Summer 2024

Deep Learning Techniques for Image Segmentation in Dermoscopic Skin Cancer Images

Norsang Lama
Missouri University of Science and Technology

Department: Electrical and Computer Engineering

DEEP LEARNING TECHNIQUES FOR IMAGE SEGMENTATION IN DERMOSCOPIC SKIN CANCER IMAGES

by

NORSANG LAMA

A DISSERTATION

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

COMPUTER ENGINEERING

2023

Approved by:

R. Joe Stanley, Advisor
William V. Stoecker
Kurt Kosbar
Maciej J. Zawodniok
V. A. Samaranayake

PUBLICATION DISSERTATION OPTION

This dissertation consists of the following three articles, formatted in the style used by the Missouri University of Science and Technology:

Paper I, found on pages 7–33, has been published in the Journal of Digital Imaging.

Paper II, found on pages 34–63, has been published in the Journal of Digital Imaging.

Paper III, found on pages 64–89, has been submitted to the IEEE Transactions on Medical Imaging.

ABSTRACT

Melanoma is recognized as the most lethal type of skin cancer, responsible for a significant proportion of skin cancer-related deaths. However, early detection of melanoma is essential for successful treatment outcomes. Computer-aided skin cancer diagnosis tools can save lives by enabling earlier detection of skin cancer. Image segmentation is a crucial step in computer-aided diagnosis as it allows the detection of critical features or regions in an image. Thus, an accurate image segmentation method is necessary to create a more precise computer-aided diagnostic tool for skin cancer diagnosis. This dissertation includes investigating and developing deep learning techniques to improve image segmentation in dermoscoopic skin lesion images.

First, a novel deep neural architecture is proposed for hair and ruler mark detection in skin lesion images. Second, a new deep learning approach is developed to segment lesion borders. Third, a novel data augmentation technique is developed to generate synthetic multi-lesion images to train a robust deep neural network for multi-lesion segmentation. The experimental results from this research achieved state-of-the-art performance on hair and ruler mark segmentation and lesion segmentation in skin lesion images.

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Dr. R. Joe Stanley, for his guidance, support, and encouragement throughout my PhD journey. Under his guidance, I was able to complete this dissertation. I also sincerely appreciate Dr. William V. Stoecker's advice, feedback, and support. His expertise in dermatology has been invaluable in progressing my research. Furthermore, I want to thank other members of my committee, Dr. Kurt Kosbar, Dr. Maciej J. Zawodniok, and Dr. V. A. Samaranake, for their insights, guidance, and feedback throughout my academic journey. Additionally, I express my gratitude to my former advisor Dr. Scott E. Umbaugh, for his support and encouragement in pursuing research and PhD education.

I want to thank my wife Binita and daughter Noryang for their unwavering love, encouragement, and moral support throughout my PhD. Thanks to my lab colleagues, Jason Hagerty, Anand Nambisan, Akanksha Maurya, Sudhir Sornupudi, and Haidar Almubarak, for their help and suggestions during the research discussion. My friends, Madhav Dhital, Milan Koirala, Purna B. Magar, and Rohini Dahal, have provided me much-needed support and encouragement.

Lastly, I am indebted to my parents, Mr. Birman Lama, and Mrs. Lalmoti Lama, for their exceptional love, support, and encouragement in fulfilling my dreams. I also want to thank my sister Usha and all my family members for their support throughout my academic journey.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

PAPER III

LIST OF TABLES

NOMENCLATURE

Symbol	Description
CAD	Computer-aided diagnosis
DL	Deep learning
SERes	Squeeze-and-excitation residual
GT	Ground truth
LAMA	Lesion-aware mixup segmentation
ISIC	International Skin Imaging Collaboration
RGB	Red, green and blue
MBConv	Mobile inverted bottleneck convolution
FC	Fully connected
W	Width
H	Height
C	Number of feature map or channel
ReLU	Rectified linear activation unit
r	Feature reduction factor
SE	Squeeze-and-excitation
Jac	Jaccard index
Dsc	Dice similarity coefficient
Prec	Precision
Rec	Recall
Acc	Accuracy

Inacc	Inaccuracy
TP	True positive
TN	True negative
FP	False positive
FN	False negative
Log	Logarithmic
Cosh	Hyperbolic cosine
Dec	Decoder
MuLe	Multi-lesion
n	Total number of patch levels
l	Patch level
h_l	Patch height
w_l	Patch width
P_l	Patch count

1. INTRODUCTION

Skin cancer is a prevalent form of cancer worldwide, resulting primarily from exposure to ultraviolet radiation (UV) from the sun or other sources such as tanning beds. The three primary types of skin cancer are melanoma, squamous cell carcinoma and basal cell carcinoma. Melanoma is responsible for most fatalities associated with skin cancer, despite representing only 1% of all skin cancer cases [1]. In the United States alone, current estimates are that 97,610 new cases of invasive melanoma and 89,070 cases of in-situ melanoma are expected to be diagnosed in 2023 [2]. Moreover, it is projected that 7,990 deaths will be attributed to melanoma – 5420 male and 2570 female. Melanoma detected at an early stage can be treated successfully; however, the five-year survival rate drops drastically from > 99% to 32 % if the melanoma has spread to distant lymph nodes or other organs [1].

Skin cancers are diagnosed using various techniques such as clinical examination, photography, dermoscopy and biopsy. Dermoscopy is a non-invasive diagnostic tool that improves non-contact skin lesion examination using a handheld device with a magnifying lens and a cross-polarized light source. Alternately, contact dermoscopy employs gel or similar skin interface to reduce reflections from the skin surface. Both techniques increase visualization of subsurface details for the clinician. Dermoscopy has shown better diagnostic accuracy of early-stage melanoma over a dermatologist's naked-eye visual inspections of lesions [3]–[5].

There has been a considerable effort to develop a computer-aided diagnosis (CAD) for skin cancer. CAD uses machine learning algorithms to analyze medical images and

provide diagnostic support to clinicians. A computer-aided diagnostic tool might not only give higher accuracy but also has the potential to reduce the time and cost associated with a skin cancer diagnosis. In recent years, deep learning (DL) techniques have dominated almost all image recognition and classification tasks [6]–[10]. DL techniques combined with dermoscopy have demonstrated higher diagnostic accuracy in the skin cancer domain than experienced dermatologists [11]–[13].

While deep learning techniques have become dominant in medical image analysis, it is essential to recognize that clinically relevant handcrafted features are also crucial. Such features can aid in building a better CAD tool and facilitate an explainable decision process. Furthermore, recent studies showed that the fusion of DL and handcrafted features boosts the diagnostic accuracy of skin lesion classifiers [12], [14]–[17]. The computation of handcrafted features usually involves image segmentation to detect crucial features or regions in the image. Thus, an accurate segmentation method is necessary to develop a precise computer-aided diagnostic tool for skin cancer diagnosis. This dissertation aims to develop better image segmentation methods using deep learning techniques for skin lesion image analysis.

1.1. PROBLEM STATEMENT

Image segmentation is a crucial step for a computer-aided diagnosis of skin cancer. Image segmentation methods using both traditional image processing approaches [18]–[23] and deep learning techniques [24]–[31] have been applied in dermoscopic skin lesion images to detect various clinical features, including hair, blood vessel, globules and lesion border. Although traditional methods showed promising results when applied to small

datasets, they typically exhibit unsatisfactory performance under challenging conditions such as lesions with varying color, low contrast between lesion and background, and images that contain artifacts such as hair, ruler marks, ink marks and stickers. Conversely, deep learning methods have overcome these challenges to some extent and improved the segmentation performance in skin lesion images. Caution is needed here because some of the recognized features, notably ink marks, can lead to DL decisions which apply only to a given image set. The ink marks can signal melanoma, but they are not an intrinsic image-set-independent diagnostic feature. Rather, such artifacts comprise an information leak. Despite the early success of DL techniques, better image segmentation methods are warranted to develop a precise computer-aided diagnostic tool for skin cancer.

This dissertation uses deep learning techniques to investigate three different image segmentation tasks in dermoscopic skin lesion images. First, deep learning-based hair and ruler segmentation is studied, as hair and ruler marks hinder accurate segmentation and detection of critical network features in skin lesion images. Hair detection is particularly challenging in dermoscopic skin lesion images as the hairs can be thin, overlapping, faded, or of similar color as skin or overlaid on a textured lesion. Second, a deep learning method is proposed to find lesion borders in skin lesion images. The existing traditional image processing and deep learning techniques face difficulties in accurately segmenting lesions due to factors such as high variation in lesion color, low contrast between the lesion and the background, and the presence of artifacts. Therefore, there is a need for novel deep learning architectures with advanced representational capabilities to address the existing techniques' current limitations. Third, a new data augmentation technique is proposed to train a robust DL model for single-lesion and multiple-lesion segmentation. Dermoscopic

skin lesion images typically contain a single lesion per image, although there are many examples where more than one lesion is present. Multiple-lesion segmentation has yet to be explored in dermoscopic image analysis. Incorporating information about multiple lesions may prove advantageous for machine learning systems utilized in computer-aided skin cancer diagnosis in various domains.

1.2. SUMMARY OF CONTRIBUTIONS

This dissertation is comprised of three journal publications as listed in the publications list. The unique contributions are summarized as follows:

1.2.1. ChimeraNet: U-Net for Hair Detection in Dermoscopic Skin Lesion Images. Detection of hair and ruler marks can serve multiple purposes. Not only can it help to eliminate noise when identifying crucial network features, but it can also yield important clinical information to enhance the accuracy of diagnosis. For instance, hair can reveal valuable details about the patient's age, gender, and body part being examined. This study presents a novel DL method, ChimeraNet, for hair and ruler mark detection in skin lesion images. ChimeraNet is a U-Net [32] based architecture which consists of two main parts – an encoder network and a decoder network. The proposed novel architecture uses a pretrained EfficientNet [9] as the encoder network; the decoder network is constructed using squeeze-and-excitation residual (SERes) structures. The squeeze-and-excitation structure was chosen as it has a better generalization capability than the plain convolution structure as it emphasizes the critical channels of feature maps while suppressing the weaker ones [33]. In addition, the skip-connections in U-Net architecture help to recover the spatial information lost during downsampling in the encoding process. Furthermore,

ChimeraNet uses an extra skip connection directly from the input to the decoder, which provides additional low-level spatial information while reconstructing a segmentation map for fine structures like hair and ruler marks. The proposed ChimeraNet yields state-of-the-art accuracy compared to previously reported classical techniques and DL methods.

1.2.2. Skin Lesion Segmentation in Dermoscopic Images with Noisy Data. Skin lesion segmentation enables the extraction of precise and detailed information about the lesion's shape, size, and texture, which helps distinguish benign lesions from malignant ones. This can aid clinicians and dermatologists in making more informed decisions regarding the treatment and management of skin cancer. In this study, we proposed a novel deep learning method for lesion segmentation in dermoscopic images. The proposed method uses a ChimeraNet architecture, described in section 1.2.1, with minor modifications to accommodate lesion segmentation task. The extra skip-connection and fifth decoder block of the original ChimeraNet were specifically designed to give extra low-level information to detect fine structures like hair and ruler marks. These extra connections were removed. Despite the growing popularity of deep learning methods in image recognition, the quality of training data remains a critical factor in achieving high performance. Thus, this study also assesses the data quality in a public benchmark dataset. ISIC 2017 [34] skin lesion segmentation dataset is the largest and most widely used dataset in skin lesion segmentation studies. Our investigation shows that the dataset contains many inaccurate or noisy ground truth (GT) segmentation masks, as GT labels were created using manual and semi-automated processes. In this study, a dermatologist (W.V.S.) visually inspected and sorted all GT masks into good, mildly noisy, and noisy categories. The experimental results show noisy labels in the training set do not lower segmentation

performance. Moreover, the proposed DL segmentation method achieved a Jaccard score of 0.807 on an official ISIC2017 test set, which surpasses the previously reported methods.

1.2.3. LAMA: Lesion-Aware Mixup Augmentation for Skin Lesion Segmentation. Multiple-skin lesion segmentation in a dermoscopic image can offer significant assistance to computer-aided diagnosis by providing relevant information on the location and characteristics of each lesion. In this research, we develop a novel lesion-aware mixup augmentation (LAMA) technique synthesizing multi-lesion samples using single-lesion images. LAMA randomly selects one or more lesions and pastes in a non-lesion area of an image. To find the non-lesion area in the image, it generates patches at multiple patch levels and creates a pool of candidate non-lesion patches by checking whether the patch contains a part of the lesion. The lesion patches are generated by cropping the lesion from each image using a bounding box enclosing the lesion. The size of the lesion varies greatly in dermoscopic lesion images. To avoid excessive resizing during mixing, LAMA groups the lesion patches into n groups based on their sizes, with each group corresponding to a specific level of non-lesion patches. The lesions are distributed in each group based on the histogram distribution of images having non-lesion patches at each patch level. Finally, LAMA randomly selects one or more lesion patches and pastes them on the non-lesion patches in the image. The DL model trained with the LAMA method improves lesion segmentation over the baseline model in single- and multi-lesion images. The proposed method finds the multiple lesions in real-life examples despite training using only synthetic multi-lesion images based on single-lesion examples.

PAPER

I. CHIMERANET: U-NET FOR HAIR DETECTION IN DERMOSCOPIC SKIN LESION IMAGES

Norsang Lama[1], Reda Kasmi[2], Jason R. Hagerty[3], R. Joe Stanley[1], Reagan Young[1], Jessica Miinch[1], Januka Nepal[3], Anand Nambisan[1], William V. Stoecker[3]

[1]Department of Electrical and Computer Engineering,
Missouri University of Science and Technology, Rolla, MO, USA
[2]Faculty of Technology, Laboratoire de Technologie Industrielle et de I'Information,
University of Bejaia, Bejaia, Algeria.
[3]S&A Technologies, Rolla, MO, 65401 USA

ABSTRACT

Hair and ruler mark structures in dermoscopic images are an obstacle preventing accurate image segmentation and detection of critical network features. Recognition and removal of hairs from images can be challenging, especially for hairs that are thin, overlapping, faded, or of similar color as skin or overlaid on a textured lesion. This paper proposes a novel deep learning (DL) technique to detect hair and ruler marks in skin lesion images. Our proposed ChimeraNet is an encoder-decoder architecture that employs pretrained EfficientNet in the encoder and squeeze-and-excitation residual (SERes) structures in the decoder. We applied this approach at multiple image sizes and evaluated it using the publicly available HAM10000 (ISIC2018 Task 3) skin lesion dataset. Our test results show that the largest image size (448x448) gave the highest accuracy of 98.23 and Jaccard index of 0.65 on the HAM10000 (ISIC 2018 Task 3) skin lesion dataset, exhibiting better performance than for two well-known deep learning approaches, U-Net and ResUNet-a. We found the Dice loss function to give the best results for all measures.

Further evaluated on 25 additional test images, the technique yields state-of-the-art accuracy compared to 8 previously reported classical techniques. We conclude that the proposed ChimeraNet architecture may enable improved detection of fine image structures. Further application of DL techniques to detect dermoscopy structures is warranted.

Keywords—Hair removal, melanoma, dermoscopy, deep learning, image segmentation, transfer learning

1. INTRODUCTION

An estimated 106,110 new cases of invasive melanoma and 101,280 new cases of in-situ melanoma will be diagnosed in 2021 in the USA [1], in addition to 2 million epitheliomas [2]. Dermoscopy is an imaging modality that renders these cancers visible when they are fully curable. However, many cases of melanoma are missed by domain experts [3], [4].

Dermoscopy is a crucial tool in the early detection of melanoma, increasing the diagnostic accuracy over clinical visual inspection in the hands of experienced physicians [5]–[7]. Yet dermatologists viewing dermoscopic images in recent studies have shown lower diagnostic accuracy than machine vision techniques [3], [4], [8], [9].

Pathan et al. published a recent review detailing both handcrafted and deep learning (DL) techniques for computer-aided diagnosis of skin lesions [10]. Recent studies have shown successful results for skin cancer diagnosis by fusing ensembles, in some cases handcrafted and DL techniques [11]–[15]. However, hair and ruler artifacts can interfere with handcrafted feature detection by mimicking pigment network and interfering with

accurate border detection [16], [17]; thus, detection of these structures is needed to maximize handcrafted feature detection accuracy.

Ian Lee et al. reviewed hair detection in dermoscopic images[18]. The methods generally employ one of three types: mathematical morphology, edge detection, and matched filters. To detect hair, the same group [18] used the top-hat transform and modified Gaussian filter to enhance hair. Xie et al. [19] proposed an algorithm focused on dark hair, using a top-hat operator and an automatic threshold. Abbas et al. [20] implemented detection of both light and dark hairs using the first derivative of Gaussian followed by morphological techniques to remove non-hair noise. Nguyen et al. [21] also detected light and dark hair using a universal matched filtering kernel, with the binary mask generated by local entropy thresholding.

Lee et al. [22] proposed Dullrazor, with dark hair identified by grayscale morphology. Subsequent operations verify long and narrow structures. Fiorese et al. [23] proposed VirtualShave, which uses top-hat filtering followed by morphological postprocessing. Koehoorn et al. [24] used a threshold-set model and a gap-detection algorithm and further postprocessing using skeletonizing. Toossi et al. [25] used Canny edge detection followed by refinement using morphologic operators. Abuzaghleh et al. [26] segment hair by a set of directional Gaussian filters. Huang et al. [27] sought to detect hairs often missed, thin hairs and hairs in shaded backgrounds, using Gaussian matched filtering. Zhou et al. [28] used Steger's algorithm and a least-square method. Inpainting algorithms employed in these studies included interpolation, in either one or multiple directions, and a fast-marching technique.

Deep learning methods have also been applied to detect hair in skin lesion images. Attia et al. [29] used a hybrid model of convolution and recurrent layers for hair segmentation. However, due to the lack of a hair mask dataset, they trained using the weakly labeled data and tested on the simulated hair. Li et al. [30] created a new hair-mask dataset to train and test U-Net [31] based hair segmentation model. To create the hair-mask dataset, first, they applied the top-hat segmentation techniques [32] and then manually removed the over-segmented regions. Moreover, they excluded the under-segmented images from the dataset. The use of weakly labeled data in the first method and selecting only over-segmented images in the second method limit the quality of data, thereby weakening the robustness of the DL network.

The approach of this paper is as follows. First, we created a well-labeled hair mask dataset by manually annotating hair and ruler marks on the public HAM10000, also ISIC 2018 Task3 lesion classification dataset [33], [34]. Many of the reported hair removal algorithms were evaluated on small sets of images. Many of these methods find only dark hair, ignoring light hair and ruler marks. Second, we proposed a novel deep learning based hair detection method called ChimeraNet that detects light hair and ruler marks as well as dark hair in dermoscopic skin lesion images. Further, we compared the performance of our proposed method against two well-known DL approaches [31], [35] and eight conventional image processing approaches [18]–[20], [22]–[24], [27], [36].

The remainder of the paper is organized as follows. Section 2 explains the image datasets and proposed method. Section 3 presents the segmentation experiments and comparison. Section 4 provides a discussion. Section 5 gives the conclusion and possible future work.

2. MATERIALS AND METHODS

2.1. IMAGE DATASETS

We used two datasets in this study. The first is the HAM10000 [33] dataset of Tschandl et al., a publicly available skin lesion dermoscopy dataset containing over 10,000 skin images for 7 diagnostic categories. All images are 8-bit RGB images of size 450×600, as shown in Figure 1. These images comprised a training set for the ISIC2018 Task 3 lesion classification challenge [34]. Since the dataset lacks ground truth hair segmentation masks, we manually drew hair masks for 1333 dermoscopy images in the dataset. The manually drawn hair masks include dark hair, white hair, and ruler marks, Figure 2. As the width of hair and ruler marks differs within an image or from one image to other, the manual hair masks were drawn with varying widths.

Figure 1. Skin lesion dermoscopy images from HAM10000 (ISIC2018 Task 3) dataset showing dark hair, light hair, and ruler marks.

The masks were evaluated and verified by a dermatologist. Some of the hair segments are very thin, in some cases 1-pixel wide, so we dilated the binary masks with a 3×3 cross-shaped structuring element.

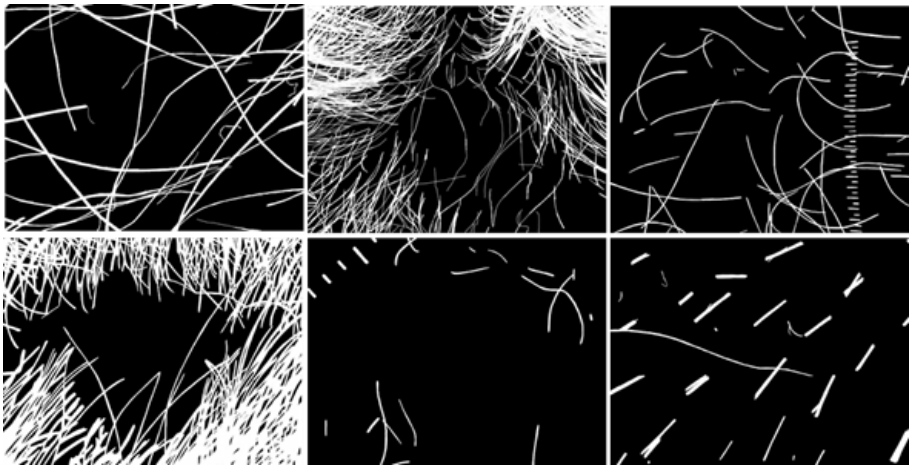


Figure 2. Manually drawn hair masks corresponding to skin lesion images shown in Figure 1. The hair masks include dark hair, white hair, and ruler marks.

