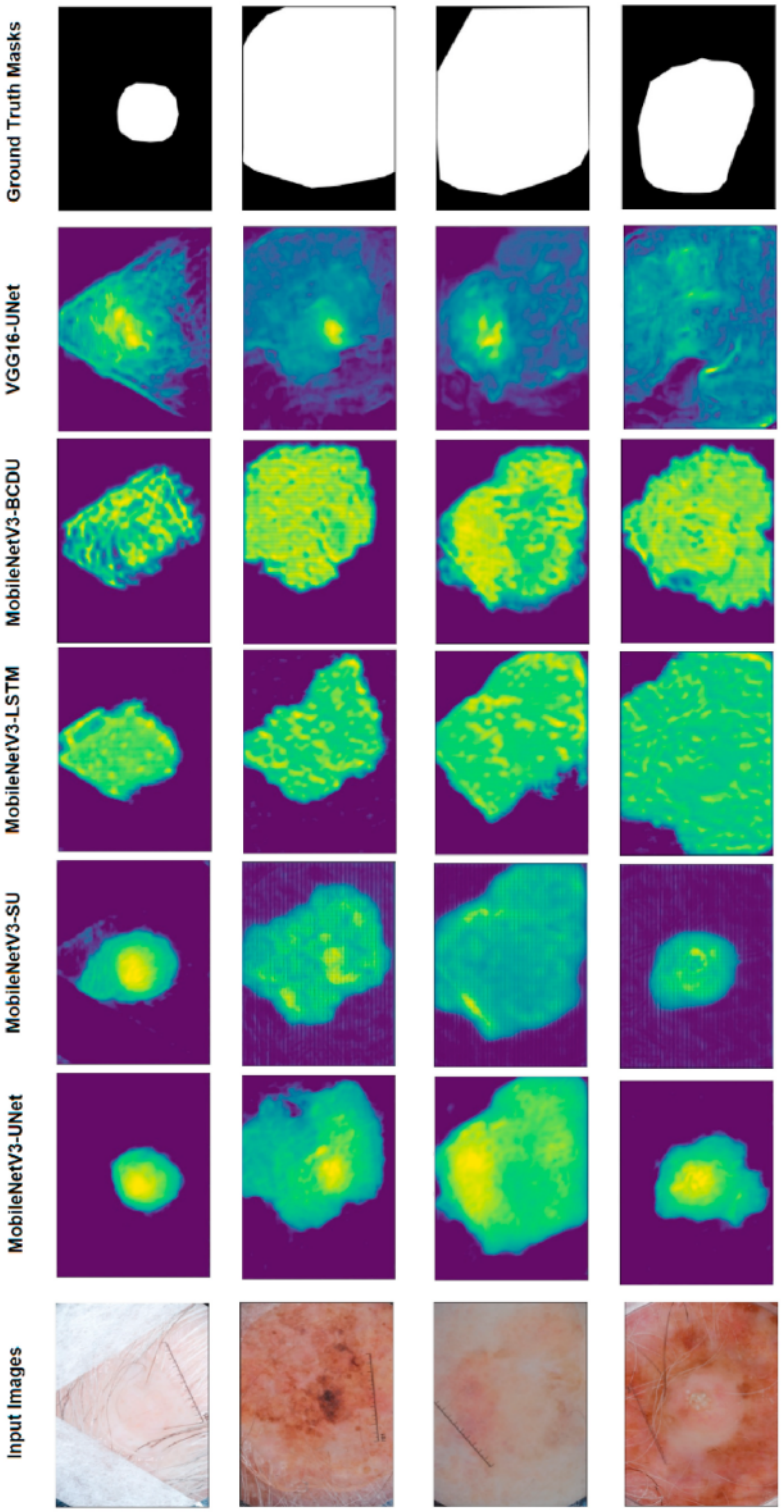

Fig. 11. Class activation mapping (CAM) of the last convolutional layers (randomly picked from one of 16 channels). From left to right, input images, MobileNetV3-UNet, MobileNetV3-Separable-UNet, MobileNetV3-LSTM-UNet, MobileNetV3-BCDU, VGG16-UNet images, and ground truth masks are shown.