We randomly divided 1333 skin lesion images from the HAM10000 dataset (ISIC2018 Task 3 Lesion Classification dataset) into image subsets of 852 for training, 214 for validation, and 267 for testing. Both training and validation images combined with manually drawn masks were used to train the deep convolutional neural networks, and the holdout test images were used to evaluate the performance of the proposed model.

The second dataset was 25 dermoscopic skin lesion images with hair and calculated hair masks from [18]. The images are RGB images of size 768×1024. We resized all images into same size as of first dataset i.e., 450×600 using bilinear interpolation method. This set was small compared to first set, so only used as a test set in this study.

2.2. DATA AUGMENTATION

Data augmentation helps convolution neural networks to generalize better and reduce the overfitting problem by adding more training samples. It performs different image transform methods on original images to generate more examples and used them for training. In this study, we selected following image transforms for data augmentation:

- Height or width shift with a range of (-0.15, +0.15)
- Horizontal or vertical flip
- Rotation with range between +90° to -90°
- Zoom with a range (-0.15, +15)
- Brightness with a range of (0.9, 1)

We performed online data augmentation to increase the number of training images by 5 times. Then, the augmented images' data range were rescaled between 0 and 1. Further, the images were normalized before feeding them to the deep network.

2.3. PROPOSED NETWORK ARCHITECTURE

U-Net [31] is a popular convolutional neural network architecture designed for biomedical image segmentation. It consists of encoder (left) and decoder (right) paths. The encoder path is a typical convolutional neural network (CNN) that extracts the high-level features from the input image. The features maps are downsampled many times at different levels, thus reducing the spatial dimension of feature maps. The decoder path, which is symmetric to the encoder path, expands the low-resolution feature map to generate a segmentation map with spatial dimensions equal to those of the input. First, the decoder concatenates the feature map and high-resolution features from the encoder. Then, it

upsamples the concatenated feature maps applying transpose convolution operations. The high-resolution features via skip-connections from the encoder are helpful in recovering the fine-grained details in the image to generate the segmentation map.

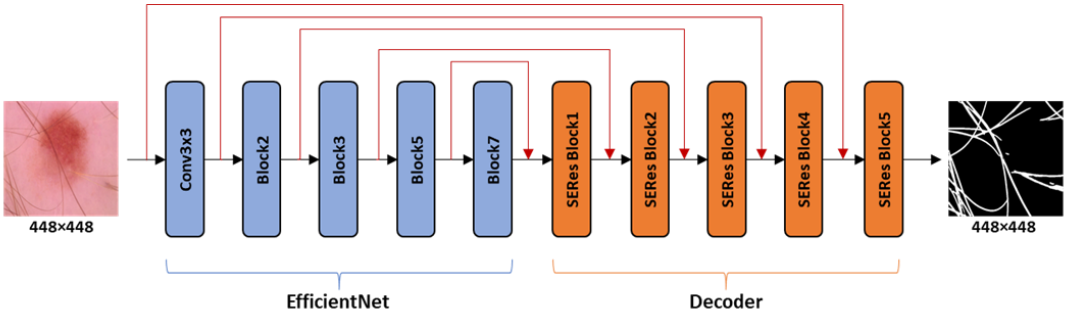


Figure 3. Proposed ChimeraNet architecture for hair segmentation. An encoder-decoder architecture with pretrained EfficientNet model as the encoder network and the decoder network comprised of five squeeze-and-excitation residual blocks. Five skip-connections (Red color) from the encoder to the decoder.

The overall pipeline of our proposed ChimeraNet architecture is shown in Figure 3. First, we used a pre-trained EfficientNet model as the encoder network in our model. Tan et al. [37] developed a family of CNN models, called EfficientNets, which achieved state-of-art top-1 accuracy in the ImageNet [38] image classification challenge. These models are composed of mobile inverted bottleneck convolution (MBConv) blocks. EfficientNet has seven blocks, from Block1 to Block7, which are composed of multiple MBConv blocks. Based on the number of MBConv blocks in Block1 to Block7, they have 8 network variants from B0-B7. EfficientNetB0 is the baseline architecture, and other variants are scaled up versions by employing a compound scaling method that uniformly scales network depth, width, and resolution with a fixed set of scaling factors. In our model, we used the EfficientNet-B5 variant of EfficientNet models trained on ImageNet. As the

models were originally developed for image classification, we could not directly employ the EfficientNet-B5 [37] as the encoder network in our model. Thus, we removed some top layers of EfficientNet-B5 that are specifically designed for classification output: a final 3x3 convolution layer, a global average pooling layer, and a fully connected (FC) layer. And we used the remaining layers, all layers from input to Block7, as the encoder part without any modification on them. Also, the U-Net architecture uses the skip-connections from the encoder network to the decoder network. Typically, the U-Net architecture has four skip-connections, and we selected the outputs of Conv3×3, Block2, Block3 and Block5 of EfficientNet-B5 as the sources of skip-connections as shown in the Figure 3. Further, we added an extra skip-connection from the input layer to the final decoder block. This adds more low-level spatial information to the decoder for precise localization of features and improves the segmentation of fine structures like hair and ruler marks which are very thin, being only a few pixels wide. The output shapes of the blocks corresponding to the skip connections and the final output of the encoder are shown in Table 1.

Table 1. Different blocks of the encoder showing the size and number of feature map.

Block Name	**Size (*W* × *H*)**	**#Feature Map (*C*)**
Input layer	448 × 448	3
Conv3×3	224 × 224	48
Block2	112 × 112	40
Block3	56 × 56	64
Block5	28 × 28	176
Block7	14 × 14	512

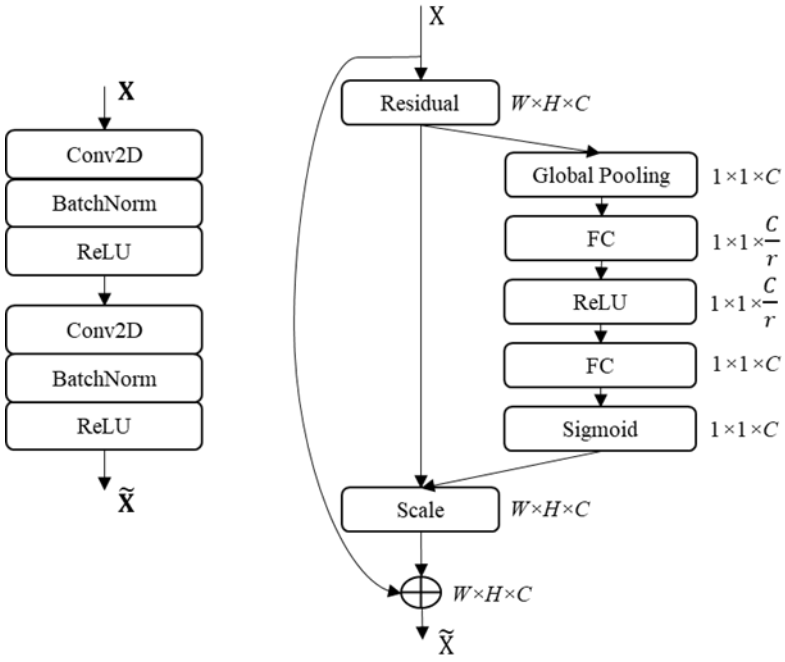


Figure 4. Convolutional block structures of the decoder. Double convolution block (left) and Squeeze-and-excitation residual block (right).

Second, we constructed a decoder network using a squeeze-and-excitation residual (SERes) structures developed by [39]. The squeeze-and-excitation (SE) block emphasizes the informative features and suppresses the weaker ones by modeling the interdependencies between channels of convolutional features. We used five SERes blocks named SERes Block1 to SERes Block5 as shown in Figure 3. Each SERes block gets two feature maps as inputs – an output feature map from previous block and a low-level feature map via skip-connection from the encoder. For example, the first SERes block (SERes Block1) of the decoder gets 14×14×512 feature input from Block7 and 28×28×176 low-level feature input from Block5 via skip-connection; here, three dimensions of the feature map represent width (W), height (H) and number of feature map or channel (C). However, the dimensions of both inputs are not same. To combine both inputs, first, the 14×14×512 feature map from

previous layer is upsampled using a transposed convolution, also called deconvolution. The transposed convolution performs both 2×2 upsampling and 3×3 convolution operations. We selected the number of filters for transposed convolution as the half of the number of input channels i.e., 256 (=512/2), and thus generates 28×28×256 feature map. Then, the two inputs are concatenated along the channel axis to form 28×28×432 feature map before feeding to SERes block. The SERes block combines an SE block with a residual structure [40], as shown in Figure 4. The residual unit in the SERes block is a double convolution block, which applies two sets of 3×3 convolution, batch normalization [41], and rectified linear unit (ReLU) operations. Again, we selected the number of filters for two convolution layers in residual unit as half of the input channels i.e., 216 (= 432/2). The residual unit outputs 28×28×216 feature map and then the squeeze-and-excitation operation is performed to scale the features along the channel axis. To find the weights for each channel of the feature map, SE first applies global average pooling to reduce the feature map to 1×1×216 and then applies non-linear operations like FC, ReLU, FC and sigmoid. The number of neurons in two FC layers are C/r and C respectively, where r is a feature reduction factor and empirically selected as r = 8. The SE generates 1×1×216 weight vector with each value in the range of 0 to 1. Then residual feature is multiplied with weight vector to scale the features and generates 28×28×216 scaled feature map. Further, the SERes block combines this scaled residual feature map with original input feature map. But the channels in original input feature map (28×28×256) and residual feature output (28×28×216) are not same so 1×1 convolution operation with 216 filters followed by batch normalization operation are performed on original input feature map. Then, SERes Block1 adds two

feature maps together and applies ReLU operation to generate the final 28×28×216 feature output.

Table 2. SERes blocks in the decoder and their output shapes.

Block Name	**Size (*W* × *H*)**	**Feature Map (*C*)**
SERes Block1	28 × 28	216
SERes Block2	56 × 56	86
SERes Block3	112 × 112	41
SERes Block4	224 × 224	34
SERes Block5	448 × 448	10

During inference, we give five different augmented versions of an input image to the trained deep network: an original image, horizontally flipped image, vertically flipped image, 90° clockwise rotated image, and 90° counterclockwise rotated image. The deep network generates segmentation output for each image and the final segmentation mask is generated by aggregating these five outputs. The final aggregated mask is the unweighted average of the five predicted masks. Further, the mask is binarized using the threshold of 0.5 to generate a binary segmentation mask.

2.4. TRAINING DETAILS

All models were built using Keras with a Tensorflow backend in Python 3 and trained using a single 32GB Nvidia V100 graphics card. The networks were trained for

200 epochs with a batch size of 16 and a constant learning rate of 0.0001 using the Adam optimization algorithm [42]. We set up an early stopping criterion with a patience of 30 epochs to stop the model from overtraining. All images are of equal size with dimensions 450×600 and were resized using bilinear interpolation into various sizes, with further details in section 3.1. We conducted experiments with five different loss functions, see section 3.2.

3. EXPERIMENTAL RESULTS AND DISCUSSION

We evaluated the performance of the proposed algorithm by comparing the predicted hair masks with the manually drawn ground truth hair masks on two test sets - 267 images from [33], [34] and 25 images from [18]. The evaluation metrics used are Jaccard index (Jac), Dice similarity coefficient (Dsc), precision (Prec), recall (Rec), accuracy (Acc) and inaccuracy (Inacc), which are defined by the Equations (1)-(6).

$$Jac = \frac{TP}{TP + FP + FN} \tag{1}$$

$$Dsc = \frac{2TP}{2TP + FP + FN} \tag{2}$$

$$Prec = \frac{TP}{TP + FP} \tag{3}$$

$$Rec = \frac{TP}{TP + FN} \tag{4}$$

$$Acc = \frac{TP + TN}{TP + FP + FP + FN} \tag{5}$$

$$Inacc = \frac{FP}{FP + TN} + \frac{FN}{FN + TP} \tag{6}$$

where TP is the number of true positives denoting the pixels correctly identified as the hair, TN is the number of true negatives denoting the pixels correctly identified as the background, FP is the number of false positives pixels denoting the background pixels incorrectly identified as the hair, and FN is the number of false negatives denoting the hair pixels incorrectly identified as the background.

3.1. HAIR SEGMENTATION PERFORMANCE FOR DIFFERENT IMAGE SIZES

In this section, we compared the hair segmentation performance of the proposed method with different input sizes using 267 test images. We selected 3 different sizes to modify the original images in the HAM10000 dataset. When the image size was doubled from 224×224 to 448×448, the Jaccard score increased by 32.65% from 0.49 to 0.65, precision increased by 14.49% from 0.69 to 0.79, and recall increased by 19.7% from 0.66 to 0.79. Table 3 shows that larger image sizes have better segmentation performance. However, a larger image size requires more memory and more computational time.

Table 3. Segmentation Test Results with Different Input Sizes.

Image Size	Acc	Inacc	Dsc	Jac	Prec	Rec
224×224	97.30	17.85	0.63	0.49	0.69	0.66
320×320	97.76	12.67	0.72	0.58	0.72	0.76
448×448	98.23	10.97	0.77	0.65	0.79	0.79

3.2. HAIR SEGMENTATION PERFORMANCE FOR DIFFERENT LOSS FUNCTIONS

For highly imbalanced data sets, such as hair images where most pixels are true negative pixels, Jadon found that the Focal-Tversky and Log-Cosh Dice loss functions yield improved results [43].

Table 4. Segmentation Test Results with Different Loss Functions.

Loss Method	Acc	Inacc	Dsc	Jac	Prec	Rec
Jaccard	98.081	10.350	0.762	0.634	0.751	0.808
Dice	98.228	10.963	0.775	0.649	0.790	0.792
Log-Cosh Dice	98.074	10.349	0.765	0.637	0.761	0.808
Tversky	98.065	9.480	0.764	0.635	0.737	0.826
Focal-Tversky	98.091	9.899	0.766	0.637	0.751	0.817

We compared the performance of five widely used loss methods for image segmentation - Dice [44], Jaccard [45], Tversky loss [46], Log-Cosh Dice loss [43], and Focal-Tversky loss [47]. As early stopping with patience of 30 epochs was set up to avoid overfitting, none of the models trained up to 200 epochs. The Dice loss method was trained most with 149 epochs. Surprisingly, the remaining four methods were stopped exactly at same epochs of 95. On 267 test images with 448×448 input resolution, we found that the Dice loss had the best overall result with a 0.649 Jac and 0.775 Dsc. Table 4 shows that the four other loss methods had similar performance in Jac and Dsc measurements with differences of less than 0.5%.

3.3. PERFORMANCE COMPARISON WITH DIFFERENT U-NET ARCHITECTURES

Next, we compared the segmentation performance of our proposed architecture with two state-of-the-art U-Net architectures using the same 448x448 input resolution and dice loss. Table 5 shows the segmentation performance of ChimeraNet compared to two DL approaches, U-Net [30], [31] and ResUNet-a [35], on 267 test images. The proposed ChimeraNet outperformed the other two methods in all the measurements. It improved the Jaccard score by 10% from 0.59 to 0.65 and the Dice score by 6.9% from 0.72 to 0.77 on U-Net even with the slightly smaller network size. Although our proposed model has a 74% smaller network size than ResUNet-a, it improved the Jaccard score by 3.2% from 0.63 to 0.65 and the precision by 2.6% from 0.77 to 0.79. Furthermore, we evaluated the performance of ChimeraNet with/out the residual squeeze-and-excitation (SERes) structure. ChimeraNet (basic) is the ChimeraNet without the SERes structure. The results show that the addition of SERes structure improved the segmentation performance without significantly increasing the number of model parameters. Nevertheless, ChimeraNet (basic) also achieved better performance compared to two other U-Net models.

Table 5. Comparison of proposed architecture with existing U-Net Architectures.

Model	**# Params**	**Acc**	**Inacc**	**Dsc**	**Jac**	**Prec**	**Rec**
U-Net [31]	31.5M	97.86	13.10	0.72	0.59	0.73	0.75
ResUNet-a [33]	52.8M	97.79	11.94	0.76	0.63	0.77	0.78
ChimeraNet (basic)	30.3M	98.18	11.27	0.77	0.64	0.78	0.79
ChimeraNet	30.4M	98.23	10.97	0.77	0.65	0.79	0.79

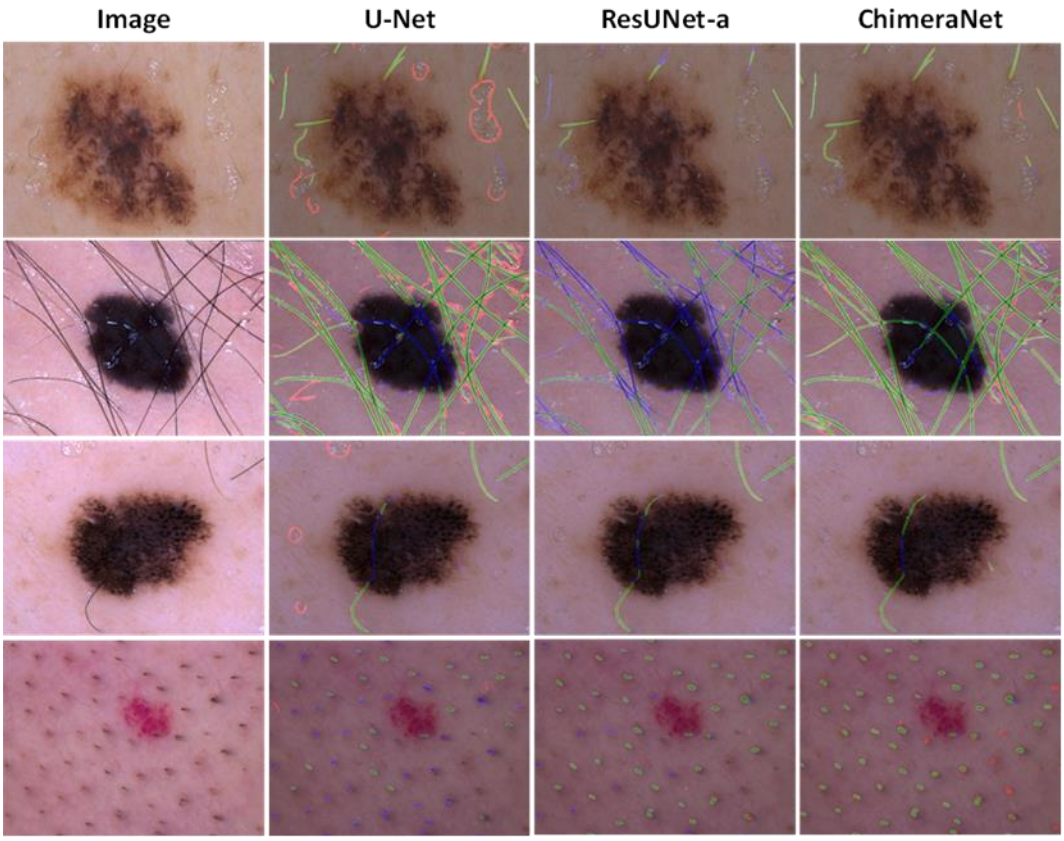


Figure 5. Hair segmentation results of proposed ChimeraNet against U-Net and ResUNet-a for HAM10000 test images. The segmentation results show true positives (GREEN), false positives (RED), and false negatives (BLUE). U-Net model finds more false positives (for example, gel bubbles) and ResUNet-a finds less hair. The proposed ChimeraNet model successfully detects hair with fewer false positives and false negatives.

In Figure 5, we compare the segmentation results of the proposed model with two DL approaches. The examples in the first, second and third rows show that U-Net performed well on long and dark hair but poorly on gel and bubble structures and falsely predicted them as hair and ruler objects. In contrast, ResUNet-a performed well on gel and bubble structures but missed many valid hair objects. In case of short hairs as shown in the fourth row, both U-Net and ResUNet-a had difficulty finding them. Furthermore, the proposed ChimeraNet performed well not only on long and short hair, but it was successful

in separating the gel and bubble structures from hair and ruler objects. It was also able to find difficult hairs such as dark hairs inside a dark lesion as shown in the second and third rows compared to U-Net and ResUNet-a architectures.