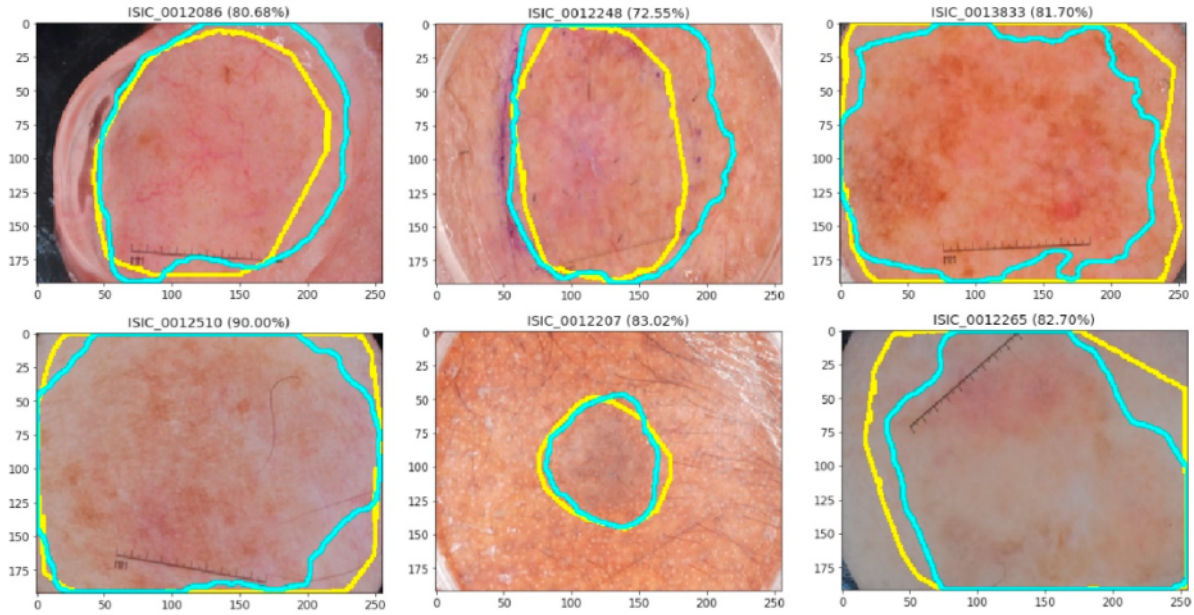

Fig. 12. Example segmentation masks with MobileNetV3-UNet for analyzing challenging images in SLS. The cyan line denotes the segmentation masks, and the yellow line represents the ground truth masks. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

set.

4.1.1. Encoder network comparison

Different encoders were compared in the U-Net architecture, such as VGG16, ResNet50, and MobileNet (multiple versions). Further, a combination of random augmentation and the SWA learning scheme were utilized. The VGG16 architecture consists of several standard convolutional blocks that resemble the vanilla U-Net encoder, and the ResNet50 architecture consists of state-of-the-art residual bottleneck [56]. However, MobileNet uses deep separable convolution layers to reduce the number of parameters. Table 5 shows that MobileNet performs better than VGG16 and ResNet50. Here, VGG16 and ResNet50 had JAI values of 75.95% and 78.16%, respectively. However, MobileNetV3, MobileNetV2, and MobileNetV1 obtained JAI values of 80.25%, 79.41%, and 79.29%, respectively. Based on the SEN and SPE data, we believe that the VGG16 encoder provides high dominance in the background than in the foreground (segmentation mask). However, MobileNet focusses on the foreground and produces better segmentation results, as reflected in the obtained ACC, DIC, and JAI values. Considering the number of parameters, MobileNetV1 utilizes less parameters than MobileNetV3 or MobileNetV2. Further, VGG16 and ResNet50 utilize higher number of parameters than all the MobileNet versions.

4.1.2. Decoder network comparison

We modified the U-Net-based decoder using other decoders, such as BCDU, UNet-LSTM, and separable-UNet. Here, MobileNetV3 was paired with the decoder, and its performance in image analysis was determined. The models for comparison were prepared with different decoders, random augmentation, and SWA. The MobileNetV3 with standard U-Net decoder outperformed other decoders with DIC, JAI, and SEN values of 87.74%, 80.25%, and 86.24%, respectively (Table 6). Moreover, MobileNetV3-Separable-UNet had ACC of 93.85% and MobileNetV3-UNet-LSTM had SPE of 97.22%. Thus, SEN values suggested that MobileNetV3-UNet was better at segmenting the skin lesion area. MobileNetV3-LSTM-UNet and MobileNetV3-BCDU utilized more parameters than MobileNetV3-UNet. Although MobileNetV3-Separable-UNet utilizes less parameters than the one with standard U-Net, the complex operation results in a large latency in computation time.

4.1.3. Random augmentation and model performance

We implemented spatial- and pixel-level augmentation methods to improve image variability. Spatial augmentation with distortion [21], augmentation using color jitter (brightness, contrast, hue, and saturation), and spatial augmentation without distortion [57] were used for comparison. The augmentation procedures were performed using the default parameters in the Albumentation library. Augmentation procedure evaluations are shown in Table 7. The use of distortion in Ref. [14] reduced the performance of MobileNetV3-UNet (JAI = 74.20%). Methods such as spatial augmentation, CLAHE, blur, and sharpening with gamma did not improve model performance. However, the proposed augmentation method enhanced model performance, and JAI of 80.25% was obtained.

The modified decoder with the proposed augmentation and without augmentation were evaluated (Table 8). Augmentation significantly improves in the performance of all methods as the JAI scores increased by ~2.55%–6.75%. The lowest improvement was recorded in the JAI score of the standard U-Net decoder, while the highest was for the BCDU decoder.

4.1.4. Learning schemes by different decoder networks

The two standard learning schemes and SWA learning with constant learning rates were compared (Table 9). SWA learning uses a constant learning rate of 1e-4 (10% of the initial learning rate) after the 180th epoch (last 20 epochs). Herein, random augmentation along with MobileNetV3 encoder and different decoders were used. The SWA learning scheme with a constant learning rate improves the scores for several decoders, except BCDU. This increase indicated that the average weights for the last 20 epochs were more generalized than the best weights from the validation data.

4.1.5. Post-processing

In this experiment, we evaluated whether FITH post-processing improves model performance. Results obtained with FITH post-processing and without post-processing (none) were compared (Table 10). The performance of MobileNetV3-UNet trained with random augmentation and SWA decreased (JAI and DIC were slightly lower by 0.02% and 0.23%, respectively) in the absence of FITH post-processing. The filled segmentation marginally improved the overall scores with the exception of SEN.

4.2. Performance comparison with other state-of-the-art methods

Among the evaluated models, MobileNetV3-UNet, comprising random augmentation, SWA, and FITH post-processing, was the best-performing encoder-decoder. We compared the performance of the proposed models with state-of-the-art models using the ISIC-2017 (Table 11), ISIC-2018 (Table 12), and PH2 (Table 13) datasets. The models were trained and tested using the training, validation, and test data from each benchmark dataset. Our proposed method, MobileNetV3-UNet, outperformed the state-of-the-art models using the ISIC-2017, ISIC-2018, and PH2 dataset.