3.4. PERFORMANCE COMPARISON WITH EXISTING HAIR SEGMENTATION ALGORITHMS

We compared the performance of our method with seven published hair segmentation algorithms and one on-press [18]–[20], [22]–[24], [27], [36]. Table 6 shows different scoring metrics computed on the second dataset (25 dermoscopic images) for all

Table 6. Results, resolution 448x448, compared with existing hair segmentation algorithms on 25 test images.

Methods & Year of publication	**Acc**	**Inacc**	**Prec**	**Rec**	**Dsc**	**Jac**
ChimeraNet	96.06	11.59	0.79	0.79	0.77	0.64
SharpRazor, 2021	93.80	22.48	0.58	0.58	0.53	0.38
Ian Lee, 2017	90.99	29.39	0.60	0.44	0.40	0.26
Xie, 2015	90.04	39.07	0.37	0.25	0.25	0.15
Koehoorn, 2015	80.13	49.11	0.08	0.14	0.07	0.03
Abbas, 2013	87.36	29.40	0.34	0.49	0.33	0.22
Huang, 2013	81.13	32.94	0.23	0.50	0.26	0.16
Fiorese, 2011	91.74	37.82	0.68	0.26	0.32	0.20
DullRazor,1997	93.15	34.87	0.66	0.31	0.38	0.25

algorithms. For methods except for SharpRazor, the hair masks reported in [16] were kindly provided by Ian Lee. Our method achieved better overall scores across all six measurements, as shown in Table 3. The most significant improvements with our method were 45% in Dice similarity coefficient, 36% in recall, and 68% in Jaccard scores from the existing best scores.

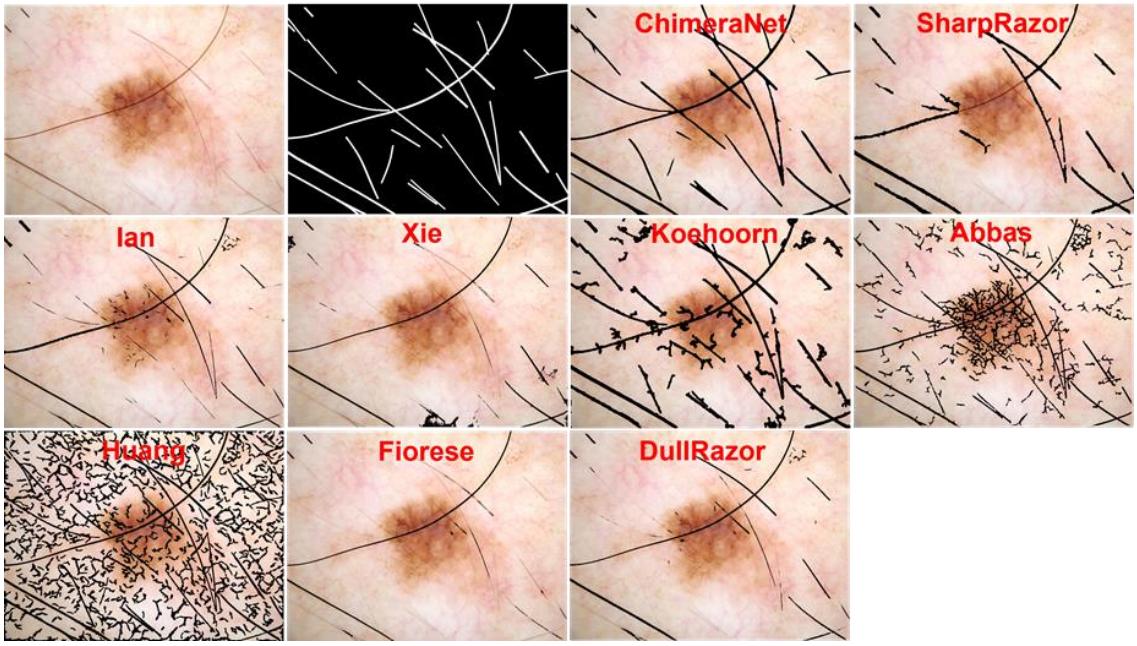


Figure 6. Lesion image, ground truth hair mask, and overlays of predicted hair mask on lesion image for nine hair detection methods. The proposed ChimeraNet method accurately detects more hair with less noise compared to other hair detection methods.

We compare reported hair detection methods with the proposed deep learning method in Figure 6. The proposed method finds thin hairs that other methods fail to detect and finds less noise.

4. DISCUSSION

After Ronneberger et al. [31] proposed the U-Net architecture for biomedical image segmentation, many new variants of U-Net have been developed. One variant used in skin image segmentation is dense residual U-Net, which replaces some convolutions with dense convolutions and appends residual convolutions to the network [48]. This approach may offer better generalization capability and may be more robust on small dataset sizes, but at the cost of slightly decreased overall performance. Diakogiannis et al. [35] proposed ResUNet-a that employs residual connections, atrous (dilated) convolutions, pyramid scene parsing pooling, and multi-tasking inference to segment monotemporal very high-resolution aerial images. Baheti et al. [49] modified U-Net architecture by employing EfficientNet [37] in the encoder network and basic convolution layers in the decoder network for scene segmentation. Here, we propose a new hair segmentation method based on U-Net architecture by employing EfficientNet-B5 as the encoder network and squeeze-and-excitation structures as the building block of the decoder network. We used the squeeze-and-excitation structure [39], which has better generalization capability than basic convolution structures because it focuses on more critical channels of feature maps. Ronneberger et al. [31] showed that the skip-connection helped to improve the semantic segmentation by adding low-level spatial information to the decoder. We put an extra skip-connection from an input to the final block of the decoder which further adds low-level spatial information for precise localization of features. The proposed architecture performed better than U-Net and ResUNet-a to detect fine structures like hair and ruler marks in skin lesion images.

We used several metrics to evaluate the performance of the DL hair detection approach. The accuracy metric counts true-negative pixels and true positive pixels equally. For most images, except on the scalp, most of the pixels are true negative pixels (not hair pixels). Thus, the accuracy metric overstates actual performance. Dice and Jaccard metrics give a better performance evaluation because they are better for scoring detection of structures with less overall representation in the images. Our DL approach gives nearly twice the Jaccard score of any classical method except for the SharpRazor method.

Although most reports consider hair pixels unwanted noise, hair and ruler structures may contribute to diagnostic accuracy. White hairs may indicate patient age. Shaved hairs may determine body location and gender, such as a male face or a female leg. Ruler marks may give an indication of the clinic where the images originated. Although hair structures still constitute noise, their detection could contribute to diagnostic accuracy, and thus they could be considered useful noise. The deep learning approach reported here, which can detect ruler marks and white hair, may yield useful information from the automatically detected structures.

5. CONCLUSION AND FUTURE WORK

In this study, we employ a novel deep learning technique to find hair and ruler structures. Using an image resolution on the order of 448×448, the reported method achieves state-of-the-art accuracy compared to other approaches. One disadvantage of the technique is that it requires many images, with the training, validation, and test sets totaling 1333 images. These training images, which required hundreds of hours to create, can

nevertheless serve as training images for future hair detection methods. In future work, we will make the training hair masks available publicly. In addition, we will employ deep learning to detect other dermoscopic features, using this new ChimeraNet model and comparing it to other models.

Despite the fairly high accuracy achieved by this method, there is a need for future work to improve the results. New approaches could include fusion methods and postprocessing using modeling or other methods. Deep learning architectures have advanced rapidly, and as techniques evolve, other models can be employed.

REFERENCES

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2021," CA Cancer J Clin, vol. 71, no. 1, pp. 7–33, 2021.

[2] H. W. Rogers, M. A. Weinstock, S. R. Feldman, and B. M. Coldiron, "Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the us population, 2012," JAMA Dermatol, vol. 151, no. 10, pp. 1081–1086, 2015, doi: 10.1001/jamadermatol.2015.1187.

[3] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.

[4] L. K. Ferris et al., "Computer-aided classification of melanocytic lesions using dermoscopic images," J Am Acad Dermatol, vol. 73, no. 5, pp. 769–776, Nov. 2015, doi: 10.1016/J.JAAD.2015.07.028.

[5] H. Pehamberger, M. Binder, A. Steiner, and K. Wolff, "In vivo epiluminescence microscopy: Improvement of early diagnosis of melanoma," Journal of Investigative Dermatology, vol. 100, no. 3 SUPPL., pp. S356–S362, 1993, doi: 10.1038/jid.1993.63.

[6] H. P. Soyer, G. Argenziano, R. Talamini, and S. Chimenti, "Is Dermoscopy Useful for the Diagnosis of Melanoma?," Arch Dermatol, vol. 137, no. 10, pp. 1361–1363, Oct. 2001, doi: 10.1001/archderm.137.10.1361.

[7] R. P. Braun, H. S. Rabinovitz, M. Oliviero, A. W. Kopf, and J. H. Saurat, “Pattern analysis: a two-step procedure for the dermoscopic diagnosis of melanoma,” Clin Dermatol, vol. 20, no. 3, pp. 236–239, May 2002, doi: 10.1016/S0738-081X(02)00216-X.

[8] M. A. Marchetti et al., “Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images,” J Am Acad Dermatol, vol. 78, no. 2, pp. 270-277.e1, Feb. 2018, doi: 10.1016/j.jaad.2017.08.016.

[9] H. A. Haenssle et al., “Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists,” Annals of Oncology, vol. 29, no. 8, pp. 1836–1842, 2018, doi: https://doi.org/10.1093/annonc/mdy166.

[10] S. Pathan, K. G. Prabhu, and P. C. 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, doi: 10.1016/J.BSPC.2017.07.010.

[11] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg, “Combining deep learning and hand-crafted features for skin lesion classification,” 2016 6th International Conference on Image Processing Theory, Tools and Applications, IPTA 2016, 2017, doi: 10.1109/IPTA.2016.7821017.

[12] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, “Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images BT - Machine Learning in Medical Imaging,” 2015, pp. 118–126.

[13] N. C. F. Codella et al., “Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images,” IBM J. Res. Dev., vol. 61, no. 4–5, pp. 5:1–5:15, Jul. 2017, doi: 10.1147/JRD.2017.2708299.

[14] I. González-Díaz, “DermaKNet: Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for Skin Lesion Diagnosis,” IEEE J Biomed Health Inform, vol. 23, no. 2, pp. 547–559, 2019, doi: 10.1109/JBHI.2018.2806962.

[15] J. R. Hagerty et al., “Deep Learning and Handcrafted Method Fusion: Higher Diagnostic Accuracy for Melanoma Dermoscopy Images,” IEEE J Biomed Health Inform, vol. 23, no. 4, pp. 1385–1391, 2019, doi: 10.1109/JBHI.2019.2891049.

[16] Q. Abbas, M. E. Celebi, and I. F. Garc\'\ia, “Hair removal methods: A comparative study for dermoscopy images,” Biomed Signal Process Control, vol. 6, no. 4, pp. 395–404, 2011.

[17] G. Celebi, Emre M.; Wen, Quan; Iyatomi, Hitoshi; Shimizu, Kouhei; Zhou, Huiyu; Schaefer, “A State-of-the-Art on Lesion Border Detection in Dermoscopy Images,” in Dermoscopy Image Analysis, J. S. Celebi, M. Emre; Mendonca, Teresa; Marques, Ed. Boca Raton: CRC Press, 2015, pp. 97–129. [Online]. Available: https://doi.org/10.1201/b19107

[18] I. Lee, X. Du, and B. Anthony, “Hair segmentation using adaptive threshold from edge and branch length measures,” Comput Biol Med, vol. 89, no. August, pp. 314–324, 2017, doi: 10.1016/j.compbiomed.2017.08.020.

[19] F. Xie, Y. Li, R. Meng, and Z. Jiang, “No-reference hair occlusion assessment for dermoscopy images based on distribution feature,” Comput Biol Med, vol. 59, pp. 106–115, 2015, doi: 10.1016/j.compbiomed.2015.01.023.

[20] Q. Abbas, I. F. Garcia, M. Emre Celebi, and W. Ahmad, “A feature-preserving hair removal algorithm for dermoscopy images,” Skin Research and Technology, vol. 19, no. 1, pp. e27--e36, 2013.

[21] N. H. Nguyen, T. K. Lee, and M. S. Atkins, “Segmentation of light and dark hair in dermoscopic images: a hybrid approach using a universal kernel,” in Medical Imaging 2010: Image Processing, 2010, vol. 7623, pp. 1436–1443.

[22] T. Lee, V. Ng, R. Gallagher, A. Coldman, and D. McLean, “Dullrazor®: A software approach to hair removal from images,” Comput Biol Med, vol. 27, no. 6, pp. 533–543, Nov. 1997, doi: 10.1016/S0010-4825(97)00020-6.

[23] M. Fiorese, E. Peserico, and A. Silletti, “VirtualShave: Automated hair removal from digital dermatoscopic images,” Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pp. 5145–5148, 2011, doi: 10.1109/IEMBS.2011.6091274.

[24] J. Koehoorn et al., “Automated digital hair removal by threshold decomposition and morphological analysis,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9082, pp. 15–26, 2015, doi: 10.1007/978-3-319-18720-4_2.

[25] M. T. B. Toossi, H. R. Pourreza, H. Zare, M.-H. Sigari, P. Layegh, and A. Azimi, “An effective hair removal algorithm for dermoscopy images,” Skin Research and Technology, vol. 19, no. 3, pp. 230–235, 2013.

[26] O. Abuzaghleh, B. D. Barkana, and M. Faezipour, “Noninvasive real-time automated skin lesion analysis system for melanoma early detection and prevention,” IEEE J Transl Eng Health Med, vol. 3, pp. 1–12, 2015.

[27] A. Huang, S. Y. Kwan, W. Y. Chang, M. Y. Liu, M. H. Chi, and G. S. Chen, “A robust hair segmentation and removal approach for clinical images of skin lesions,” Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pp. 3315–3318, 2013, doi: 10.1109/EMBC.2013.6610250.

[28] H. Zhou et al., “Feature-preserving artifact removal from dermoscopy images,” in Medical Imaging 2008: Image Processing, 2008, vol. 6914, p. 69141B.

[29] M. Attia, M. Hossny, H. Zhou, S. Nahavandi, H. Asadi, and A. Yazdabadi, “Digital hair segmentation using hybrid convolutional and recurrent neural networks architecture,” Comput Methods Programs Biomed, vol. 177, pp. 17–30, 2019, doi: 10.1016/j.cmpb.2019.05.010.

[30] W. Li, A. N. Joseph Raj, T. Tjahjadi, and Z. Zhuang, “Digital hair removal by deep learning for skin lesion segmentation,” Pattern Recognit, vol. 117, 2021, doi: 10.1016/j.patcog.2021.107994.

[31] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation.” [Online]. Available: http://lmb.informatik.uni-freiburg.de/

[32] F. Rodriguez, E. Maire, P. Courjault-Radé, and J. Darrozes, “The Black Top Hat function applied to a DEM: A tool to estimate recent incision in a mountainous watershed (Estibère Watershed, Central Pyrenees),” Geophys Res Lett, vol. 29, no. 6, pp. 1–9, 2002.

[33] P. Tschandl, C. Rosendahl, and H. Kittler, “Data Descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions Background & Summary,” Nature Publishing Group, 2018, doi: 10.1038/sdata.2018.161.

[34] N. Codella et al., “Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC).”

[35] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, “ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 162, no. March 2019, pp. 94–114, 2020, doi: 10.1016/j.isprsjprs.2020.01.013.

[36] R. Kasmi et al., “SharpRazor: Automatic Removal of Hair and Ruler Marks from Dermoscopy Images,” Skin Research and Technology, 2021.

[37] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in International conference on machine learning, 2019, pp. 6105–6114.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition, 2009, pp. 248–255.

[39] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[40] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[41] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.”

[42] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[43] S. Jadon, “A survey of loss functions for semantic segmentation,” 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2020, 2020, doi: 10.1109/CIBCB48159.2020.9277638.

[44] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, “Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations,” in Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, 2017, pp. 240–248.

[45] M. A. Rahman and Y. Wang, “Optimizing intersection-over-union in deep neural networks for image segmentation,” in International symposium on visual computing, 2016, pp. 234–244.

[46] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, “Tversky loss function for image segmentation using 3D fully convolutional deep networks,” in International workshop on machine learning in medical imaging, 2017, pp. 379–387.

[47] N. Abraham and N. M. Khan, “A novel focal tversky loss function with improved attention u-net for lesion segmentation,” in 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), 2019, pp. 683–687.

[48] N. C. F. Codella et al., “Segmentation of both diseased and healthy skin from clinical photographs in a primary care setting,” in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 3414–3417.

[49] B. Baheti, S. Innani, S. Gajre, and S. Talbar, “Eff-UNet: A novel architecture for semantic segmentation in unstructured environment,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol. 2020-June, pp. 1473–1481, 2020, doi: 10.1109/CVPRW50498.2020.00187.

II. SKIN LESION SEGMENTATION IN DERMOSCOPIC IMAGES WITH NOISY DATA

Norsang Lama[1], Jason R. Hagerty[2], R. Joe Stanley[1], William V. Stoecker[2]

[1]Department of Electrical and Computer Engineering,
Missouri University of Science and Technology, Rolla, MO, USA
[2]S&A Technologies, Rolla, MO, 65401 USA

ABSTRACT

We propose a deep learning approach to segment the skin lesion in dermoscopic images. The proposed network architecture uses a pretrained EfficientNet model in the encoder and squeeze-and-excitation residual structures in the decoder. We applied this approach on the publicly available International Skin Imaging Collaboration (ISIC) 2017 Challenge skin lesion segmentation dataset. This benchmark dataset has been widely used in previous studies. We observed many inaccurate or noisy ground truth labels. To reduce noisy data, we manually sorted all ground truth labels into three categories – good, mildly noisy, and noisy labels. Further, we investigated the effect of such noisy labels in training and test sets. Our test results show that the proposed method achieved Jaccard scores of 0.807 on the official ISIC 2017 test set and 0.832 on the curated ISIC 2017 test set, exhibiting better performance than previously reported methods. Further, the experimental results showed that the noisy labels in the training set did not lower the segmentation performance. However, the noisy labels in the test set adversely affected the evaluation scores. We recommend that the noisy labels should be avoided in the test set in future studies for accurate evaluation of the segmentation algorithms.

Keywords— Melanoma, dermoscopy, deep learning, image segmentation, noisy data

1. INTRODUCTION

An estimated 99,780 new cases of invasive melanoma and 97,920 in-situ melanoma will be diagnosed in 2022 in the United States [1]. Dermoscopy is an imaging modality to aid dermatologists for the early detection of skin cancer and can improve diagnostic accuracy over clinical visual inspection by the experienced domain expert [2]–[4].

Computer vision techniques have improved appreciably in recent years [5]–[9] and have been successfully applied to many medical imaging problems [10]–[13]. In the skin cancer domain, deep learning techniques combined with dermoscopy have higher diagnostic accuracy than experienced dermatologists [10], [14]–[17]. Pathan et al. published a recent review detailing both handcrafted and deep learning (DL) techniques for computer-aided diagnosis of skin lesions [18]. Although deep learning eliminates a tedious feature engineering process, recent studies show that the fusion of deep learning and handcrafted features can improve accuracy in skin cancer diagnosis [17], [19]–[22]. Handcrafted features are not as straightforward as the deep learning method, and they require a lesion border to define the region of interest. Accurate calculation of handcrafted lesion features depends upon correct detection of the lesion border [22]. Thus, lesion segmentation is an important step in computer-aided diagnosis of skin cancer.

Traditional image processing methods were applied to segment the skin lesion in dermoscopic images [23]–[25]. These methods performed well on small sets but generated

unsatisfactory results when applied to challenging conditions such as low contrast between lesion and background, lesions with different colors, and images with artifacts like hair, ruler marks, gel bubbles, and ink markers, etc. Deep learning techniques have overcome these challenges to some extent and improved border detection in skin lesion images [26]–[30].

Al-Masni et al. [26] proposed a deep full-resolution convolutional network for skin lesion segmentation. Unlike U-Net [31], this method does not employ upsampling or downsampling operations so that the feature maps always have the same resolution from the input to the output. The deep learning methods require little preprocessing and the RGB color images are directly fed to the network. However, recent studies showed that adding more input color channels improves skin lesion segmentation; Yuan et al. [28] combined three RGB channels, three HSV channels and one L channel of CIELAB color space and input 7-channel images to their deep neural network model and showed improved results. Ozturk et al. [30] also used 7 channels in their deep learning method, however, their approach was slightly different. The first input layer took three RGB channels and four additional channels (S of HSV color space, I of YIQ color space, B of CBR color space and Z of XYZ color space) were fed to deeper intermediate layers.

Xie et al. [29] created a high-resolution feature block (HRFB) having three branches – a normal convolutional branch, a spatial attention branch, and a channel attention branch. Tong et al. [32] used an extended U-Net architecture and proposed ASCU-Net by employing a triple attention mechanism of attention gate [33], spatial attention module, and channel attention module.