Res-UNet [22] combines the ResNet50 encoder with U-Net, which is similar to our proposed encoder-decoder concept. However, augmentation involves multiplying the training data with a balanced binary class to prevent overfitting. SU-SWA [21] is a U-Net architecture with Xception as an encoder, but it uses separable block for all parts. However, when the separable-UNet decoder that contains separable blocks (light in parameters but complex in operation) was paired with MobileNetV3, it could not outperform the standard U-Net decoder. Ensemble-A [53] (in ISIC-2017), Ensemble-S [53] (in PH2), DCL-PSI [52] use an ensemble method. Although these models are highly complex, the method utilized is quite robust. DL with auxiliary task [13] uses multi-task learning efficiently without an encoder-decoder architecture; however, compared to the previous ensemble methods, their performance is comparable in SLS. ASCU-Net [12] relies on attention modules. Quantitatively, its performance was lower than that of other models. DAGAN [11], a form of generative adversarial network (GAN)-based method, uses encoders to form segmentation. Lie et al. (2020) had utilized skip connection and dense convolution U-Net (UNet-SCDC), which was heavy, but suitable for SLS. The proposed MobileNetV3-UNet outperformed the state-of-the-art models when evaluated using the ISIC-2017 and PH2 datasets. Here, we observed DIC and JAI values of 87.74%, 80.25% and 95.18%, 91.08% for the ISIC-2017 and PH2 datasets, respectively.

The attention module, recurrent layer [54], and bidirectional convolutional LSTM based [28] modifications of the UNet-based architecture are used for medical image segmentation. Double-UNet [23] combining two architectures into one possessed considerable parameters. The proposed MobileNetV3 with BCDU, UNet-LSTM, separable-UNet, and standard U-Net decoders outperformed the state-of-the-art models when evaluations were performed using the ISIC-2018 dataset. MobileNetV3-Separable-UNet had the best ACC of 94.85%, while MobileNetV3-UNet had the best-performing DIC, JAI, and SEN values of 90.98%, 83.44%, and 90.89%, respectively. Further, our proposed method, which used efficient and lightweight models, outperformed the state-of-the-art models in all the datasets (ISIC-2017, ISIC-2018, and PH2) considered. MobileNetV3 can reduce millions of parameters and improve model performance. The proposed model outperformed the state-of-the-art models in terms of segmentation performance and computational efficiency. Moreover, the proposed MobileNetV3-UNet model obtained low ACC and SEN values, compared to the other methods. Therefore, the proposed model segmented skin cancer areas efficiently with better margins than the other methods. We obtained DIC and JAI values of 87.74%, 80.25% and 95.18%, 91.08% for ISIC-2017 and PH2 datasets, respectively. Thus, although the ACC value was low, it outperformed the existing methods in terms of DIC and JAI. In addition, the model still utilized a relatively small number of parameters and had a computation time of approximately 8 million in 0.0199 s per image.

5. Discussion

Based on our analysis, MobileNetV3 (encoder) significantly improved segmentation performance and computational efficiency. However, for capturing stage to stage features, the standard U-Net decoder was more robust than other decoders. The encoder plays an essential role in feature extraction; thus, the information obtained depends on it. Using the depthwise separable convolution block reduced the number of parameters, while the inverted residual bottleneck and NAS module improved the performance of the deep learning network. However, with a complex decoder layer, the model might lose generalizability. The LSTM-UNet and BCDU models utilized more parameters than the standard U-Net (Table 6). However, the separable-UNet had fewer parameters than the standard U-Net, but had a large latency owing to the large operation. The lightweight architecture performed better for SLS.

Using augmentation, the model learns a higher diversity within the data more robustly without overfitting. For SLS, augmentation is important due to the small data size and the presence of challenging images. Representative training and validation JAI of MobileNetV3-UNet without and with augmentation are shown in Fig. 9. The validation score may not be as high as the training score. Overfitting occurs when validation statistics decline, while training increases, as shown in Fig. 9a. As shown in Fig. 9b, augmentation can address this problem, where the validation and training scores are closer. Further, there is a wide local optimum that we effectively overcame using SWA. Augmentation affects the model training process. Without augmentation, JAI between the training and validation data do not align along the epochs, which results in model overfitting. Fig. 10 shows that augmentation significantly affected the robustness of the developed model. Models trained without augmentation detect healthy skin as lesions, or vice versa. Without augmentation, the training dataset tend to be similar, and the level of ambiguity remains high.

The use of SWA also affects the training process, which is difficult to determine towards the global optimum, so that the model with the SWA scheme can be more robust by averaging the local optimum. Post-processing only fills the hole by cleaning the messy segmentation results. With FITH post-processing, the performance of the segmentation results slightly improved.

Fig. 11 shows that the ability of MobileNetV3 is better than VGG16 in obtaining important features, as shown by class activation mapping (CAM) in one of the images. The performance of the Standard U-Net decoder is slightly improved from the other proposed decoders. Fig. 11 demonstrates an activation mapping plot of several challenging images for each model in final convolution after the sigmoid function. Several models have been represented in class activation mapping (CAM) for images of difficult lesions. These results indicate that the scores obtained in the experiment are representative of CAM. In addition, each model has its characteristics in obtaining map features. MobileNetV3-UNet can handle challenging cases of SLS, as shown in Fig. 12. Challenges that emerged in previous studies [21], such as difficult lesion images, were all well segmented by MobileNetV3-UNet. This result indicates that the segmentation capability of the MobileNetV3-UNet model is both lightweight and accurate.

6. Conclusions

In this study, lightweight encoder-decoder model based on MobileNetV3-UNet was developed for automatic SLS. Here, we comprehensively studied encoder-decoder utilization, data augmentation, SWA learning schemes, and post-processing methods to determine the best-performing model. MobileNetV3 encoder performed best for

SLS in terms of both accuracy and computational complexity. This architecture outperformed previous versions of MobileNet, ResNet50, and VGG16. Based on our analysis, the lightweight architecture works well in SLS. A standard convolution block instead of a depthwise separable convolution block can reduce the number of parameters. In addition, an inverted residual bottleneck and NAS module improved deep learning performance. This architecture could be beneficial for implementing deep learning in CAD systems, specifically for real-time image semantic segmentation. In the decoder, standard U-Net can rebuild features from stage to stage to obtain important information from the context and global information from MobileNetV3. Additionally, the proposed method outperformed the state-of-the-art models and several modified decoders (UNet-LSTM, BCDU, and Separable-UNet). The noise and challenges in SLS data were addressed using an enhanced extraction feature, random augmentation without distortion, which allowed the model to be more robust. In addition, the constant learning rate of SWA can improve model generalizability. Further, FITH post-processing improved segmentation results marginally. In the future, a comprehensive study on segmentation using MobileNetV3-UNet for other fields can be performed. The effect of decoder modifications can also be evaluated. Moreover, applications of the developed model in CAD systems should be studied to improve their performance in SLS.

Funding

Ministry of Research Technology and Higher Education of the Republic of Indonesia under scheme Fundamental Research Grant numbers: 257-22/UN7.6.1/PP/2020, and Ministry of Research and Technology National Research and Innovation Agency under scheme Fundamental Research Grant numbers: 257-22/UN7.6.1/PP/2021.

Research data

Source code for the developed models is available at https://github.com/bowoadi/lightweight_sls.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by Ministry of Research Technology and Higher Education of the Republic of Indonesia under scheme Fundamental Research Grant numbers: 257-22/UN7.6.1/PP/2020, and Ministry of Research and Technology National Research and Innovation Agency under scheme Fundamental Research Grant numbers: 257-22/UN7.6.1/PP/2021.

References

[1] Capdehourat G, Corez A, Bazzano A, Alonso R, Musé P. Toward a combined tool to assist dermatologists in melanoma detection from dermoscopic images of pigmented skin lesions. Pattern Recogn Lett 2011;32(16):2187–96. https://doi.org/10.1016/j.patrec.2011.06.015.

[2] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA*. Cancer J. Clin. 2021;71(1):7–33. https://doi.org/10.3322/caac.21654.

[3] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thurn S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542 (7639):115–8. https://doi.org/10.1038/nature21056.

[4] Binder M. Epiluminescence microscopy. A useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. Arch Dermatol 1995; 131(3):286–91. https://doi.org/10.1001/archderm.131.3.286.

[5] Vestergaard ME, Macaskill P, Holt PE, Menzies SW. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. Br J Dermatol 2008;159(3):669–76. https://doi.org/10.1111/j.1365-2133.2008.08713.x.