A transfer learning approach has also been applied to skin lesion segmentation problems. Kadry et al. [34] and Rajinikanth [35] employed a pretrained VGG [7] network to encode the important features from the skin lesion image and then upsampled the feature maps repeatedly to generate the segmentation mask. Zafar et al. [36] employed a ResNet-50 [9] architecture pretrained on ImageNet [37] as the encoder network in their U-Net architecture. A similar method by Tschandl et al. [27] also used a ResNet-34 architecture as the encoding layers and investigated the effect of random weight initialization versus domain-specific or ImageNet pretraining. Nawaz et al. [38] presented a U-Net architecture using DenseNet [39] encoder to segment melanoma lesion of varying colors and sizes. Nguyen et al. [40] integrated a pretrained EfficientNet-B4 [41] and the residual blocks in their U-Net architecture. Despite the early success of deep learning methods on skin lesion segmentation, many current architectures still fail to produce satisfactory results on challenging conditions like low skin-versus-lesion contrast and presence of image artifacts like hair or ruler marks, ink markers and gel bubbles. Another concern we found was the presence of inaccurate or noisy ground truth (GT) masks in the benchmark ISIC 2017 [42] skin lesion segmentation datasets used in previous studies. These noisy GTs in the benchmark dataset warrant investigation to determine their effect on skin lesion segmentation.

In this study, we propose a deep learning method to improve skin lesion segmentation in dermoscopic images. The proposed method uses a modified ChimeraNet [43] architecture that was used to detect hair and ruler marks in dermoscopic images. The main contributions of this paper are as follows:

- The proposed method achieves state-of-the-art segmentation performance on the ISIC 2017 skin lesion segmentation dataset. This segmentation improvement can benefit conventional analysis of the lesion, which depends on accurate segmentation.
- We identify noisy or inaccurate ground truth labels in the benchmark public dataset.
- We investigate the effect of pruning the noisy or inaccurate ground truth labels from the dataset.

2. MATERIALS AND METHODS

This section discusses the materials and methods used in this study. Our proposed method has three stages – annotation curation, training, and evaluation or inference. First, a dermatologist or specialist assesses the quality of ground truth annotations in a benchmark public dataset. Second, a U-Net segmentation model is trained using the curated training set. Finally, the trained model is evaluated on the curated test set. The overall flow diagram of the proposed method is shown in Figure 1.

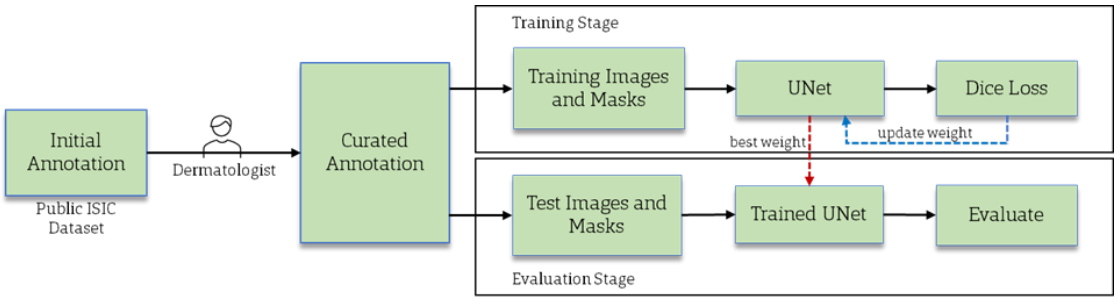


Figure 1. The overall flow diagram of a proposed skin lesion segmentation method.

2.1. IMAGE DATASETS

The dataset used in this study is the publicly available ISIC 2017 [42] skin lesion segmentation dataset. It is a large skin lesion segmentation dataset released as a part of the 2017 International Skin Imaging Collaboration (ISIC) Challenge. It provides 2750 dermoscopic skin lesion images with lesion boundary masks - 2000 training, 150 validation, and 600 test images. The ground truth (GT) lesion boundary masks were determined under the supervision of expert clinicians using both manual annotation and semi-automated process as shown in Figure 2. The images are 8-bit RGB images with varying height and width ranging from a few hundred pixels to a few thousand pixels. As the dataset provides a single train-validation split, we combined the official training and validation sets to create a single training set of 2150 images to run 5-fold cross-validation experiments. The official 600 test images were used as a holdout test set to evaluate the performance of our proposed method against the state-of-the-art methods.

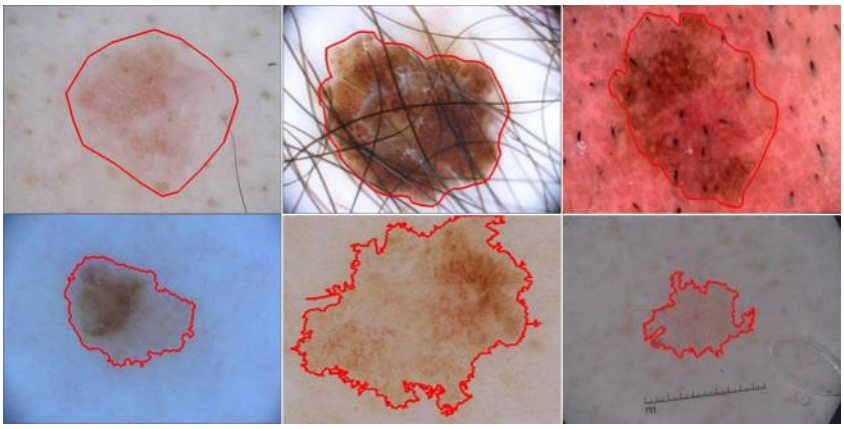


Figure 2. Skin lesion dermoscopy images with ground truth lesion boundary (RED) from publicly available ISIC skin lesion datasets. The masks are manually drawn (first row) or generated using a semi-automated process (second row).

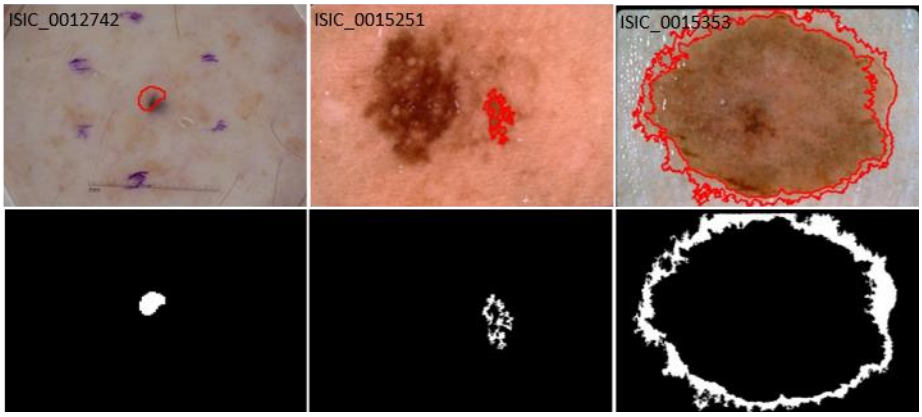


Figure 3. Examples of inaccurate or noisy ground truths on ISIC lesion segmentation dataset. Overlays show GT lesion boundaries on lesion images (top row) and ground truth lesion segmentation mask (bottom row). The lesion boundary (RED) fails to cover the whole lesion in all examples.

Table 1. Number of images with good, mildly noisy, and noisy lesion boundary labels in ISIC 2017 train and test sets.

Image Set	**Good**	**Mildly Noisy**	**Noisy**
Train + Validation (2150)	1982	149	19
Test (600)	493	87	20
Total (2750)	2475	236	39

As the GT masks were created using both manual and semi-automated processes, we found some of the ground truth masks, especially those determined automatically, were inaccurate (Figure 3). The noisy labels or inaccurate examples in the training set might affect the model adversely, reducing accuracy. Conversely, noisy labels might aid performance by increasing the number of training examples or regularizing the overparameterized deep learning model. Also, the noisy labels in the test set might not

demonstrate a true evaluation of the model. Thus, all 2750 GT masks, including both train and test sets, were re-evaluated by a dermatologist and categorized into three categories - good, mildly noisy, and noisy. The number of GT masks in each category after reevaluation are shown in Table 1.

2.2. DATA AUGMENTATION

Data augmentation can be applied during the training of deep neural networks to increase the number of training images without adding new images. Augmentation will result in better generalization of deep network models and reduce the overfitting problem. Data augmentation performs different image transform methods on the original training images to generate more examples for training. In this study, we selected the following image transforms for data augmentation:

- Height or width shift with a range of (-0.15, +0.15)
- Horizontal or vertical flip
- Rotation with range between +90° to -90°
- Zoom with a range (-0.15, +15)
- Brightness with a range of (0.85, 1.15)
- Contrast with a range of (0.85, 1.15)

Furthermore, all images were resized to 448×448 and the image pixel values were rescaled between 0 and 1. Finally, the images were normalized before feeding them to the deep network.

2.3. NETWORK ARCHITECTURE

In this study, we used a modified U-Net [31] convolutional neural network (CNN) architecture for skin lesion segmentation by Lama et al. [43]. The proposed encoder-decoder based image segmentation model, named ChimeraNet, uses a pretrained EfficientNet [41]model in the encoder and squeeze-and-excitation [44] structures in the decoder. Further, we applied a dilated convolution [45] operation in place of a regular convolution operation in these squeeze-and-excitation residual blocks. As artifacts like hair, ruler marks, and purple marks hinder the detection of important features from skin lesion images [23], [46], we adopted the CNN architecture that was already successful in segmenting fine structures like hair and ruler marks from the skin lesion images. However, a few minor modifications were performed on the original ChimeraNet[43] model to accommodate skin lesion segmentation task. The overall pipeline of the proposed UNet architecture is shown in Figure 4.

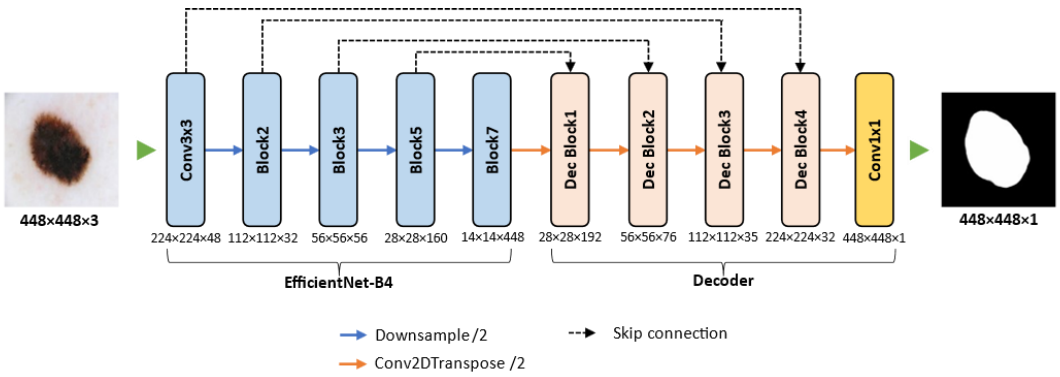


Figure 4. Proposed architecture for skin lesion segmentation. An encoder-decoder architecture with pretrained EfficientNet model as the encoder network and the decoder network comprised of four squeeze-and-excitation residual blocks.

2.3.1. Encoder. In the encoder part, we used the EfficientNet [41] model pretrained on the ImageNet [37] image classification challenge dataset. EfficientNets are composed of mobile inverted bottleneck convolution (MBConv) structures and have 8 network variants from EfficientNet-B0 to EfficientNet-B7. These networks use multiple MBConv blocks grouped together to form seven larger blocks named Block1 to Block7 as given in Table 2. EfficientNetB0 is the baseline architecture, and other variants are scaled up by employing a compound scaling method that uniformly scales network depth, width, and resolution with a fixed set of scaling factors. In the proposed model, we used the pretrained EfficientNet-B4 variant of EfficientNet models as the encoder network. Like many CNN architectures, the EfficientNet model downsamples the feature map repeatedly while extracting the most useful features from the image. The spatial dimension of the final feature map gets much smaller than the original dimension of an input image. The dimensions of feature maps at different levels are given in Table 2.

Conversely, the decoder network needs to expand these low-resolution feature maps to generate the segmentation map with spatial dimensions equal to those of the input image. The U-Net architecture uses the skip-connections to recover the spatial information lost in the encoder due to downsampling process. For precise localization of features, the skip-connection feeds the high-resolution output feature maps at various levels in the encoder to the decoder by skipping some blocks as shown in Figure 4. In the proposed method, we used the outputs of Conv3×3, Block2, Block3 and Block5 as sources of the skip-connections. These blocks are selected for skip connections because the size of output feature maps is downsampled by a factor of 2 in the subsequent block. The output

dimensions of each block corresponding to the skip connections and the final output of the encoder are given in Table 2.

Table 2. Different blocks of EfficientNet-B4 model and their output feature map sizes and the number of channels.

Block Name	**Feature Map Size (W × H)**	**#Feature Map (C)**
Input layer	448 × 448	3
Conv3×3	224 × 224	48
Block 1	224 × 224	24
Block 2	112 × 112	32
Block 3	56 × 56	56
Block 4	28 × 28	112
Block 5	28 × 28	160
Block 6	14 × 14	272
Block 7	14 × 14	448

2.3.2. Decoder. The decoder network is constructed using a squeeze-and-excitation residual (SERes) structure [44] as shown in Figure 5. The SERes block has a better feature representation capability than the plain convolution block as it emphasizes the informative features and suppresses the weaker ones by modeling the interdependencies between channels of convolutional features. The decoder network has 4 blocks named Dec Block1 to Dec Block4, as shown in Figure 4. Each decoder block is composed of a SERes block and gets two feature maps as inputs – an output feature map from the previous stage and a

low-level feature map via a skip-connection from the encoder. For example, the first block (Dec Block1) of the decoder gets 14×14×448 feature input from the previous stage, the final output of the encoder, and 28×28×160 low-level feature input from Block5 via a skip-connection. Here, three dimensions of the feature map represent width (W), height (H) and number of feature map or channel (C). Both feature inputs are concatenated before feeding to the SERes block. However, the dimensions of both inputs are not the same. To combine both inputs, first, the 14×14×448 feature map from the previous stage is upsampled using a transposed convolution, also called deconvolution. The transposed convolution performs 2×2 upsampling followed by a 3×3 convolution operation. We selected the number of filters for transposed convolution as half of the number of input channels, i.e. 224 (=448/2), thus generating a 28×28×224 feature map. Then, the two inputs are concatenated along the channel axis to form 28×28×384 feature map before feeding to the SERes block. The SERes block combines an SE block with a residual structure [44], as shown in Figure 5. The residual unit in the SERes block is a double convolution block, which applies two sets of 3×3 dilated convolutions (dilation rate = 2), batch normalization [47], and rectified linear unit (ReLU) operations. Again, we selected the number of filters for two convolution layers in the residual unit as half of the input channels, i.e. 192 (= 384/2). The residual unit outputs a 28×28×192 feature map, and then the squeeze-and-excitation operation is performed to scale the features along the channel axis. To find the weights for each channel of the feature map, SE first applies global average pooling to reduce the feature map to 1×1×192 and then applies non-linear operations like FC, ReLU, FC, and sigmoid. The number of neurons in two FC layers are C/r and C, respectively, where r is a feature reduction factor and empirically selected as r = 8. The SE generates a 1×1×192 weight vector with each value

in the range of 0 to 1. Then the residual feature is multiplied with a weight vector to scale the features and generate a 28×28×192 scaled feature map.

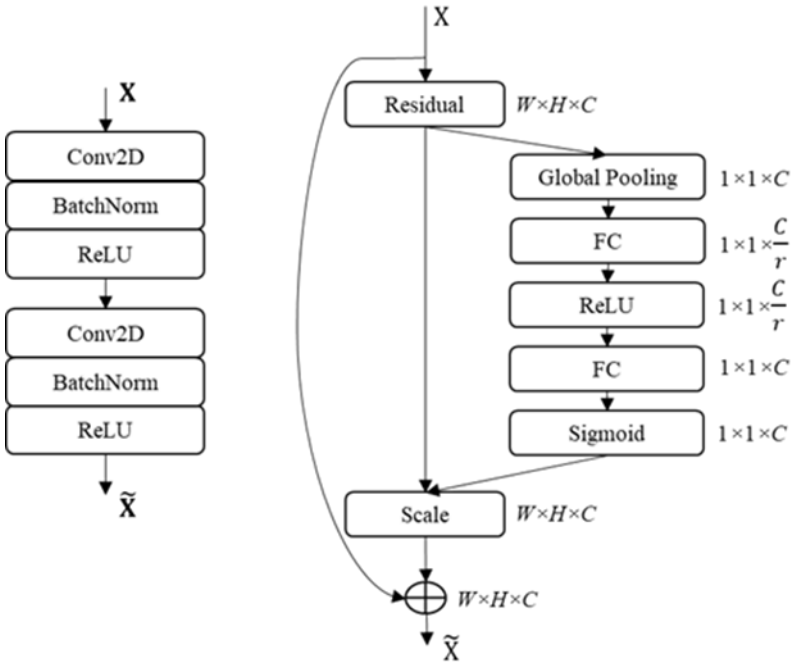


Figure 5. Structures of convolution blocks in the decoder network. Double convolution block (left) and squeeze-and-excitation residual block (right).

Further, the SERes block combines this scaled residual feature map with the original input feature map. However, the channels in the original input feature map (28×28×224) and residual feature output (28×28×192) are not the same, so a 1×1 convolution operation with 192 filters followed by a batch normalization operation are performed on the original input feature map. Then, the SERes block adds two feature maps together and applies the ReLU operation to generate the final 28×28×192 feature output.

Similarly, the remaining decoder blocks (Dec Block2 to Dec Block4) apply the same set of operations as Dec Block1. Only the size and the number of feature maps are

different, as shown in Table 3. The number of feature maps (C) corresponds to the number of convolutional filters applied in each SERes block. Also, the dropout operation with 0.4 probability was applied after each decoder block to regularize the network from the overfitting during the training. The output resolution of the final decoder block, Dec Block4, is still smaller than the original input resolution so it is upsampled by a factor of 2. Finally, 1×1 convolution and a sigmoid function are applied to generate the final segmentation map of size 448×448×1. The 1x1 convolution reduces the number of channels to the desired number of classes, and the sigmoid operation converts all pixel values to the range between 0 and 1. Each pixel value in the segmentation map represents the probability score of that pixel belonging to the skin lesion.

Table 3. SERes blocks in the decoder and their output sizes.

Block Name	**Size (W × H)**	**Feature Map (C)**
Dec Block1	28 × 28	192
Dec Block2	56 × 56	76
Dec Block3	112 × 112	35
Dec Block4	224 × 224	32

During inference, we give five different augmented versions of an input image to the trained deep network: an original image, a horizontally flipped image, a vertically flipped image, a 90° clockwise rotated image, and a 90° counterclockwise rotated image. The deep network generates the segmentation output for each image, and the final segmentation mask is generated by aggregating these five outputs using the unweighted

average of the five predicted masks. The mask is binarized using the threshold of 0.5 to generate the final segmentation mask.

2.4. TRAINING DETAILS

All models were built using Keras with a Tensorflow backend in Python 3 and trained using a single 32GB Nvidia V100 graphics card. We used a 5-fold cross-validation method to tune the hyperparameters, which are shown in Table 4. The networks were trained using a Dice [48] loss function and Adam [49] optimization algorithm. To reduce overfitting of a deep neural network model, we used data augmentation (see details in section 2.2), a dropout layer, and an early stopping technique. The dropout probability of 0.4 was selected for the dropout layers in each decoder block. All images were resized to 448x448 using bilinear interpolation.

Table 4. Training hyperparameters.

Parameter	**Value**
image size	448x448
learning rate	0.0001
batch size	10
epoch	200
dropout probability	0.4
optimizer	Adam
loss method	dice
early stopping patience	30

3. EXPERIMENTAL RESULTS

We evaluated the performance of the proposed method by comparing the predicted lesion segmentation masks with the provided ground truth masks on the official ISIC 2017 [42] skin lesion segmentation dataset having 600 test images. In addition, the proposed method was also evaluated on curated ISIC 2017 test sets. The evaluation metrics used are Jaccard index (Jac), Dice similarity coefficient (Dsc), and accuracy (Acc).

3.1. SEGMENTATION PERFORMANCE OF THE PROPOSED AND STATE-OF-THE-ART METHODS ON ISIC 2017 TEST IMAGES

In this section, we compared the lesion segmentation performance of the proposed method on 600 ISIC 2017 test images with the previously reported methods, as shown in Table 5. The proposed method achieved the highest Jaccard score of 0.807 compared to the state-of-the-art methods [53].

In Figure 6, we show the segmentation results of the proposed method on ISIC 2017 test images. The segmentation results showed that the proposed method successfully finds the lesion border despite the presence of hair, ruler marks, ink marker, and sticker artifacts. Also, the proposed method accurately segments the skin lesion from the background in challenging images having low contrast between the skin and the lesion (see row 3). The predicted masks have smooth lesion borders (blue) compared to the jagged ground truth lesion borders (red) generated by semi-automated processes (see third column).

Table 5. Performance comparison with other lesion segmentation methods on the original ISIC 2017 test dataset.

Methods	**Year**	**Jac**	**Dsc**	**Acc**
Al-Masni et al. [26]	2018	0.771	0.871	0.940
Tschandl et al. [27]	2019	0.768	0.851	
Yuan and Lo [28]	2019	0.765	0.849	0.934
Navarro et al.[50]	2019	0.769	0.854	0.955
Xie et al.[29]	2020	0.783	0.862	0.938
Ozturk and Ozkaya [30]	2020	0.783	0.886	0.953
Shan et al.[51]	2020	0.763	0.846	0.937
Kaymak et al.[52]	2020	0.725	0.841	0.939
Nguyen et al.[40]	2020	0.781	0.861	
Zafar et al.[36]	2020	0.772	0.858	
Goyal et al. [53]	2020	0.793	0.871	
Tong et al.[32]	2021	0.742		0.926
Chen et al.[54]	2022	0.8036	0.8704	0.9471
Ashraf et al.[55]	2022	0.8005		
Our Method		0.807	0.880	0.948

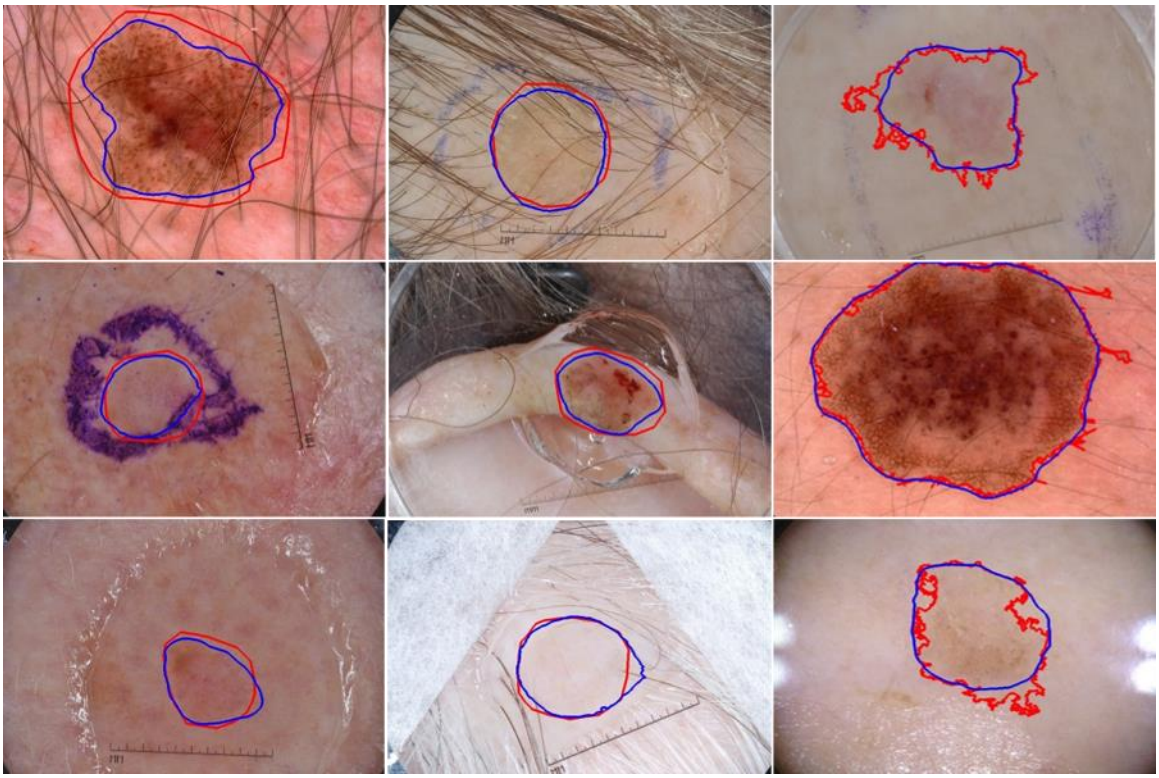


Figure 6. Segmentation results of the proposed method on ISIC 2017 test set. Overlays of ground truth lesion boundary (RED) and predicted lesion boundary (BLUE) on skin lesion images. Lesion border predictions are accurate even in the presence of artifacts like hair, ruler marks, ink markers, etc.

3.2. EFFECT OF PRUNING THE NOISY GT LABELS FROM ISIC 2017 DATASET

In this section, we investigated the effect of pruning the noisy ground truth (GT) labels from both training and test sets of the ISIC 2017 lesion segmentation dataset. Table 6 shows the segmentation performance of the proposed method on 600 test images before and after pruning the noisy labels from the dataset.

First, we removed the noisy GT labels from the training set. When 19 noisy labels were removed from the training set of 2150 images, there was no significant change in the performance per Jaccard scores (0.807 vs. 0.806) on 600 test images. However, when both

mildly noisy and noisy labels (168 images) were removed, the performance slightly decreased from a Jaccard score of 0.807 to 0.802 on 600 test images. Larger training sets provide more examples, advantageous for training the deep learning model even if the labels are noisy or mildly noisy.

Second, we removed the noisy GT labels from 600 ISIC 2017 test images. The model trained on the full training set (2150 images) improved the Jaccard score by 0.01 from 0.807 to 0.817 (a 1 % improvement) when 20 noisy labels were removed. Further, when both noisy (= 20) and mildly noisy (= 87) GT labels were removed, we achieved the highest Jaccard score of 0.832, which is 2.5 % improvement from 0.807.

Table 6. Segmentation performance comparison of the proposed method before and after pruning noisy and mildly noisy GT labels from ISIC 2017 train and test sets

Train Pruned (N_{train})	**Test Pruned (N_{test})**	**Jac**	**Dsc**	**Acc**
None (2150)	*None* (600)	0.807	0.880	94.779
	Noisy (580)	0.817	0.889	95.536
	Noisy + Mildly Noisy (493)	0.832	0.900	96.393
Noisy (2131)	*None* (600)	0.806	0.878	94.723
	Noisy (580)	0.815	0.887	95.518
	Noisy + Mildly Noisy (493)	0.827	0.895	96.210
Noisy + Mildly Noisy (1982)	*None* (600)	0.802	0.875	94.658
	Noisy (580)	0.812	0.885	95.398
	Noisy + Mildly Noisy (493)	0.824	0.893	96.073

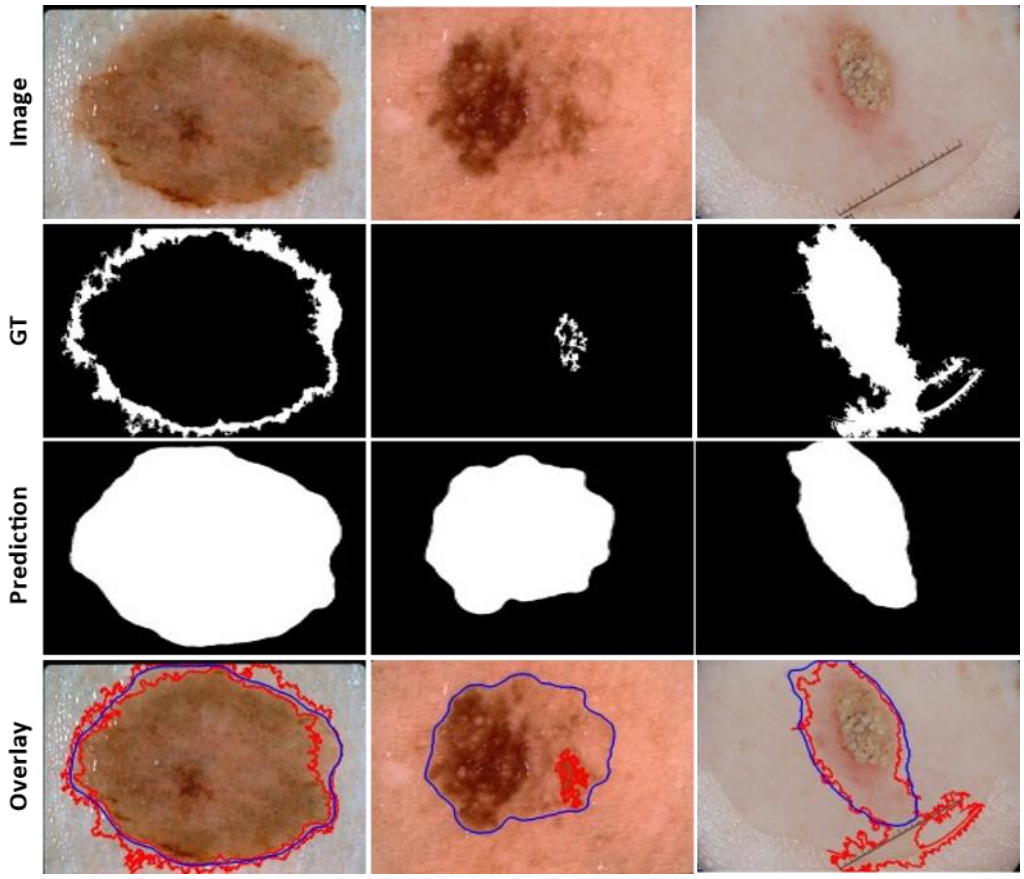


Figure 7. Segmentation results of the proposed method on examples having noisy (or inaccurate) ground truth (GT) on an official ISIC 2017 test set. The predicted lesion borders (BLUE) cover the lesion area more accurately than the GT lesion border (RED).

In Figure 7, we showed the segmentation results of the proposed method on the ISIC2017 test images having noisy or inaccurate ground truth masks. The overlays of the predicted lesion boundary (indicated by blue line) and the ground truth lesion boundary (indicated by red line) on the third-row show that the predicted segmentation covers the lesion area more accurately than the ground truth lesion mask.

4. DISCUSSIONS

In this study, we demonstrated that our proposed deep learning method successfully detects the skin lesion boundary on most dermoscopic skin lesion images. We scored segmentation performance using the Jaccard index, dice similarity coefficient, and accuracy. As the accuracy metric counts true-negative pixels and true-positive pixels equally, accuracy overstates the actual performance when positive (lesion) and negative (background) pixels are highly imbalanced. Accordingly, accuracy is less useful than the other methods in assessing segmentation performance for lesions which occupy a small area of the image.

The proposed network architecture in this study was the same model, ChimeraNet, used in detecting hair and ruler marks in dermoscopic images [43]. The encoder-decoder architecture uses a pre-trained EfficientNet [41] model as the encoder network and a squeeze-and-excitation residual [44] structure as the convolutional block to construct the decoder network. The proposed method performed better than the state-of-the-art methods on the skin lesion segmentation task, with the highest Jaccard scores of 0.807 on the official ISIC2017 test set and 0.832 on the curated ISIC2017 test set. A very similar method was employed by Nguyen et al. [40], with a pretrained EfficientNet [41] model as an encoder network. Our proposed model improved the Jaccard score of Nguyen by 0.026, from 0.781 to 0.807, (a 3.2% improvement) on the ISIC 2017 test set. The main difference was the use of a different decoder network which employs a squeeze-and-excitation residual structure and dilated convolution operations. The squeeze-and-excitation convolutional structure improved the segmentation performance by focusing on more critical channels of the

feature maps [44], resulting in a better feature representation and generalization than the basic convolutional blocks. The dilated convolution operations give the larger receptive field without increasing the number of filter parameters. Lama et al. [43] compared loss functions and various U-Net architectures and found that the U-Net architecture presented here was best in the dermoscopy domain. Thus, ablation studies comparing various architectures are not repeated here.

Although the ISIC 2017 skin lesion segmentation dataset is the largest publicly available and most-used dataset in skin lesion segmentation studies in the deep learning era, we found many inaccurate ground truth masks in the dataset. These inaccurate GT masks might affect the segmentation performance of the deep learning model. Thus, our dermatologist manually reevaluated all ground truth masks and graded them into three categories – good, mildly noisy, and noisy. Then we conducted multiple experiments to analyze the effect of removing the noisy or inaccurate ground truth masks from both training and test sets. The results in Table 6 show that the model trained on the complete training set performed slightly better than the model trained on the curated training set (after removing noisy and mildly noisy examples). The full training set model had 0.807 Jaccard and 0.880 Dice scores on 600 ISIC2017 test images, while the curated training set model only achieved 0.802 Jaccard and 0.875 Dice scores. These experimental results show that the presence of noisy or inaccurate labels in the training set does not reduce the model's performance. Instead, some noisy or inaccurate labels in the training set might provide a regularization effect for the overparameterized deep learning model and thus generalize better, aside from the beneficial effect of a more extensive training set. Conversely, the noisy or inaccurate GTs in the test set adversely affected the evaluation scores. The Jaccard

and Dice similarity scores were improved from 0.807 to 0.832 and 0.88 to 0.90, respectively when the noisy and mildly noisy GT labels were removed from the official ISIC 2017 lesion segmentation test set. This result shows that the segmentation performance is significantly underestimated when evaluated on the test set having noisy or inaccurate GT labels. As many previous studies have used the official test set to evaluate their method against the state-of-the-art methods, comparisons might not be fair and accurate.

Image segmentations created by ChimeraNet deep learning are subjectively improved compared to both automatic and manual borders. The excessive jaggedness of the automatic borders is remedied with the new technique. The manual borders, characterized by straight lines joined at points, are smoothed. Both types of distortion in the ground truth segmentations, excessive jaggedness, and straight-line junctions, are non-physiologic and may lead to error in handcrafted feature analysis which depends upon an accurate border.

There are limitations to this study. Only one dermatologist scored the accuracy of the segmentations. No new noisy segmentations were added to the benchmark ISIC 2017 dataset to see the effect of noisy data at different proportions in the training set. The experiments were conducted using only the available noisy data in the original dataset. Further, we did not create new ground truths for the noisy or inaccurate ground truths.

5. CONCLUSION

In this study, we employed a novel deep learning technique to segment skin lesions in dermoscopic images. The proposed method performed better than the previous state-of-the-art methods. We observed the presence of noisy or inaccurate ground truth labels in a large benchmark dataset. With help of a dermatologist, we manually re-evaluated the ground truth masks. Further, we investigated the effect of noisy ground truth labels in the benchmark dataset. Our experimental results show that more training data, including noisy data, yields better performance than the condensed curated data. However, the noisy data adversely affects the evaluation scores when present in the test data. The test scores were improved when the noisy or inaccurate labels were removed from the official test set. We recommend that future researchers avoid the noisy data in the test set for a fair and accurate evaluation of their lesion segmentation algorithms.

REFERENCES

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, “Cancer statistics, 2022,” CA Cancer J Clin, vol. 72, no. 1, pp. 7–33, 2022, doi: https://doi.org/10.3322/caac.21708.

[2] H. Pehamberger, M. Binder, A. Steiner, and K. Wolff, “In vivo epiluminescence microscopy: Improvement of early diagnosis of melanoma,” Journal of Investigative Dermatology, vol. 100, no. 3 SUPPL., pp. S356–S362, 1993, doi: 10.1038/jid.1993.63.

[3] H. P. Soyer, G. Argenziano, R. Talamini, and S. Chimenti, “Is Dermoscopy Useful for the Diagnosis of Melanoma?,” Arch Dermatol, vol. 137, no. 10, pp. 1361–1363, Oct. 2001, doi: 10.1001/archderm.137.10.1361.

[4] R. P. Braun, H. S. Rabinovitz, M. Oliviero, A. W. Kopf, and J. H. Saurat, “Pattern analysis: a two-step procedure for the dermoscopic diagnosis of melanoma,” Clin Dermatol, vol. 20, no. 3, pp. 236–239, May 2002, doi: 10.1016/S0738-081X(02)00216-X.

[5] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in Advances in Neural Information and Processing Systems (NIPS), vol. 25, 2012, pp. 1097–1105.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[8] I. Goodfellow et al., “Generative adversarial networks,” Commun ACM, vol. 63, no. 11, pp. 139–144, 2020.

[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[10] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.

[11] V. Gulshan et al., “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.

[12] S. Sornapudi et al., “Deep learning nuclei detection in digitized histology images by superpixels,” J Pathol Inform, vol. 9, no. 1, p. 5, 2018.

[13] G. Litjens et al., “A survey on deep learning in medical image analysis,” Med Image Anal, vol. 42, pp. 60–88, 2017, doi: https://doi.org/10.1016/j.media.2017.07.005.

[14] L. K. Ferris et al., “Computer-aided classification of melanocytic lesions using dermoscopic images,” J Am Acad Dermatol, vol. 73, no. 5, pp. 769–776, Nov. 2015, doi: 10.1016/J.JAAD.2015.07.028.

[15] M. A. Marchetti et al., “Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images,” J Am Acad Dermatol, vol. 78, no. 2, pp. 270-277.e1, Feb. 2018, doi: 10.1016/j.jaad.2017.08.016.

[16] H. A. Haenssle et al., “Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists,” Annals of Oncology, vol. 29, no. 8, pp. 1836–1842, 2018, doi: https://doi.org/10.1093/annonc/mdy166.

[17] N. C. F. Codella et al., “Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images,” IBM J. Res. Dev., vol. 61, no. 4–5, pp. 5:1–5:15, Jul. 2017, doi: 10.1147/JRD.2017.2708299.

[18] S. Pathan, K. G. Prabhu, and P. C. 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, doi: 10.1016/J.BSPC.2017.07.010.

[19] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg, “Combining deep learning and hand-crafted features for skin lesion classification,” 2016 6th International Conference on Image Processing Theory, Tools and Applications, IPTA 2016, 2017, doi: 10.1109/IPTA.2016.7821017.

[20] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, “Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images BT - Machine Learning in Medical Imaging,” 2015, pp. 118–126.

[21] I. González-Díaz, “DermaKNet: Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for Skin Lesion Diagnosis,” IEEE J Biomed Health Inform, vol. 23, no. 2, pp. 547–559, 2019, doi: 10.1109/JBHI.2018.2806962.

[22] J. R. Hagerty et al., “Deep Learning and Handcrafted Method Fusion: Higher Diagnostic Accuracy for Melanoma Dermoscopy Images,” IEEE J Biomed Health Inform, vol. 23, no. 4, pp. 1385–1391, 2019, doi: 10.1109/JBHI.2019.2891049.

[23] G. Celebi, Emre M.; Wen, Quan; Iyatomi, Hitoshi; Shimizu, Kouhei; Zhou, Huiyu; Schaefer, “A State-of-the-Art on Lesion Border Detection in Dermoscopy Images,” in Dermoscopy Image Analysis, J. S. Celebi, M. Emre; Mendonca, Teresa; Marques, Ed. Boca Raton: CRC Press, 2015, pp. 97–129. [Online]. Available: https://doi.org/10.1201/b19107

[24] N. K. Mishra et al., “Automatic lesion border selection in dermoscopy images using morphology and color features,” Skin Research and Technology, vol. 25, no. 4, pp. 544–552, 2019.

[25] M. E. Celebi, H. Iyatomi, G. Schaefer, and W. v Stoecker, “Lesion border detection in dermoscopy images,” Computerized Medical Imaging and Graphics, vol. 33, no. 2, pp. 148–153, 2009, doi: https://doi.org/10.1016/j.compmedimag.2008.11.002.

[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, 2018, doi: 10.1016/j.cmpb.2018.05.027.

[27] P. Tschandl, C. Sinz, and H. Kittler, “Domain-specific classification-pretrained fully convolutional network encoders for skin lesion segmentation,” Comput Biol Med, vol. 104, pp. 111–116, 2019, doi: https://doi.org/10.1016/j.compbiomed.2018.11.010.

[28] Y. Yuan and Y. C. Lo, “Improving Dermoscopic Image Segmentation With Enhanced Convolutional-Deconvolutional Networks,” IEEE J Biomed Health Inform, vol. 23, no. 2, pp. 519–526, 2019, doi: 10.1109/JBHI.2017.2787487.

[29] F. Xie, J. Yang, J. Liu, Z. Jiang, Y. Zheng, and Y. Wang, “Skin lesion segmentation using high-resolution convolutional neural network,” Comput Methods Programs Biomed, vol. 186, p. 105241, 2020, doi: https://doi.org/10.1016/j.cmpb.2019.105241.

[30] Ş. Öztürk and U. Özkaya, “Skin Lesion Segmentation with Improved Convolutional Neural Network,” J Digit Imaging, vol. 33, no. 4, pp. 958–970, 2020, doi: 10.1007/s10278-020-00343-z.

[31] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation.” [Online]. Available: http://lmb.informatik.uni-freiburg.de/

[32] X. Tong, J. Wei, B. Sun, S. Su, Z. Zuo, and P. Wu, “Ascu-net: Attention gate, spatial and channel attention u-net for skin lesion segmentation,” Diagnostics, vol. 11, no. 3, 2021, doi: 10.3390/diagnostics11030501.

[33] O. Oktay et al., “Attention u-net: Learning where to look for the pancreas,” arXiv preprint arXiv:1804.03999, 2018.