[6] Celebi ME, Kingravi HA, Uddin B, Iyatomi H, Aslandogan A, Stoecker WA, et al. A methodological approach to the classification of dermoscopy images. Comput Med Imag Graph 2007;31(6):362–73. https://doi.org/10.1016/j.compmedimag.2007.01.003.

[7] Celebi ME, Iyatomi H, Schaefer G, Stoecker WV. Lesion border detection in dermoscopy images. Comput Med Imag Graph 2009;33(2):148–53. https://doi.org/10.1016/j.compmedimag.2008.11.002.

[8] Silveira M, Jacinto CN, Marques JS, AndrÉ RSM, Mendoca T, Yamauchi S, et al. Comparison of segmentation methods for melanoma diagnosis in dermoscopy images. IEEE J. Sel. Top. Signal Process. 2009;3(1):35–45. https://doi.org/10.1109/JSTSP.2008.2011119.

[9] Wong A, Scharcanski J, Fieguth P. Automatic skin lesion segmentation via iterative stochastic region merging. IEEE Trans Inf Technol Biomed 2011;15(6):929–36. https://doi.org/10.1109/TITB.2011.2157829.

[10] Yuan Y, Chao M, Lo YC. Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance. IEEE Trans Med Imag 2017;36(9): 1876–86. https://doi.org/10.1109/TMI.2017.2695227.

[11] Lei B, Xia Z, Jiang F, Jiang X, Ge Z, Xu Y, et al. Skin lesion segmentation via generative adversarial networks with dual discriminators. Med Image Anal 2020; 64:101716. https://doi.org/10.1016/j.media.2020.101716.

[12] Tong X, Wei J, Sun B, Su S, Zuo Z, Wu P. ASCU-net: attention gate, spatial and channel attention U-net for skin lesion segmentation. Diagnostics 2021;11(3):501. https://doi.org/10.3390/diagnostics11030501.

[13] Liu L, Tsui YY, Mandal M. Skin lesion segmentation using deep learning with auxiliary task. J. Imaging 2021;7(4). https://doi.org/10.3390/jimaging7040067.

[14] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.

[15] Xing Y, Zhong L, Zhong X. An encoder-decoder network based FCN architecture for semantic segmentation. Wireless Commun Mobile Comput 2020:2020. https://doi.org/10.1155/2020/8861886.

[16] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2015;9351:234–41. https://doi.org/10.1007/978-3-319-24574-4_28.

[17] Siam M, Gamal M, Abdel-Razek M, Yogamani S, Jagersand M, Zhang H. A comparative study of real-time semantic segmentation for autonomous driving. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. 2018. https://doi.org/10.1109/CVPRW.2018.00101. 2018-June:700–10.

[18] Howard A, Sandler M, Chen B, Wang W, Chen LC, Tan M, et al. Searching for mobileNetV3. Proc. IEEE Int. Conf. Comput. Vis. 2019. https://doi.org/10.1109/ICCV.2019.00140. 2019-Octob: 1314–24.

[19] Drozdzal M, Vorontsov E, Chartrand G, Kadoury S, Pal C. The importance of skip connections in biomedical image segmentation. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2016;10008. https://doi.org/10.1007/978-3-319-46976-8_19. LNCS:179–87.

[20] Deng J, Fei-Fei L, ImageNet Li K. Constructing a large-scale image database. J Vis 2010;9(8):1037. https://doi.org/10.1167/9.8.1037.

[21] Tang P, Liang Q, Yan X, Xiang S, Sun W, Zhang D, et al. Efficient skin lesion segmentation using separable-Unet with stochastic weight averaging. Comput Methods Progr Biomed 2019;178:289–301. https://doi.org/10.1016/j.cmpb.2019.07.005.

[22] Zafar K, Gilani SO, Waris A, Ahmed A, Jamil M, Khan MN, et al. Skin lesion segmentation from dermoscopic images using convolutional neural network. Sensors 2020;20(6). https://doi.org/10.3390/s20061601.