[34] S. Kadry, D. Taniar, R. Damaševičius, V. Rajinikanth, and I. A. Lawal, "Extraction of abnormal skin lesion from dermoscopy image using VGG-SegNet," in 2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII), 2021, pp. 1–5.

[35] V. Rajinikanth, S. Kadry, R. Damaševičius, D. Sankaran, M. A. Mohammed, and S. Chander, "Skin melanoma segmentation using VGG-UNet with Adam/SGD optimizer: a study," in 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), 2022, pp. 982–986.

[36] K. Zafar et al., "Skin lesion segmentation from dermoscopic images using convolutional neural network," Sensors (Switzerland), vol. 20, no. 6, pp. 1–14, 2020, doi: 10.3390/s20061601.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition, 2009, pp. 248–255.

[38] M. Nawaz et al., "Melanoma segmentation: a framework of improved DenseNet77 and UNET convolutional neural network," Int J Imaging Syst Technol, vol. 32, no. 6, pp. 2137–2153, 2022.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[40] D. K. Nguyen, T. T. Tran, C. P. Nguyen, and V. T. Pham, "Skin Lesion Segmentation based on Integrating EfficientNet and Residual block into U-Net Neural Network," Proceedings of 2020 5th International Conference on Green Technology and Sustainable Development, GTSD 2020, pp. 366–371, 2020, doi: 10.1109/GTSD50082.2020.9303084.

[41] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International conference on machine learning, 2019, pp. 6105–6114.

[42] N. C. F. Codella 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)," Proceedings - International Symposium on Biomedical Imaging, vol. 2018-April, pp. 168–172, 2018, doi: 10.1109/ISBI.2018.8363547.

[43] N. Lama et al., "ChimeraNet: U-Net for Hair Detection in Dermoscopic Skin Lesion Images," J Digit Imaging, no. 0123456789, 2022, doi: 10.1007/s10278-022-00740-6.

[44] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[45] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” arXiv preprint arXiv:1511.07122, 2015.

[46] Q. Abbas, M. E. Celebi, and I. F. Garc\’\ia, “Hair removal methods: A comparative study for dermoscopy images,” Biomed Signal Process Control, vol. 6, no. 4, pp. 395–404, 2011.

[47] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.”

[48] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, “Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations,” in Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, 2017, pp. 240–248.

[49] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[50] F. Navarro, M. Escudero-Viñolo, and J. Bescós, “Accurate Segmentation and Registration of Skin Lesion Images to Evaluate Lesion Change,” IEEE J Biomed Health Inform, vol. 23, no. 2, pp. 501–508, 2019, doi: 10.1109/JBHI.2018.2825251.

[51] P. Shan, Y. Wang, C. Fu, W. Song, and J. Chen, “Automatic skin lesion segmentation based on FC-DPN,” Comput Biol Med, vol. 123, no. April, p. 103762, 2020, doi: 10.1016/j.compbiomed.2020.103762.

[52] R. Kaymak, C. Kaymak, and A. Ucar, “Skin lesion segmentation using fully convolutional networks: A comparative experimental study,” Expert Syst Appl, vol. 161, p. 113742, 2020, doi: 10.1016/j.eswa.2020.113742.

[53] M. Goyal, A. Oakley, P. Bansal, D. Dancey, and M. H. Yap, “Skin Lesion Segmentation in Dermoscopic Images with Ensemble Deep Learning Methods,” IEEE Access, vol. 8, pp. 4171–4181, 2020, doi: 10.1109/ACCESS.2019.2960504.

[54] P. Chen, S. Huang, and Q. Yue, “Skin Lesion Segmentation Using Recurrent Attentional Convolutional Networks,” IEEE Access, vol. 10, no. September, pp. 94007–94018, 2022, doi: 10.1109/ACCESS.2022.3204280.

[55] H. Ashraf, A. Waris, M. F. Ghafoor, S. O. Gilani, and I. K. Niazi, “Melanoma segmentation using deep learning with test-time augmentations and conditional random fields,” Sci Rep, vol. 12, no. 1, pp. 1–16, 2022, doi: 10.1038/s41598-022-07885-y.

III. LAMA: LESION-AWARE MIXUP AUGMENTATION FOR SKIN LESION SEGMENTATION

Norsang Lama[1], Joe Stanley[1], Binita Lama[3], Akanksha Maurya[1], Anand Nambisan[1], Jason R. Hagerty[3], Thanh Phan[2], R. William V. Stoecker[3]

[1]Department of Electrical and Computer Engineering,
Missouri University of Science and Technology, Rolla, MO, 65409, USA
[2]Department of Biological Sciences, College of Arts, Sciences, and Education,
Missouri University of Science and Technology, Rolla, MO 65409, USA
[3]S&A Technologies, Rolla, MO, 65401, USA

ABSTRACT

Deep learning can, in experimental image environments, exceed specialists' diagnostic accuracy. However, little progress has been made in accurate segmentation of multiple lesions. Thus, information present in multiple-lesion images, available to specialists, is not retrievable by machine learning. We propose a simple and effective data augmentation technique for skin lesion segmentation in dermoscopic images. The lesion-aware mixup augmentation (LAMA) method generates a synthetic multi-lesion image by mixing two or more lesion images from the training set. To train the deep neural network with the proposed LAMA method, we used the publicly available International Skin Imaging Collaboration (ISIC) 2017 Challenge skin lesion segmentation dataset. As none of the previous skin lesion datasets (including ISIC 2017) have considered multiple lesions per image, we created a new multi-lesion (MuLe) segmentation dataset utilizing publicly available ISIC 2020 skin lesion images with multiple lesions per image. MuLe was used as a test set to evaluate the effectiveness of the proposed method. Our test results show that the proposed method improved the Jaccard score from 0.687 to 0.744 and the Dice score

from 0.7923 to 0.8321 over a baseline model on MuLe test images. On the single-lesion ISIC 2017 test images, LAMA improved the baseline model's segmentation performance by raising the Jaccard score of 0.7947 to 0.8013 and the Dice score of 0.8714 to 0.8766. The experimental results showed that LAMA improved the segmentation accuracy on both single-lesion and multi-lesion dermoscopic images. Although the deep learning model was trained using only synthetic multi-lesion images from the LAMA method, the model successfully detected lesion boundaries on real multi-lesion images. These results show that LAMA can help to train a robust single- or multi-lesion segmentation model.

Keywords — Melanoma, dermoscopy, deep learning, image segmentation, data augmentation

1. INTRODUCTION

An estimated 97,610 new cases of invasive melanoma and 89,070 in-situ melanoma are expected to be diagnosed in 2023 in the United States [1]. Dermoscopy is a powerful imaging technique that assists dermatologists in early skin cancer detection, resulting in improved diagnostic accuracy compared to visual inspection by a domain expert [2]–[4].

Recent advancements in deep learning techniques [5]–[12] have proven immensely beneficial in medical image analysis and have been successfully applied to various medical imaging problems [13]–[16]. In skin cancer, deep learning techniques in combination with dermoscopy have demonstrated higher diagnostic accuracy than experienced dermatologists [13], [17]–[20]. Pathan et al. published a recent review describing handcrafted and deep learning (DL) techniques for computer-aided skin cancer diagnosis

[21]. Recent studies show that the fusion of handcrafted and deep learning techniques boosts the diagnostic accuracy of skin cancer [20], [22]–[26]. The handcrafted features usually require the lesion borders to facilitate their computation [23], [24]. The precise calculation of handcrafted lesion features depends on accurate detection of the lesion border. Therefore, lesion segmentation is a crucial step in computer-aided skin cancer diagnosis.

Skin lesion segmentation methods have been developed using both conventional image processing approaches [27]–[29] and deep learning techniques [30]–[35]. While traditional methods showed promising results on small sets, they tend to perform poorly on challenging conditions such as low contrast between lesion and background, lesions with different colors, and images containing artifacts like hair, ruler marks, gel bubbles, and ink marks. In contrast, deep learning techniques, specifically convolutional neural networks (CNN), seem to perform better in detecting lesion borders in the presence of these challenges.

Data augmentation is a widely used technique in deep learning to artificially increase the size of the training dataset and improve the generalization of the model. Augmentation involves randomly applying image transformations such as rotation, translation, scaling, flipping, color jittering, etc. In addition to these basic transformations, various data augmentation strategies have been developed for image classification, such as Cutout [36], Mixup [37], [38], and Cutmix[39]. Cutout randomly masks out rectangular regions of the input image, which encourages the model to rely on other parts of the image and improves its robustness to occlusion. Mixup generates new training samples by linearly interpolating between pairs of images and their corresponding labels, which improves the

model's ability to generalize to new input distributions. Cutmix combines the cutout and mixup methods by replacing a rectangular region in one image with a corresponding region from another, which can improve the model's ability to learn fine-grained features while preserving the overall structure of the input. Although these methods have proven successful for image classification, they cannot be adopted for skin lesion segmentation as these methods do not consider the location or geometry of the objects (lesion).

This study introduces a novel lesion-aware mixup augmentation (LAMA) method for skin lesion segmentation in dermoscopic images. The LAMA method can generate a new synthetic multi-lesion image by mixing two or more single-lesion images from the training set. Unlike Cutmix, LAMA randomly selects two or more lesions from a set of lesions and paste pastes the new lesions in the non-lesion area of an image. To find the non-lesion area in an image, it applies a multi-level patch generation and qualification process, which is faster as it needs to execute only once at the beginning of training. Also, the lesion size in dermoscopic images varies greatly, with some lesions covering the whole image while others might cover only a small part of the image. To avoid excessive resizing of lesions during mixing, LAMA categorizes the lesions into n groups based on their sizes, with each group corresponding to a specific level of non-lesion patches. And LAMA samples the patch level or patch size while mixing based on the histogram distribution of images having non-lesion patches at each patch level in the training set. We evaluated the proposed method on a benchmark ISIC 2017 [40] skin lesion segmentation dataset. Also, as none of the publicly available datasets or the datasets used in previous studies considered multiple lesions per image, we created a new multi-lesion (MuLe) segmentation dataset using the publicly available ISIC 2020 [41] melanoma classification dataset to evaluate the

effectiveness of the proposed method on real examples having multiple lesions in each image. The experimental results showed that the LAMA method combined with ChimeraNet [35], [42], a U-Net [11] model with EfficientNet [12] encoder, improved the segmentation performance on both the multi-lesion MuLe dataset and the single-lesion ISIC 2017 dataset.

2. LAMA

This section describes our proposed lesion-aware mixup augmentation (LAMA) method for skin lesion segmentation. It generates a new synthetic multi-lesion image by mixing two or more skin lesion images to train the deep neural network. LAMA randomly selects one or more lesions from a training set of single-lesion images and pastes them on the non-lesion area of an image. To find a non-lesion region in the image, LAMA uses a multi-level patch generation and qualification process. As skin lesion size varies greatly, the lesion patches are categorized into n groups based on their sizes, with each group corresponding to a specific level of non-lesion patches. This grouping avoids an excessive resizing of lesion patches when mixed with non-lesion patches. In the following subsections, we describe each process in detail.

2.1. FINDING NON-LESION PATCHES

The candidate patches with various sizes are generated from a non-lesion region of a lesion image. A patch is randomly selected from a pool of candidate patches to paste a new lesion. For an image x $\in R^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$ where C = number of colors (3) and a lesion mask

y $\in R^{H\times W}$, we define n different patch levels $l = 2, 3, \ldots, n+1$ and generate P_l non-overlapping patches of sizes $h_l \times w_l$ for each patch level l as shown in Figure 1, where patch height $h_l = {}^{H}/_{l}$, patch width $w_l = {}^{W}/_{l}$ and patch count $P_l = l^2$.

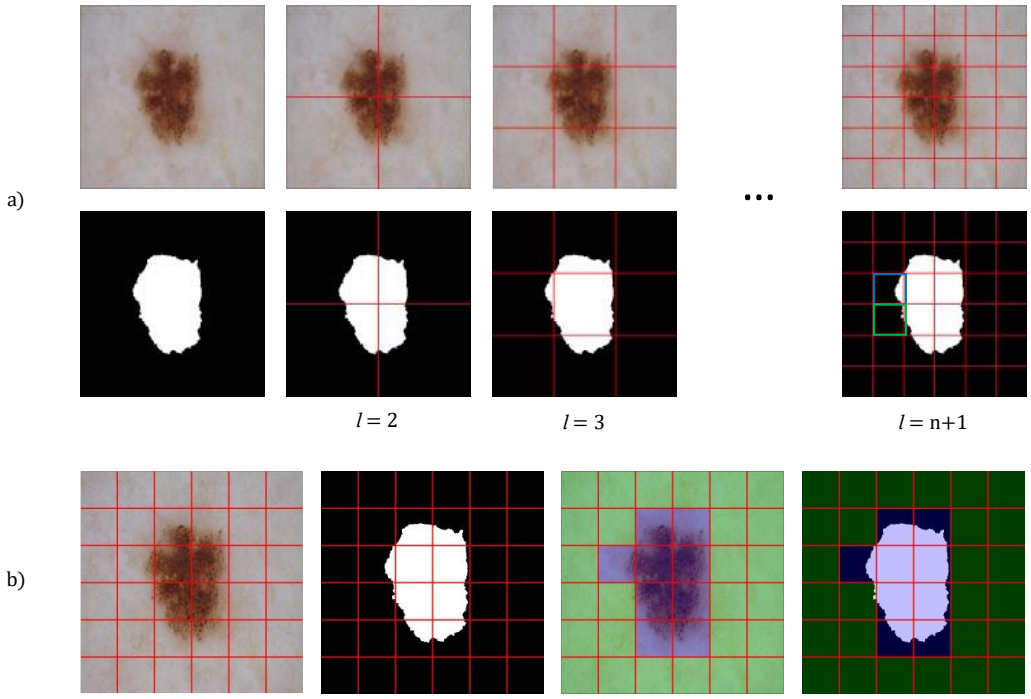
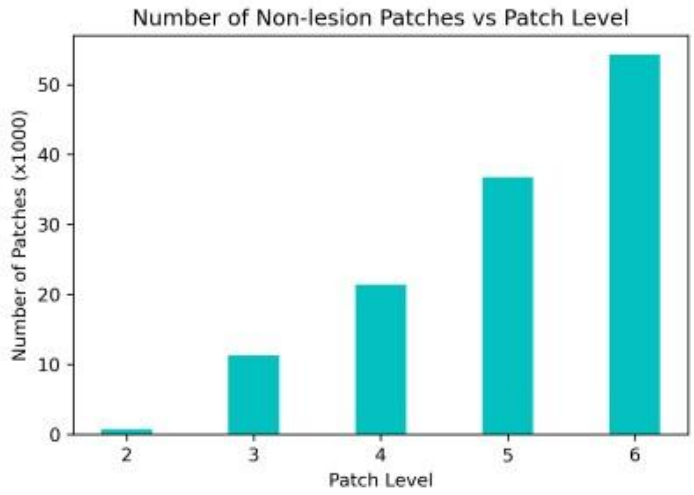


Figure 1. Non-lesion patch generation process. a) Creating patches of size $h_l \times w_l$ at n different patch levels from $l = 2, 3, \ldots, n+1$. b) Finding non-lesion patches at patch level $l = 6$. Non-lesion patches that satisfy the criteria $T_{plpa} < 0.1$ and $T_{plla} < 0.05$ are overlayed with GREEN color. The patches overlayed with BLUE color do not satisfy the criteria and are not added to the pool of candidate patches. In the rightmost mask of (a), the blue patch has $T_{plpa} = 0.17$, so it fails the first criterion and is a lesion patch. However, the green patch has $T_{plpa} = 0.02$, so it satisfies the first criterion and is a non-lesion patch.

To minimize occlusion of an existing lesion by the new lesion, a patch p will be only added to a pool of candidate patches if it satisfies two criteria - a total lesion area within the patch is less than 10% of a total patch area (i.e., $T_{plpa} < 0.1$), and a total lesion

area within the patch is less than 5% of a total lesion area (i.e., $T_{plla} < 0.05$). As the patch count is proportional to a square of patch level l and a lesion usually covers a large part of the image, most candidate non-lesion patches will belong to higher patch levels, i.e., smaller patch sizes, as shown in Figure 2 (a). The distribution of images having non-lesion patches vs. patch level in Figure 2 (b) shows that only a few images have larger non-lesion patches.

a)



b)

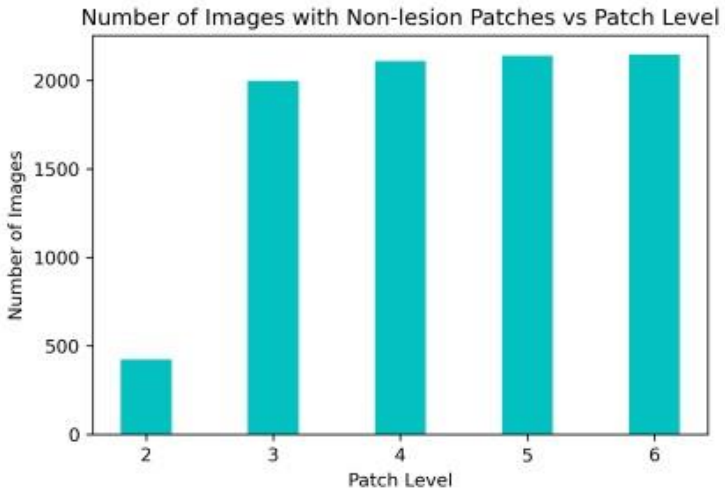


Figure 2. Number of non-lesion patches and images having non-lesion patches for different patch levels (n = 1-5) on ISIC 2017 lesion segmentation training set.

2.2. LESION GROUPS

Skin lesions have different shapes and sizes. Some lesions cover most of the image while others might occupy a tiny part of the image as shown in Figure 3. To avoid excessive resizing of lesions during mixing, all lesions from a training set are grouped into n groups based on their sizes, with each group corresponding to a specific patch level. A lesion from group g_i only mixes with patch level l_i And we compute lesion size in terms of an area of a rectangular bounding box enclosing the lesion. As the non-lesion patches in each patch level are highly imbalanced, as shown in Figure 2 (a), we divide the lesions into n groups based on the histogram of images having non-lesion patches for n patch levels, which is more balanced, as shown in Figure 2 (b). The number of lesions per group is shown in Figure 4.

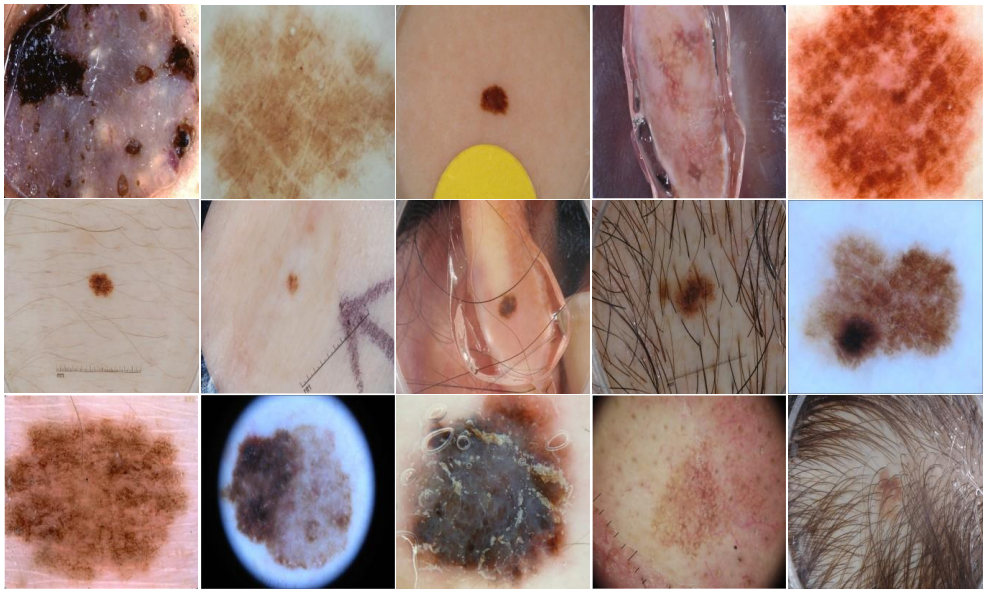


Figure 2. Skin Lesion Images in ISIC 2017 lesion segmentation dataset with varying lesion shapes, sizes, and colors.

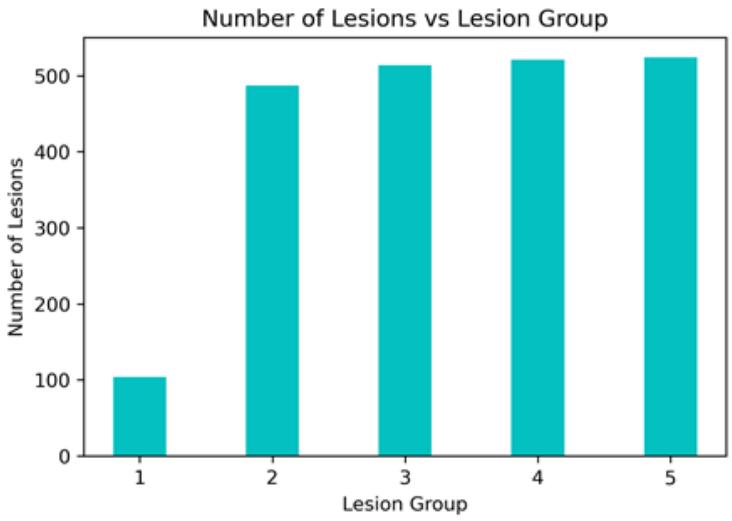


Figure 4. Number of lesions per group in n groups ($n = 1 - 5$).

2.3. MIXING A NEW LESION

To paste a new lesion in a non-lesion patch, first, we randomly select a lesion group g_i with a weighted probability based on the lesion group distribution in Figure 4. Then, a new lesion is randomly selected from the lesion group g_i. The selected lesion is cropped using a rectangular bounding box with extended height or width (extended by 5%) to ensure some background around the lesion. The cropped lesion is further augmented using random image transformations such as rotate by 90°, transpose, vertical or horizontal flip, and Gaussian noise. Next, we randomly select a non-lesion patch from the patch level l_i to paste the augmented lesion crop. The augmented lesion crop needs to be resized to match the size of a non-lesion patch. Finally, the lesion is pasted on the selected non-lesion patch location, and a new training image is created with mixed lesions, as shown in Figure 5.

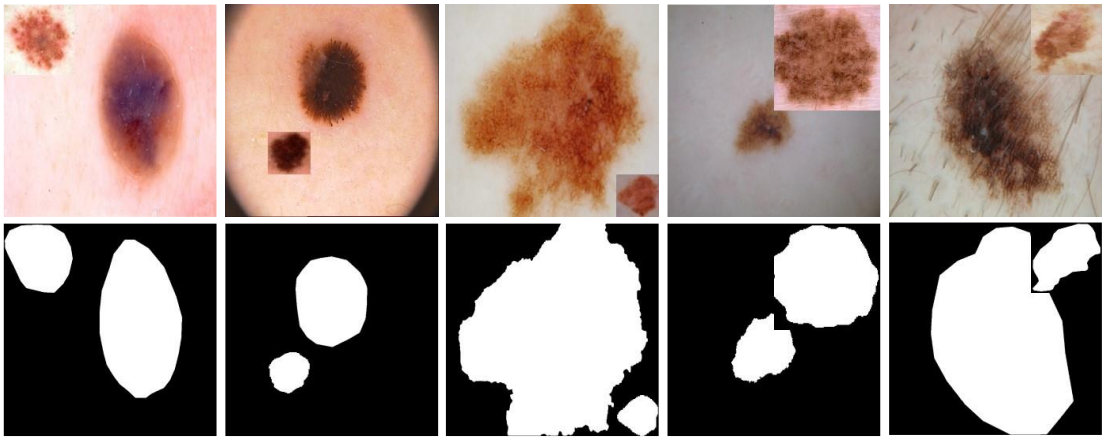


Figure 5. Augmented Lesion images and masks after mixing a single extra lesion using our proposed LAMA method. Some part of the existing lesion is allowed to be occluded as in last two columns.

2.4. ADDING MULTIPLE LESIONS

When adding more lesions in a mixed image, we use a heatmap mask $m \in I^{H \times W}$ to track a non-lesion region of the image so that multiple lesions are not pasted in the same non-lesion patch location. In this mask, a pixel value of 0 represents a non-lesion pixel and a pixel value of 1 represents a lesion pixel. With each new lesion addition, we update the heatmap m by setting a pixel of m to 1 if that pixel belongs to a lesion. The non-lesion patch will be removed from the pool if it does not satisfy an additional criterion of a heatmap area within a patch is less than 5% of a total patch area ($T_{phpa} < 0.05$). And when adding multiple lesions, we vary the number of new lesions by randomly selecting a number between 1 and the maximum number of new lesions (=5) rather than using a fixed number as shown in Figure 6.

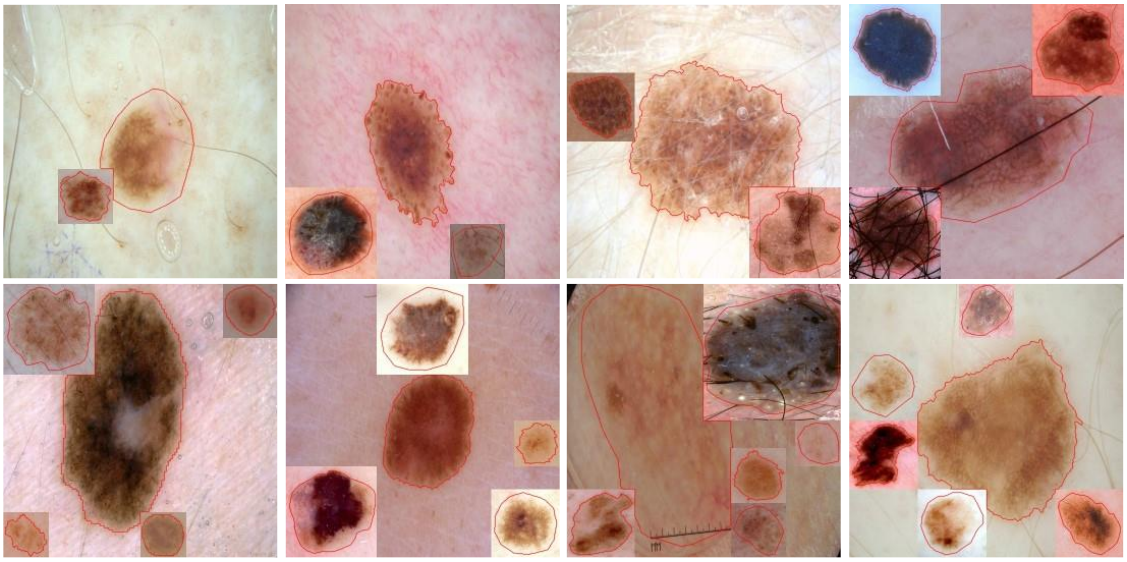


Figure 6. Augmented images after mixing multiple lesions from the proposed method. The lesion boundary (RED line) shows that new lesions do not occlude other lesions. The number of additional lesions varies between 1 and 5.

3. EXPERIMENTS

In this section, we evaluate the performance of a proposed method on the ISIC skin lesion datasets. For our experiments, we used a ChimeraNet architecture from the previous studies [35], [42] that achieved a state-of-art performance on skin lesion segmentation. The architecture uses a pretrained EfficientNet [12] as encoder and a squeeze-and-excitation residual (SERes) structure as decoder. In addition, the decoder network applies a transposed convolution to upsample the feature maps from the encoder repeatedly to generate a final segmentation mask the same size as the input image. In this study, we used ChimeraNet (basic), a simplified version of the model, by removing the SERes structures while only keeping the transposed convolutions in the decoder network.

Table 1. Other image transformations for data augmentation.

Method	**Parameter Values**
Rotate	(-90°, 90°)
Horizontal and vertical Shift	(-0.15, 0.15)
Scaling	(-0.15, 0.15)
Brightness	(-0.15, 0.15)
Contrast	(-0.15, 0.15)
Hue (HSV)	(-10, 10)
Saturation (HSV)	(-20, 20)
Value (HSV)	(-10, 10)
Motion blur	5
Median blur	5
Gaussian blur	5
Gaussian noise	var_limit = (5, 25)
Optical distortion	distort_limit = 1
Grid distortion	num_steps = 6, distort_limit = 1
CLAHE	clip_limit = 4
Transpose	
Horizontal and vertical flip	
RGB shift	
Channel shuffle	

As data augmentation has been an essential step for deep network training and uses more than one image transformation, we combined our augmentation method with other image transforms shown in Table 1. Other techniques such as batch normalization [43], drop out, and early stopping are also used to train the network efficiently. The training hyperparameters are shown in Table 2.

All experiments were implemented using Keras with a Tensorflow backend in Python 3 and trained using a single 32GB Nvidia V100 graphics card. And we use the Albumentations [44] library for the additional image transforms for data augmentation.

Table 2. Training hyperparameters.

Parameter	**Value**
input size	448×448
learning rate	0.0001
batch size	10
epoch	0
dropout probability	0.5
optimizer	Adam [45]
loss method	dice [46]
interpolation	bilinear
early stopping patience	5
number of patch levels	5
maximum number of added lesions	5
lesion bounding box extend	5%

3.1. MULTIPLE-LESION SEGMENTATION

The current publicly available skin lesion segmentation datasets have only one lesion per image. We created a new multi-lesion (MuLe) dataset by randomly selecting 203 images from ISIC skin lesion classification images that might have multiple lesions. Our domain expert annotator manually created the ground truth segmentation masks for these images, and the ground truths were verified by a dermatologist (W.V.S.). As this dataset is small, we only used it as a hold-out test set to evaluate the performance of our proposed method. To train the network, we used the publicly available ISIC 2017 [40] skin lesion segmentation dataset, with one lesion per image. It provides 2750 dermoscopic skin lesion images with ground truth lesion segmentation masks - 2000 training, 150 validation, and 600 test images. As the ISIC 2017 dataset provides a single train-validation split, we combined the official training and validation sets of ISIC2017 to create a single training set of 2150 images to run 5-fold cross-validation experiments. All the images were resized to 540×540 using a bilinear interpolation method. During training, we only applied our proposed method on the training set to generate the synthetic multi-lesion images using the single-lesion images. Figure 7 shows the validation loss plot for the DL models trained with or without our proposed LAMA method. The loss plot shows that the model with LAMA has lower loss than the baseline model. The baseline model without LAMA trained for 21 epochs and exited early when the validation loss stopped improving, while the model with LAMA trained longer with improved validation loss.

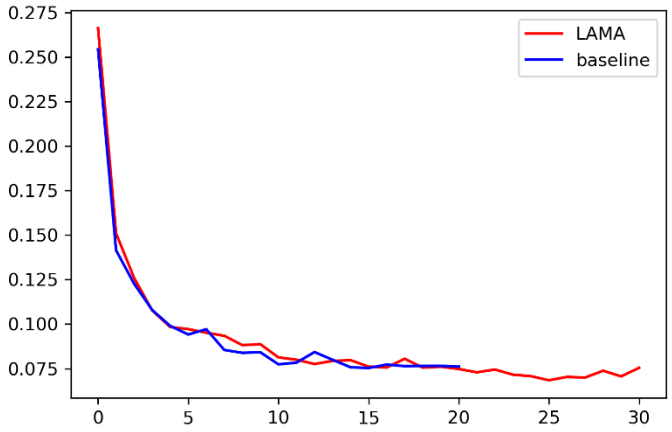


Figure 7. Validation loss plot for ISIC 2017 skin lesion segmentation dataset. The lesion segmentation model trained with the proposed LAMA method has lower loss than the plain model.

Table 3. Performance of the proposed method on multi-lesion (MuLe) test images.

Method	**Jaccard**	**Dice**	**Accuracy**	**Precision**	**Recall**
ChimaraNet (baseline)	0.687	0.7923	0.9666	0.8551	0.8149
ChimeraNet + LAMA	0.7440	0.8321	0.9733	0.8435	0.8846

Table 3 shows the performance of the proposed method on 206 multi-lesion (MuLe) test images trained on the ISIC 2017 lesion segmentation dataset. The ChimeraNet model with LAMA improved overall segmentation performance on real multi-lesion images, although the model was trained using only the synthetic multi-lesion images. The largest improvements were for recall of 0.8149 to 0.8846, Jaccard score of 0.687 to 0.7444, and Dice score of 0.7923 to 0.8321. Precision fell from 0.8551 to 0.8435. Figure 8 shows the

segmentation results of the ChimeraNet model combined with or without the proposed LAMA method. The proposed method successfully finds the multiple lesions even though the model was trained using the synthetic multi-lesion images.

3.2. ISIC 2017 SKIN LESION SEGMENTATION

In this section, we evaluate the lesion segmentation performance of the proposed method on 600 ISIC 2017 test images. All the images in this official test set have a single lesion per image. We use the same trained model in section 3.1 that was trained using the synthetic multi-lesion training images.

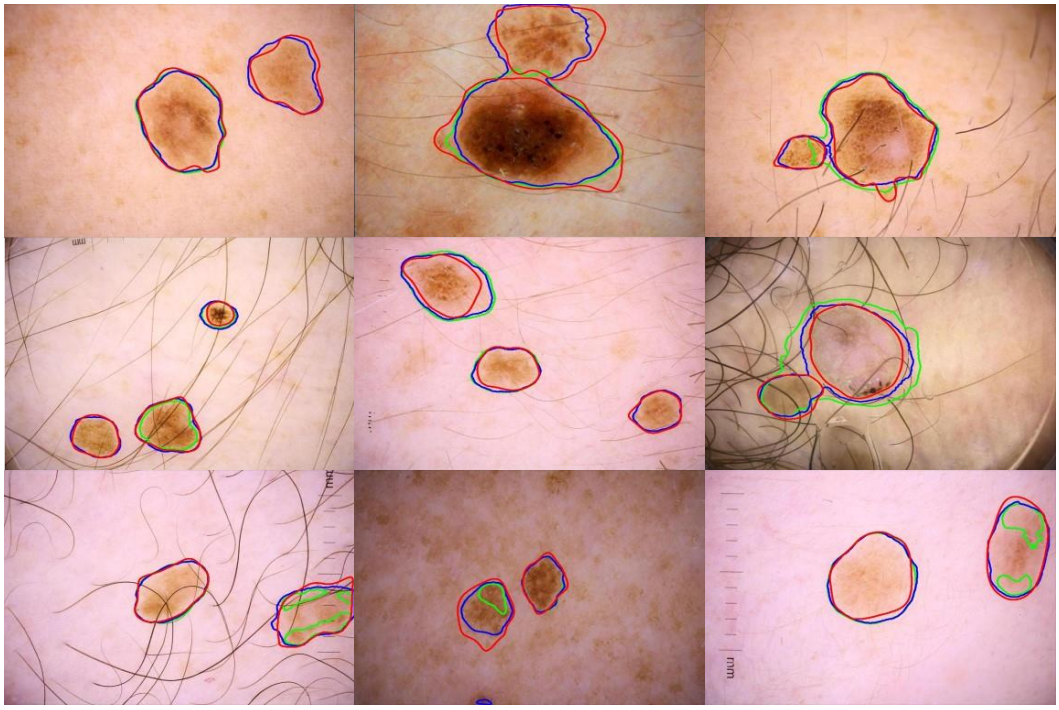


Figure 8. Segmentation results of the proposed method on multi-lesion (MuLe) test set. Overlay of ground truth boundary (RED), predicted boundary with LAMA (BLUE), and predicted boundary without LAMA (GREEN), i.e., the baseline model, on skin lesion images. The baseline model misses a part or some of the multiple lesions.

Table 4. Performance comparison with other lesion segmentation methods on 600 ISIC 2017 test images.

Methods	**Year**	**Jaccard**	**Dice**	**Accuracy**
Al-Masni et al. [30]	2018	0.771	0.871	0.940
Tschandl et al. [31]	2019	0.768	0.851	
Yuan and Lo [32]	2019	0.765	0.849	0.934
Navarro et al.[47]	2019	0.769	0.854	0.955
Xie et al.[33]	2020	0.783	0.862	0.938
Ozturk and Ozkaya [34]	2020	0.783	0.886	0.953
Shan et al.[48]	2020	0.763	0.846	0.937
Kaymak et al.[49]	2020	0.725	0.841	0.939
Nguyen et al.[50]	2020	0.781	0.861	
Zafar et al.[51]	2020	0.772	0.858	
Goyal et al. [52]	2020	0.793	0.871	
Tong et al.[53]	2021	0.742		0.926
Chen et al.[54]	2022	0.8036	0.8704	0.947
Ashraf et al.[55]	2022	0.8005		
Lama et al. [35]	2023	0.807	0.880	0.948
ChimeraNet (baseline)		0.7947	0.8714	0.944
ChimeraNet + LAMA		0.8013	0.8766	0.948

Table 4 shows that the ChimeraNet model with the proposed LAMA method improved all measurements - Jaccard score from 0.7947 to 0.8013, dice score from 0.8714 to 0.8766, and accuracy from 0.944 to 0.948. The proposed method shows robustness by

achieving a Jaccard score of 0.8013, comparable to state-of-the-art methods, even though the model was trained using synthetic multi-lesion images.

In Figure 9, we show the segmentation results of the proposed method on ISIC 2017 test images. The segmentation results show that the proposed method accurately segments the skin lesion even in challenging conditions like images with artifacts or low contrast.

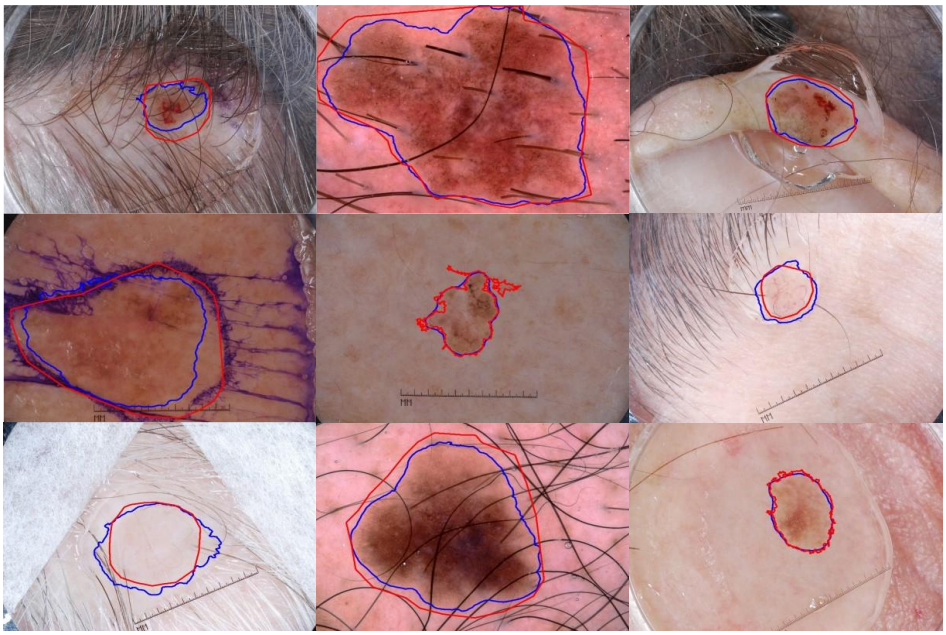


Figure 9. Segmentation results of the proposed method on ISIC 2017 test set. Overlays of ground truth lesion boundary (RED) and predicted lesion boundary (BLUE) on skin lesion images. LAMA finds lesion borders accurately in challenging images like low skin-and-lesion contrast and in the presence of artifacts like hair, ruler marks, and ink marks.

4. DISCUSSION

In this study, we demonstrated that our proposed lesion-aware mixup augmentation (LAMA) method helps train a robust skin lesion segmentation model. Our LAMA method

successfully generated synthetic multi-lesion images by mixing two or more lesion images to train the convolutional neural network. As the currently available public ISIC skin lesion segmentation datasets have only one lesion per image, we created a new multi-lesion (MuLe) test dataset to evaluate the effectiveness of the proposed method on real examples that have one or more lesions in each image. The ChimeraNet model trained with synthetic multi-lesion images successfully detected the lesion boundaries of multiple lesions in real images. LAMA improved the Jaccard score from 0.687 to 0.744, Dice score from 0.7923 to 0.8321, and recall from 0.8149 to 0.8846 on 203 MuLe test images. Also, the LAMA method improved the Jaccard score of 0.7947 to 0.8013 and Dice score of 0.8714 to 8766 when evaluated on the ISIC 2017 skin lesion segmentation test set. Although the model was trained using synthetic images to detect multiple lesions per image, the model achieved results comparable to state-of-the-methods on single-lesion test images. LAMA segmentation improvements are most noticeable on challenging images, as shown in Figure 6.

Data augmentation methods like Mixup [37] and Cutmix [39] have improved the generalization capability of a deep neural network by providing a regularization effect during training. Mixup is specifically designed for the image classification task and cannot be adopted for image segmentation. The proposed LAMA method is very similar to Cutmix, as both randomly replace the patches from one image with patches from another. However, CutMix does not consider the location and geometry of the objects (for example, lesions in our case) while selecting the patches of the images. Also, it mixes only two images, so it cannot generate a synthetic image with more than two lesions for our problem. The proposed LAMA solves these shortcomings of Cutmix by creating the lesion patches

using the bounding box covering the whole lesion and then pasting one or more lesions in the non-lesion area of the image. LAMA finds the non-lesion region of the image using a multi-level patch generation and selection scheme. This process is optimized, running only once at the beginning of training, and using the same set for all epochs. The other challenge for mixing skin lesions is due to a large variability in lesion sizes. To avoid extreme resizing of lesions during mixing, all lesion patches are grouped into n groups based on their sizes, and each group corresponds to one of the n patch levels (or patch sizes). This ensures that the lesion patch and non-lesion patch have a similar patch size. In addition, the proposed LAMA method augments the lesion patches before pasting them on the non-lesion area of an image such that it helps to generate more variation of synthetic images.