[23] Jha D, Riegler MA, Johansen D, Halvorsen P, Johansen HD, DoubleU-Net “. A deep convolutional neural network for medical image segmentation. Proc. - IEEE Symp. Comput. Med. Syst. 2020. https://doi.org/10.1109/CBMS49503.2020.00111. 2020-July:558–64.

[24] Galdran A, et al. Data-driven color augmentation techniques for deep skin image analysis. ” *arXiv*; 2017.

[25] Codella NCF, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, 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). Proc. - Int. Symp. Biomed. Imaging 2018. https://doi.org/10.1109/ISBI.2018.8363547. 2018-April:168–72.

[26] Codella N, et al. Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC). ” *arXiv*; 2019.

[27] Mendonça T, Ferreira PM, Marçal ARS, et al. PH2: a public database for the analysis of dermoscopic images. In: Raton B, editor. In *dermoscopy image analysis*. USA Florida: CRC Press; 2015. p. 419–39.

[28] Azad R, Asadi-Aghbolaghi M, Fathy M, Escalera S. Bi-directional ConvLSTM U-net with densley connected convolutions. 2019.

[29] Asadi-Aghbolaghi M, Azad R, Fathy M, Escalera S. Multi-level context gating of embedded collective knowledge for medical image segmentation. ” *arXiv*; 2020.

[30] Buslaev A, Parinov A, Khvedchenya E, Iglovikov V, Kalinin AA, Albumentations “. Fast and flexible image augmentations. ” *arXiv*; 2018.

[31] Howard AG, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. ” *arXiv*; 2017.

[32] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. IEEE Comput Soc Conf Comput Vis Pattern Recogn 2018:4510–20. https://doi.org/10.1109/CVPR.2018.00474.

[33] Tan M, Chen B, Pang R, Vasudevan V, Le QV. Mnasnet: platform-aware neural architecture search for mobile. IEEE Comput Soc Conf Comput Vis Pattern Recogn 2019. https://doi.org/10.1109/CVPR.2019.00293. 2019-June:2815–23.

[34] Song H, Wang W, Zhao S, Shen J, Lam KM. Pyramid dilated deeper ConvLSTM for video salient object detection. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2018;11215. https://doi.org/10.1007/978-3-030-01252-6_44. LNCS:744–760.

[35] Chollet F. Deep learning with separable convolutions. arXiv Prepr. arXiv1610 2016:1–14. 02357.

[36] Izmailov P, Podoprikhin D, Garipov T, Vetrov D, Wilson AG. Averaging weights leads to wider optima and better generalization. UAI 2018 34th Conf. Uncertain. Artif. Intell. 2018;2:876–85. 2018.

[37] Chollet Francois, Keras ", GitHub " [Online]. Available: https://keras.io; 2015.

[38] Kingma DP, Ba JL, Adam ". A method for stochastic optimization," *3rd*. Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. 2015:1–15.

[39] Humayun J, Malik AS, Kamel N. "Multilevel thresholding for segmentation of pigmented skin lesions," *2011*. IEEE Int. Conf. Imaging Syst. Tech. IST 2011 - Proc. 2011:310–4. https://doi.org/10.1109/IST.2011.5962214.

[40] Garnavi R, Aldeen M, Celebi ME, Varigos, Finch S. Border detection in dermoscopy images using hybrid thresholding on optimized color channels. Comput Med Imag Graph 2011;35(2):105–15. https://doi.org/10.1016/j.compmedimag.2010.08.001.

[41] Emre Celebi M, Kingravi HA, Iyatomi H, Alp Aslandogan Y, Stoecker WV, Moss RH, et al. Border detection in dermoscopy images using statistical region merging. Skin Res Technol 2008;14(3):347–53.

[42] Tang J. A multi-direction GVF snake for the segmentation of skin cancer images. Pattern Recogn 2009;42(6):1172–9. https://doi.org/10.1016/j.patcog.2008.09.007.

[43] Abbas Q, Fondón I, Sarmiento A, Emre Celebi M. An improved segmentation method for non-melanoma skin lesions using active contour model. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2014;8815:193–200. https://doi.org/10.1007/978-3-319-11755-3_22.