Although the significant color variability in skin or background in skin lesion images can cause a noticeable distinction between the original image background and newly added patches when these images are mixed, as shown in Figure 6, the experimental results have revealed that deep neural networks can successfully detect the actual boundaries of lesions and are not affected by such sharp demarcations.

The proposed method can train segmentation of multiple lesions, potentially supplying additional information to machine learning systems. The ugly duckling sign [56], for example, shows that a given lesion differs from the patient's other lesions. This sign, most applicable to clinical images, may also appear in dermoscopy images. Machine learning lacks such information. The LAMA technique can synthesize multi-lesion samples from single-lesion images to train for the segmentation of multiple lesions, potentially providing new information for machine learning.

5. CONCLUSION

In this study, we proposed a novel lesion-aware mixup augmentation (LAMA) method to train a robust deep neural network for skin lesion segmentation in dermoscopic images. None of the previous studies on skin lesion segmentation considered more than one lesion in dermoscopic skin lesion images. Therefore, we created a new multi-lesion (MuLe) segmentation dataset using publicly available ISIC skin lesion images to evaluate our proposed method. The proposed LAMA method effectively produced synthetic multi-lesion images by utilizing a training set of single-lesion images and their corresponding ground truth masks. The experimental results show that the ChimeraNet lesion segmentation model trained with LAMA not only successfully detects multiple lesions on real-life examples but also enhances the segmentation performance of single-lesion images. Further study of the LAMA technique is warranted.

REFERENCES

[1] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, “Cancer statistics, 2023,” CA Cancer J Clin, vol. 73, no. 1, pp. 17–48, 2023, doi: https://doi.org/10.3322/caac.21763.

[2] H. Pehamberger, M. Binder, A. Steiner, and K. Wolff, “In vivo epiluminescence microscopy: Improvement of early diagnosis of melanoma,” Journal of Investigative Dermatology, vol. 100, no. 3 SUPPL., pp. S356–S362, 1993, doi: 10.1038/jid.1993.63.

[3] H. P. Soyer, G. Argenziano, R. Talamini, and S. Chimenti, “Is Dermoscopy Useful for the Diagnosis of Melanoma?,” Arch Dermatol, vol. 137, no. 10, pp. 1361–1363, Oct. 2001, doi: 10.1001/archderm.137.10.1361.

[4] R. P. Braun, H. S. Rabinovitz, M. Oliviero, A. W. Kopf, and J. H. Saurat, “Pattern analysis: a two-step procedure for the dermoscopic diagnosis of melanoma,” Clin Dermatol, vol. 20, no. 3, pp. 236–239, May 2002, doi: 10.1016/S0738-081X(02)00216-X.

[5] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in Advances in Neural Information and Processing Systems (NIPS), 2012, pp. 1097–1105.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[8] I. Goodfellow et al., “Generative adversarial networks,” Commun ACM, vol. 63, no. 11, pp. 139–144, 2020.

[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[10] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv preprint arXiv:2010.11929, 2020.

[11] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation.” [Online]. Available: http://lmb.informatik.uni-freiburg.de/

[12] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in International conference on machine learning, 2019, pp. 6105–6114.

[13] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.

[14] V. Gulshan et al., “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.

[15] S. Sornapudi et al., “Deep learning nuclei detection in digitized histology images by superpixels,” J Pathol Inform, vol. 9, no. 1, p. 5, 2018.

[16] G. Litjens et al., “A survey on deep learning in medical image analysis,” Med Image Anal, vol. 42, pp. 60–88, 2017, doi: https://doi.org/10.1016/j.media.2017.07.005.

[17] L. K. Ferris et al., “Computer-aided classification of melanocytic lesions using dermoscopic images,” J Am Acad Dermatol, vol. 73, no. 5, pp. 769–776, Nov. 2015, doi: 10.1016/J.JAAD.2015.07.028.

[18] M. A. Marchetti et al., “Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images,” J Am Acad Dermatol, vol. 78, no. 2, pp. 270-277.e1, Feb. 2018, doi: 10.1016/j.jaad.2017.08.016.

[19] H. A. Haenssle et al., “Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists,” Annals of Oncology, vol. 29, no. 8, pp. 1836–1842, 2018, doi: https://doi.org/10.1093/annonc/mdy166.

[20] N. C. F. Codella et al., “Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images,” IBM J. Res. Dev., vol. 61, no. 4–5, pp. 5:1–5:15, Jul. 2017, doi: 10.1147/JRD.2017.2708299.

[21] S. Pathan, K. G. Prabhu, and P. C. 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, doi: 10.1016/J.BSPC.2017.07.010.

[22] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg, “Combining deep learning and hand-crafted features for skin lesion classification,” 2016 6th International Conference on Image Processing Theory, Tools and Applications, IPTA 2016, 2017, doi: 10.1109/IPTA.2016.7821017.

[23] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, “Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images BT - Machine Learning in Medical Imaging,” L. Zhou, L. Wang, Q. Wang, and Y. Shi, Eds., Cham: Springer International Publishing, 2015, pp. 118–126.

[24] I. González-Díaz, “DermaKNet: Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for Skin Lesion Diagnosis,” IEEE J Biomed Health Inform, vol. 23, no. 2, pp. 547–559, 2019, doi: 10.1109/JBHI.2018.2806962.

[25] J. R. Hagerty et al., “Deep Learning and Handcrafted Method Fusion: Higher Diagnostic Accuracy for Melanoma Dermoscopy Images,” IEEE J Biomed Health Inform, vol. 23, no. 4, pp. 1385–1391, 2019, doi: 10.1109/JBHI.2019.2891049.

[26] A. K. Nambisan et al., “Improving Automatic Melanoma Diagnosis Using Deep Learning-Based Segmentation of Irregular Networks,” Cancers (Basel), vol. 15, no. 4, 2023, doi: 10.3390/cancers15041259.

[27] G. Celebi, Emre M.; Wen, Quan; Iyatomi, Hitoshi; Shimizu, Kouhei; Zhou, Huiyu; Schaefer, “A State-of-the-Art on Lesion Border Detection in Dermoscopy Images,” in Dermoscopy Image Analysis, J. S. Celebi, M. Emre; Mendonca, Teresa; Marques, Ed., Boca Raton: CRC Press, 2015, pp. 97–129. [Online]. Available: https://doi.org/10.1201/b19107

[28] N. K. Mishra et al., “Automatic lesion border selection in dermoscopy images using morphology and color features,” Skin Research and Technology, vol. 25, no. 4, pp. 544–552, 2019.

[29] M. E. Celebi, H. Iyatomi, G. Schaefer, and W. v Stoecker, “Lesion border detection in dermoscopy images,” Computerized Medical Imaging and Graphics, vol. 33, no. 2, pp. 148–153, 2009, doi: https://doi.org/10.1016/j.compmedimag.2008.11.002.

[30] 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, 2018, doi: 10.1016/j.cmpb.2018.05.027.

[31] P. Tschandl, C. Sinz, and H. Kittler, “Domain-specific classification-pretrained fully convolutional network encoders for skin lesion segmentation,” Comput Biol Med, vol. 104, pp. 111–116, 2019, doi: https://doi.org/10.1016/j.compbiomed.2018.11.010.

[32] Y. Yuan and Y. C. Lo, “Improving Dermoscopic Image Segmentation With Enhanced Convolutional-Deconvolutional Networks,” IEEE J Biomed Health Inform, vol. 23, no. 2, pp. 519–526, 2019, doi: 10.1109/JBHI.2017.2787487.

[33] F. Xie, J. Yang, J. Liu, Z. Jiang, Y. Zheng, and Y. Wang, “Skin lesion segmentation using high-resolution convolutional neural network,” Comput Methods Programs Biomed, vol. 186, p. 105241, 2020, doi: https://doi.org/10.1016/j.cmpb.2019.105241.

[34] Ş. Öztürk and U. Özkaya, “Skin Lesion Segmentation with Improved Convolutional Neural Network,” J Digit Imaging, vol. 33, no. 4, pp. 958–970, 2020, doi: 10.1007/s10278-020-00343-z.

[35] N. Lama, J. Hagerty, A. Nambisan, R. J. Stanley, and W. V. Stoecker, “Skin Lesion Segmentation in Dermoscopic Images with Noisy Data,” J Digit Imaging, 2023, doi: 10.1007/s10278-023-00819-8.

[36] T. DeVries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout,” arXiv preprint arXiv:1708.04552, 2017.

[37] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” arXiv preprint arXiv:1710.09412, 2017.

[38] Y. Tokozume, Y. Ushiku, and T. Harada, “Between-class learning for image classification,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5486–5494.

[39] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, “Cutmix: Regularization strategy to train strong classifiers with localizable features,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6023–6032.

[40] N. C. F. Codella 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),” Proceedings - International Symposium on Biomedical Imaging, vol. 2018-April, pp. 168–172, 2018, doi: 10.1109/ISBI.2018.8363547.

[41] V. Rotemberg et al., “A patient-centric dataset of images and metadata for identifying melanomas using clinical context,” Sci Data, vol. 8, no. 1, p. 34, 2021.

[42] N. Lama et al., “ChimeraNet: U-Net for Hair Detection in Dermoscopic Skin Lesion Images,” J Digit Imaging, no. 0123456789, 2022, doi: 10.1007/s10278-022-00740-6.

[43] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.”

[44] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: fast and flexible image augmentations,” Information, vol. 11, no. 2, p. 125, 2020.

[45] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[46] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, “Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations,” in Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, 2017, pp. 240–248.

[47] F. Navarro, M. Escudero-Viñolo, and J. Bescós, “Accurate Segmentation and Registration of Skin Lesion Images to Evaluate Lesion Change,” IEEE J Biomed Health Inform, vol. 23, no. 2, pp. 501–508, 2019, doi: 10.1109/JBHI.2018.2825251.

[48] P. Shan, Y. Wang, C. Fu, W. Song, and J. Chen, “Automatic skin lesion segmentation based on FC-DPN,” Comput Biol Med, vol. 123, no. April, p. 103762, 2020, doi: 10.1016/j.compbiomed.2020.103762.

[49] R. Kaymak, C. Kaymak, and A. Ucar, “Skin lesion segmentation using fully convolutional networks: A comparative experimental study,” Expert Syst Appl, vol. 161, p. 113742, 2020, doi: 10.1016/j.eswa.2020.113742.

[50] D. K. Nguyen, T. T. Tran, C. P. Nguyen, and V. T. Pham, “Skin Lesion Segmentation based on Integrating EfficientNet and Residual block into U-Net Neural Network,” Proceedings of 2020 5th International Conference on Green Technology and Sustainable Development, GTSD 2020, pp. 366–371, 2020, doi: 10.1109/GTSD50082.2020.9303084.

[51] K. Zafar et al., “Skin lesion segmentation from dermoscopic images using convolutional neural network,” Sensors (Switzerland), vol. 20, no. 6, pp. 1–14, 2020, doi: 10.3390/s20061601.

[52] M. Goyal, A. Oakley, P. Bansal, D. Dancey, and M. H. Yap, “Skin Lesion Segmentation in Dermoscopic Images with Ensemble Deep Learning Methods,” IEEE Access, vol. 8, pp. 4171–4181, 2020, doi: 10.1109/ACCESS.2019.2960504.

[53] X. Tong, J. Wei, B. Sun, S. Su, Z. Zuo, and P. Wu, “Ascu-net: Attention gate, spatial and channel attention u-net for skin lesion segmentation,” Diagnostics, vol. 11, no. 3, 2021, doi: 10.3390/diagnostics11030501.

[54] P. Chen, S. Huang, and Q. Yue, “Skin Lesion Segmentation Using Recurrent Attentional Convolutional Networks,” IEEE Access, vol. 10, no. September, pp. 94007–94018, 2022, doi: 10.1109/ACCESS.2022.3204280.

[55] H. Ashraf, A. Waris, M. F. Ghafoor, S. O. Gilani, and I. K. Niazi, “Melanoma segmentation using deep learning with test-time augmentations and conditional random fields,” Sci Rep, vol. 12, no. 1, pp. 1–16, 2022, doi: 10.1038/s41598-022-07885-y.

[56] A. Scope, et al. The "ugly duckling" sign: agreement between observers. Archives of Dermatology, vol. 144, no. 1, pp. 58-64. doi: 10.1001/archdermatol.2007.15.

SECTION

2. SUMMARY AND CONCLUSIONS

Developing accurate image segmentation algorithms is crucial for computer-aided diagnosis of skin cancer. This dissertation proposes novel deep learning methods to improve segmentation accuracy on three different segmentation tasks in dermoscopic skin lesion images. First, a novel DL architecture, named ChimeraNet, was developed for hair and ruler mark detection. ChimeraNet was further modified to accommodate the lesion segmentation task. Furthermore, noisy data's effect on training data was investigated as the benchmark lesion segmentation dataset contains many noisy or inaccurate ground truth labels. Finally, a novel data augmentation technique was proposed to generate synthetic multi-lesion images by mixing one or more single lesions. Multi-lesion image generation with the proposed lesion-aware mixup augmentation (LAMA) method enables multi-lesion segmentation in dermoscopic lesion images. Experimental results showed that the proposed DL methods have state-of-the-art performance in various segmentation tasks in dermoscopic skin lesion images. Furthermore, these DL techniques are simple and efficient, so they can be easily transferred to other medical image segmentation problems.

BIBLIOGRAPHY

[1] American Cancer Society, “Cancer Facts & Figures 2023,” 2023. Accessed: Apr. 24, 2023. [Online]. Available: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/2023-cancer-facts-figures.html

[2] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, “Cancer statistics, 2023,” CA Cancer J Clin, vol. 73, no. 1, pp. 17–48, 2023, doi: https://doi.org/10.3322/caac.21763.

[3] H. Pehamberger, M. Binder, A. Steiner, and K. Wolff, “In vivo epiluminescence microscopy: Improvement of early diagnosis of melanoma,” Journal of Investigative Dermatology, vol. 100, no. 3 SUPPL., pp. S356–S362, 1993, doi: 10.1038/jid.1993.63.

[4] H. P. Soyer, G. Argenziano, R. Talamini, and S. Chimenti, “Is Dermoscopy Useful for the Diagnosis of Melanoma?,” Arch Dermatol, vol. 137, no. 10, pp. 1361–1363, Oct. 2001, doi: 10.1001/archderm.137.10.1361.

[5] R. P. Braun, H. S. Rabinovitz, M. Oliviero, A. W. Kopf, and J. H. Saurat, “Pattern analysis: a two-step procedure for the dermoscopic diagnosis of melanoma,” Clin Dermatol, vol. 20, no. 3, pp. 236–239, May 2002, doi: 10.1016/S0738-081X(02)00216-X.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Commun ACM, vol. 60, no. 6, pp. 84–90, 2017.

[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[9] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in International conference on machine learning, 2019, pp. 6105–6114.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[11] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.

[12] N. C. F. Codella et al., “Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images,” IBM J. Res. Dev., vol. 61, no. 4–5, pp. 5:1–5:15, Jul. 2017, doi: 10.1147/JRD.2017.2708299.

[13] M. A. Marchetti et al., “Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images,” J Am Acad Dermatol, vol. 78, no. 2, pp. 270-277.e1, Feb. 2018, doi: 10.1016/j.jaad.2017.08.016.

[14] J. R. Hagerty et al., “Deep Learning and Handcrafted Method Fusion: Higher Diagnostic Accuracy for Melanoma Dermoscopy Images,” IEEE J Biomed Health Inform, vol. 23, no. 4, pp. 1385–1391, 2019, doi: 10.1109/JBHI.2019.2891049.

[15] A. K. Nambisan et al., “Improving Automatic Melanoma Diagnosis Using Deep Learning-Based Segmentation of Irregular Networks,” Cancers (Basel), vol. 15, no. 4, 2023, doi: 10.3390/cancers15041259.

[16] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg, “Combining deep learning and hand-crafted features for skin lesion classification,” 2016 6th International Conference on Image Processing Theory, Tools and Applications, IPTA 2016, 2017, doi: 10.1109/IPTA.2016.7821017.

[17] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, “Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images BT - Machine Learning in Medical Imaging,” L. Zhou, L. Wang, Q. Wang, and Y. Shi, Eds., Cham: Springer International Publishing, 2015, pp. 118–126.

[18] T. Lee, V. Ng, R. Gallagher, A. Coldman, and D. McLean, “Dullrazor®: A software approach to hair removal from images,” Comput Biol Med, vol. 27, no. 6, pp. 533–543, Nov. 1997, doi: 10.1016/S0010-4825(97)00020-6.

[19] Q. Abbas, I. F. Garcia, M. Emre Celebi, and W. Ahmad, “A feature-preserving hair removal algorithm for dermoscopy images,” Skin Research and Technology, vol. 19, no. 1, pp. e27--e36, 2013.

[20] R. Kasmi et al., “SharpRazor: Automatic Removal of Hair and Ruler Marks from Dermoscopy Images,” Skin Research and Technology, 2021.

[21] N. K. Mishra et al., “Automatic lesion border selection in dermoscopy images using morphology and color features,” Skin Research and Technology, vol. 25, no. 4, pp. 544–552, 2019.

[22] G. Celebi, Emre M.; Wen, Quan; Iyatomi, Hitoshi; Shimizu, Kouhei; Zhou, Huiyu; Schaefer, “A State-of-the-Art on Lesion Border Detection in Dermoscopy Images,” in Dermoscopy Image Analysis, J. S. Celebi, M. Emre; Mendonca, Teresa; Marques, Ed., Boca Raton: CRC Press, 2015, pp. 97–129. [Online]. Available: https://doi.org/10.1201/b19107

[23] M. E. Celebi, H. Iyatomi, G. Schaefer, and W. V Stoecker, “Lesion border detection in dermoscopy images,” Computerized Medical Imaging and Graphics, vol. 33, no. 2, pp. 148–153, 2009, doi: https://doi.org/10.1016/j.compmedimag.2008.11.002.

[24] W. Li, A. N. Joseph Raj, T. Tjahjadi, and Z. Zhuang, “Digital hair removal by deep learning for skin lesion segmentation,” Pattern Recognit, vol. 117, 2021, doi: 10.1016/j.patcog.2021.107994.

[25] M. Attia, M. Hossny, H. Zhou, S. Nahavandi, H. Asadi, and A. Yazdabadi, “Digital hair segmentation using hybrid convolutional and recurrent neural networks architecture,” Comput Methods Programs Biomed, vol. 177, pp. 17–30, 2019, doi: 10.1016/j.cmpb.2019.05.010.

[26] Ş. Öztürk and U. Özkaya, “Skin Lesion Segmentation with Improved Convolutional Neural Network,” J Digit Imaging, vol. 33, no. 4, pp. 958–970, 2020, doi: 10.1007/s10278-020-00343-z.

[27] F. Xie, J. Yang, J. Liu, Z. Jiang, Y. Zheng, and Y. Wang, “Skin lesion segmentation using high-resolution convolutional neural network,” Comput Methods Programs Biomed, vol. 186, p. 105241, 2020, doi: https://doi.org/10.1016/j.cmpb.2019.105241.

[28] R. Kaymak, C. Kaymak, and A. Ucar, “Skin lesion segmentation using fully convolutional networks: A comparative experimental study,” Expert Syst Appl, vol. 161, p. 113742, 2020, doi: 10.1016/j.eswa.2020.113742.

[29] 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, 2018, doi: 10.1016/j.cmpb.2018.05.027.

[30] A. Maurya et al., “A deep learning approach to detect blood vessels in basal cell carcinoma,” Skin Research and Technology, vol. 28, no. 4, pp. 571–576, 2022.

[31] A. K. Nambisan et al., “Deep learning-based dot and globule segmentation with pixel and blob-based metrics for evaluation,” Intelligent Systems with Applications, vol. 16, p. 200126, 2022, doi: https://doi.org/10.1016/j.iswa.2022.200126.

[32] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation.” [Online]. Available: http://lmb.informatik.uni-freiburg.de/

[33] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[34] N. C. F. Codella 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),” Proceedings - International Symposium on Biomedical Imaging, vol. 2018-April, pp. 168–172, 2018, doi: 10.1109/ISBI.2018.8363547.

VITA

Norsang Lama was born in Thulopakhar, Sindhupalchok, Nepal. He received his bachelor's degree in Electronics and Communication Engineering from Tribhuvan University in 2012. He worked as a Lecturer at Department of Electronics and Computer Engineering, Kantipur Engineering College, Nepal from December 2011 to July 2015. In August 2015, he joined Southern Illinois University Edwardsville to pursue his graduate studies and received his master's degree in electrical engineering in July 2017. He received his PhD degree in Computer Engineering from Missouri University of Science and Technology in July 2023. His research interests were computer vision, image analysis, machine learning, and deep learning.