[44] Ganster H, Pinz A, Röhrer R, Wildling E, Binder M, Kittler H. Automated melanoma recognition. IEEE Trans Med Imag 2001;20(3):233–9. https://doi.org/10.1109/42.918473.

[45] Ali AR, Couceiro MS, Hassenian AE. Melanoma detection using fuzzy C-means clustering coupled with mathematical morphology," *2014 14th*. Int. Conf. Hybrid Intell. Syst. HIS 2014:73–8. https://doi.org/10.1109/HIS.2014.7086175.

[46] Xie F, Bovik AC. Automatic segmentation of dermoscopy images using self-generating neural networks seeded by genetic algorithm. Pattern Recogn 2013;46(3):1012–9. https://doi.org/10.1016/j.patcog.2012.08.012.

[47] He Y, Xie F. Automatic skin lesion segmentation based on texture analysis and supervised learning. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2013;7725. https://doi.org/10.1007/978-3-642-37444-9_26. LNCS:330–41.

[48] Bi L, Kim J, Ahn E, Kumar Feng D, Fulham M. Step-wise integration of deep class-specific learning for dermoscopic image segmentation. Pattern Recogn 2019;85:78–89. https://doi.org/10.1016/j.patcog.2018.08.001.

[49] Yu L, Chen H, Dou Q, Qin J, Heng PA. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imag 2017;36(4):994–1004. https://doi.org/10.1109/TMI.2016.2642839.

[50] Bi L, Kim J, Ahn E, Feng D, ". Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. " *arXiv*; 2017.

[51] Bi L, Kim J, Ahn Kumar A, Fulham M, Feng D. Dermoscopic image segmentation via multistage fully convolutional networks. IEEE Trans Biomed Eng 2017;64(9):2065–74. https://doi.org/10.1109/TBME.2017.2712771.

[52] Bi L, Kim J, Ahn E, Kumar A, Feng D, Fulham M. Step-wise integration of deep class-specific learning for dermoscopic image segmentation. Pattern Recogn 2019;85:78–89. https://doi.org/10.1016/j.patcog.2018.08.001.

[53] Goyal M, Oakley A, Bansal P, Dancey D, Yap MH. Skin lesion segmentation in dermoscopic images with ensemble deep learning methods. IEEE Access 2019;8:4171–81. https://doi.org/10.1109/ACCESS.2019.2960504.

[54] Alom MZ, Yakopcic C, Hasan M, Taha TM, Asari VK. Recurrent residual U-Net for medical image segmentation. J Med Imaging 2019;6(1):1. https://doi.org/10.1117/1.jmi.6.1.014006.

[55] Hartanto CA, Wibowo A. "Development of mobile skin cancer detection using faster R-CNN and MobileNet v2 model," *7th*. Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2020 - Proc. 2020:58–63. https://doi.org/10.1109/ICITACEE50144.2020.9239197.

[56] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. IEEE Comput Soc Conf Comput Vis Pattern Recogn 2016. https://doi.org/10.1109/CVPR.2016.90. 2016-Decem:770–8.

[57] Wei Z, Song H, Chen L, Li Q, Han G. Attention-based denseunet network with adversarial training for skin lesion segmentation. IEEE Access 2019;7:136616–29. https://doi.org/10.1109/ACCESS.2019.2940794.

C2

ORIGINALITY REPORT

4%	4%	4%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	personalpages.manchester.ac.uk Internet Source	1%
2	ahmedhosny.com Internet Source	1%
3	Lokesh Singh, Rekh Ram Janghel, Satya Prakash Sahu. "A hybrid feature fusion strategy for early fusion and majority voting for late fusion towards melanocytic skin lesion detection", International Journal of Imaging Systems and Technology, 2021 Publication	1%
4	www.cfp.ca Internet Source	1%
5	journals.sums.ac.ir Internet Source	1%
6	eprints.soton.ac.uk Internet Source	1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	Off		